written and edited by Ronald E. Kumon

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.

The LibreTexts mission is to unite students, faculty and scholars in a cooperative effort to develop an easy-to-use online platform for the construction, customization, and dissemination of OER content to reduce the burdens of unreasonable textbook costs to our students and society. The LibreTexts project is a multi-institutional collaborative venture to develop the next generation of open-access texts to improve postsecondary education at all levels of higher learning by developing an Open Access Resource environment. The project currently consists of 14 independently operating and interconnected libraries that are constantly being optimized by students, faculty, and outside experts to supplant conventional paper-based books. These free textbook alternatives are organized within a central environment that is both vertically (from advance to basic level) and horizontally (across different fields) integrated.

The LibreTexts libraries are Powered by NICE CXOne and are supported by the Department of Education Open Textbook Pilot Project, 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.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org. More information on our activities can be found via Facebook (https://facebook.com/Libretexts), Twitter (https://twitter.com/libretexts), or our blog (http://Blog.Libretexts.org).

This text was compiled on 04/15/2025

TABLE OF CONTENTS

18: Calculation of Magnetic Quantities from Currents

19: Alternating-Current (AC) Circuits

20: Maxwell's Equations

21: Electrical Transmission Lines

Detailed Licensing

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing****.*

Preface

Welcome to *Electricity and Magnetism for Amateur Radio and Wireless Technology*! This book is designed to teach you about the physics of electricity and magnetism at the introductory undergraduate university level. This book distinguishes itself from other physics textbooks on this topic by placing the subject into the specific context of amateur radio and other wireless technology. Ideally, you will be learning about electricity and magnetism and also preparing to become an amateur radio operator at an entry-level (e.g., Technician class in the United States) at the same time.

No prior experience or knowledge is assumed about electricity or magnetism, but you are expected to have some background in the concepts and practice of Newtonian physics at a college or university level (e.g., OpenStax University Physics, vol. 1, or College Physics, 2nd ed., Ch. 1–8). This book also assumes some prior knowledge of algebra, geometry, trigonometry, vectors, and differential and integral calculus to understand the mathematical content fully. A background in vector calculus and differential equations will be helpful, but not necessary, as those topics are developed in the text as needed. That said, the book also includes numerous discussions of physics concepts that do not require mathematics to understand or apply, so do not let your math background get in the way of the interesting and amazing physics that awaits!

This project has been inspired and supported by numerous sources. I want to acknowledge the following people and organizations:

- Terrell Teaching Innovation Grant for 2023-24 from the Center for Excellence in Teaching & Learning at Kettering University (Cindy Donner, Director)
- Department of Natural Sciences at Kettering University (Prof. Dan Ludwigsen, Department Head, and colleagues in the department)
- Department of Electrical and Computer Engineering at Kettering University (Prof. Mark Thompson, Department Head, and faculty)
- Kettering University Amateur Radio and Electronics Club (including recent presidents, Ruth Willet, KM4LAO; Quinn Austin, KE8TKJ; Jake Dancel, KE8TMJ; Jason Aurilia, KE8YLQ, and faculty member, Prof. Larry Navarre, N8LJN)
- American Radio Relay League, including their Teachers Institute on Wireless Technology, Collegiate Amateur Radio Program, and numerous publications (including Steve Goodgame, K5ATA; Eliza Croarkin, KC1TAP; Marvin Turner, W0MET; Tommy Gober, N5DUX; Wayne Greene, KB4DSF; Bob Inderbitzen, NQ1R)
- Livonia Amateur Radio Club (LARC, K8UNS)
- Ham Radio Science Citizen Initiative (HamSCI)
- Ham Radio Question Pool (Russ Olsen)
- National Conference of Volunteer Examiner Coordinators (NCVEC)
- OpenStax project from Rice University (Much of the core content of the book is a derivative of their textbook, University Physics, vol. 2)
- LibreTexts project (including Prof. Delmar Larsen, Josh Halpern, Eric Kean, Jennifer Rogers, and the rest of the team)
- Various authors of relevant open-educational textbooks including
  - Steven W. Ellingson, Electromagnetics, vol 1; Electromagnetics, vol. 2; Radio Systems Engineering. revised 1st ed.
  - Tony R. Kuphaldt, Electrical Circuits I (DC), Electrical Circuits II (AC), Electrical Circuits III (Semiconductors),
  - Michael Steer, Fundamentals of Microwave and RF Design
  - Don H. Johnson, Electrical Engineering

Finally, I would like to thank my family for their patience while I have been spending so much time on this project!

Ronald E. Kumon (K8DTJ)
Associate Professor of Physics
Kettering University

CHAPTER OVERVIEW

1: Preliminary Concepts

1.1: What is Electricity and Magnetism?

Learning Objectives

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

- Define electricity and magnetism.
- Recognize that electricity and magnetism are different aspects of electromagnetism.

Electricity is a branch of physics concerned with the effects "associated with the presence and motion of matter possessing an electric charge" [1]. As we will see, an **electric charge** is associated with an **electric field** that can affect other electric charges around it. Common natural phenomena that involve electricity, including lightning and static electricity (Fig. 1.1.1). Electricity also plays a central role in many modern technologies that are pervasive through everyday life, including electric heating, electric lighting, and electric circuits in cellular phones, computers, appliances, automobiles, and other devices that use electrical power. Indeed, it is hard to wake up in the morning, get through your day, and get back to bed at night without encountering multiple electric devices or phenomena!

Figure 1.1.1: Lightning [2] and static electricity [3] are examples of a natural electrical phenomenon.

Magnetism is a branch of physics concerned with the effects associated with a **magnetic field**. In contrast to electric charges, magnetic fields do not arise from separate "magnetic charges." Instead, we will see that they are caused by the motion of electric charges or their intrinsic magnetic properties. While magnetic fields are perhaps most commonly associated with permanent iron magnets (Fig. 1.1.2), a temporary magnetic field can be created by electric charges moving through a coil of wire. Like electricity, magnetism also plays a critical role in modern technologies, including motors, generators, relays, solenoids, loudspeakers, hard drives, and many other examples [4].

Figure 1.1.2: Iron filings are attracted to an permanent bar magnet. [5]

While we will start our study of electricity and magnetism by examining them separately for simplicity, it turns out that it will eventually be better to think about the two phenomena as different aspects of one phenomenon called **electromagnetism** [6]. Oscillating charges will generate **electromagnetic waves**, which are oscillations in the **electromagnetic field** that travel from one location to another. The **frequency** of the wave is its number of oscillations per second. Radio waves, infrared light, visible light, ultraviolet light, x-rays, gamma rays are all examples of electromagnetic waves with different frequency ranges in the **electromagnetic spectrum** [7].

Given the wide range of electromagnetic technologies, it is not possible to include them all in an introductory textbook. In this book, we will be focusing on wireless technology and Amateur Radio as a means to show how the principles of electricity and magnetism can be used in practice.

References

1. Wikipedia contributors. Electricity [Internet]. Wikipedia, The Free Encyclopedia.
2. Wikimedia Commons contributors. File:London MMB »1E6 Lightning.jpg [Internet]. Wikimedia Commons. (CC BY-SA 4.0; mattbuck)
3. OpenStax, Prelude to Electric Charges and Fields, University Physics II. [Internet]. LibreTexts. (CC-BY 4.0; Modification of work by Sean McGrath)
4. Wikipedia contributors. Magnetism [Internet]. Wikipedia, The Free Encyclopedia.
5. Wikimedia Commons contributors. File:Metal shavings around bar magnet (close-up).jpg [Internet]. Wikimedia Commons. (CC BY-SA 4.0; Maciej J. Mrowinski)
6. Wikipedia contributors. Electromagnetism [Internet]. Wikipedia, The Free Encyclopedia.
7. Wikipedia contributors. Electromagnetic radiation [Internet]. Wikipedia, The Free Encyclopedia.

1.2: Wireless Technology and Amateur Radio - What and Why?

Learning Objectives

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

- Define wireless technology, and provide some examples of it.
- Define Amateur Radio, and provide some examples of how Amateur Radio is used in practice.

Wireless technology is equipment or methods that enable information to be transferred between two locations without an electrical wire, optical fiber, or other guiding medium [1]. Wireless equipment most commonly uses electromagnetic waves to send signals through air or empty space from a sender to a receiver. Contemporary examples of wireless technology include cellular telephones, broadcast and satellite televisions, broadcast AM/FM radio receivers, remote garage door openers, remote controls, cordless telephones and headphones, wireless computer mice and keyboards, Global Positioning System (GPS) receivers, handheld radios ("walkie-talkies"), Amateur Radio transceivers, and other devices that use wireless communication protocols like WiFi and Bluetooth [1] (Fig. 1.2.1). The development of wireless technology combined with the development of small, high-capacity batteries has led to the widespread adoption of wireless devices due to their convenience and portability.

Figure 1.2.1: Examples of wireless technology include (left to right) cellular phone [2], global positioning system (GPS) receiver [3], and wireless mouse and keyboard [4].

Amateur Radio, also known as **ham radio**, is one of several personal radio services which use wireless technology to enable licensed operators to communicate. As suggested by the name, Amateur Radio operators are restricted to using radio for non-commercial purposes [5]. In contrast to other personal radio services (see Table 1.2.1) which are typically limited to local use, amateur radio stations can often communicate with other stations around the world given their higher allowed power levels and wider access to the electromagnetic spectrum. To gain access to these privileges, Amateur Radio operators must typically be licensed by their local government agency (e.g., the Federal Communications Commission (FCC) [6] in the United States) and demonstrate some understanding of the key concepts of electricity and magnetism, electronics, radio equipment, radio wave propagation, radio-frequency safety, and local operating regulations and conventions. It is this overlap between the physics of electromagnetism and the topics of Amateur Radio that make it natural to consider both in the same context.

Table 1.2.1. Types of personal radio services in the United States [7]. Licenses are required to use the GMRS.

Service	Channels	Intended Use	Typical Range
Citizens Band	40	Private/Business	16 km (10 miles)
Marine VHF	50	Maritime	32 km (20 miles)
Family Radio Service (FRS) & General Mobile Radio Service (GMRS)	22	Personal	3.2 km (2 miles)
Multi-Use Radio Service (MURS)	5	Personal	8 km (5 miles)

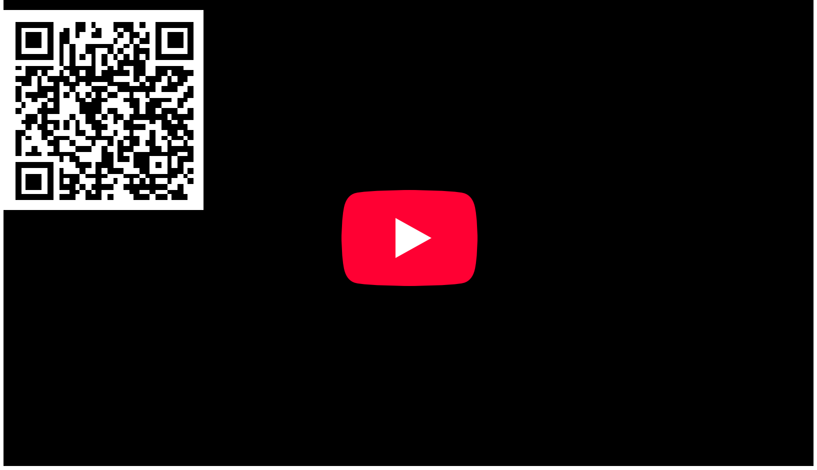
Amateur Radio operators have also developed interesting and novel ways to use radio signals (see Example 1.2.1).

Example 1.2.1

Amateur Radio operators have expanded the service and hobby into a wide variety of directions:

- **Wireless Experimentation & Technical Development**
  As wireless technology developed, Amateur Radio operators were involved in its progress. For example, in the United States, the first Amateur Radio licenses were issued in 1912 [7], not long after the first demonstration of radio communication in the 1890s and early 1900s [8]. Over time, Amateur Radio operators have worked to develop and refine techniques to send images, video, text messages, email, data packets, and other information formats via radio signals. They have also develop and test methods to successfully communicate messages under conditions with a lot of noise or with low power.
- **Emergency Communications**
  Amateur Radio operators are also known for their use of radio in emergency communications and other public service applications [9]. For example, Amateur Radio networks are often activated during hurricanes and other extreme weather events when other means of communications like cellular networks are down due to power outages. In the United States, programs like the American Radio Relay League's Amateur Radio Emergency Service (ARES) and the government's Radio Amateur Civil Emergency Service (RACES) prepare operators to work with local agencies in times of crisis and report severe weather conditions to the National Weather Services via the SKYWARN organization.
- **Radiosport (Contesting)**
  Amateur Radio operators have developed many different ways to practice their operating skills in competition with each other [10]. During contests, operators try to make as many two-way contacts with other operators within a fixed duration of time, often with different categories (e.g., operating teams, operating at low power, operating in different part of the spectrum). Formal or informal contests are also held to test operators' ability to receive or transmit quickly using Morse code, find hidden transmitters (Amateur Radio Direction Finding or "foxhunting"), or quickly deploy, contact, and move to a new location repeatedly (Rapid Deployment of Amateur Radio).
- **Remote Communication**
  Amateur Radio operators practice their transmitting and receiving skills by making contacts with many different geographic regions as possible [11]. Operators may try to work all countries, states, counties, or other entities from their own local station. Some operators take part in expeditions to set up stations in remote locations to make contacts ("DXpeditions" where "DX" is radio shorthand for "distant") or in specific kinds of geographic entities (e.g., Parks on the Air (POTA), Summits on the Air (SOTA), Islands on the Air (IOTA)). Other operators specialize in making long-distance contacts via orbiting satellites or via temporary conditions like aurora (Northern or Southern lights) or meteor showers that would otherwise be difficult to perform.
- **Radio Clubs**
  Amateur Radio operators have formed a variety of organizations to support Amateur Radio [12]. These clubs provide networking, formal or informal education about radio, licensing exams, "hamfests" to buy or sell radio gear, contesting events, field days, and other opportunities. Organizations have been developed to support operators in a common geographic area, with common background (e.g., Collegiate Amateur Radio Program (CARP) for college and university students, Youth on the Air (YOTA) for operators under 25 years old, Young Ladies Radio League for women of all ages) or common interests (e.g, Morse code, radiosport, remote communication, satellite communication). Many countries have national organizations (e.g., the American Radio Relay League, Radio Amateurs of Canada, Radio Society of Great Britain) that help to design and administer license exams, advocate for public policies to protect the Amateur Radio service, perform technical testing of radio equipment, and develop educational materials.
- **Citizen Science**
  While much more is known about radio waves than in the early 1900s, there is still much to be learned, particularly about how radio waves travel from location to location through the atmosphere and ionosphere surrounding the earth. Collaborations between the professional scientists & engineers at universities and government agencies and the broader community of citizen Amateur Radio operators can enable accelerated development of new radio technology and improved understanding of radio wave propagation. For example, the Ham Radio Science Citizen Investigation (HamSCI) [13] has studied how radio wave propagation changes during a total solar eclipse by crowd-sourcing contact reports from the Amateur Radio community.

These videos illustrate some examples of these various aspects of Amateur Radio.



References

1. Wikipedia contributors. Wireless [Internet]. Wikipedia, The Free Encyclopedia.
2. Freestocks.org. iPhone on a wooden bench [Internet]. Freestocks.org. (CC0)
3. Bird M. Turned on black gps monitor [Internet]. Pexels.com. (Free to use)
4. SHVETS production. Black entrepreneur typing on convenient wireless keyboard [Internet]. Pexels.com (Free to use)
5. Wikipedia contributors. Amateur radio [Internet]. Wikipedia, The Free Encyclopedia.
6. Federal Communications Commission. Amateur radio service [Internet]. Federal Communications Commission.
7. American Radio Relay League (ARRL). The ARRL ham radio license manual. Newington, CT: ARRL (2022).
8. Wikipedia contributors. Wireless telegraphy [Internet]. Wikipedia, The Free Encyclopedia.
9. American Radio Relay League (ARRL). Public service [Internet]. ARRL.
10. Wikipedia contributors. Radiosport [Internet]. Wikipedia, The Free Encyclopedia.
11. Wikipedia contributors. DXing [Internet]. Wikipedia, The Free Encyclopedia.
12. American Radio Relay League (ARRL). Clubs [Internet]. ARRL.
13. Ham Radio Science Citizen Investigation (HamSCI). HamSCI [Internet]. HamSCI.

1.3: Units

Learning Objectives

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

- Identify the meaning of commonly used prefixes for units.
- Identify SI units associated with electricity and magnetism and their abbreviations.

The term "unit" refers to the measure used to express a physical quantity. For example, the mean radius of the Earth is about 6,371,000 meters; in this case the unit is the meter.

A number like "6,371,000" becomes a bit cumbersome to write, so it is common to use a prefix to modify the unit. For example, the radius of the Earth is more commonly said to be 6371 kilometers, where one kilometer is understood to mean 1000 meters. It is common practice to use prefixes, such as "kilo-," that yield values in the range 0.001 to 10, 000. A list of standard prefixes appears in Table $1.3.1$.

Table $1.3.1$: Prefixes used to modify units.

Prefix	Multiply by:	Multiply by:
exa	E	10^{18}
peta	P	10^{15}
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kil	k	10^{3}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

Writing out the names of units can also become tedious. For this reason, it is common to use standard abbreviations, e.g., "6731 km" as opposed to "6371 kilometers," where "k" is the standard abbreviation for the prefix "kilo" and "m" is the standard abbreviation for "meter." Table $1.3.2$ shows a list of commonly used base units and their abbreviations.

Table $1.3.2$: Some units that are commonly used in electromagnetics.

Unit	Abbreviation	Quantifies:
ampere	A	electric current
coulomb	C	electric charge
farad	F	capacitance
henry	H	inductance
hertz	Hz	frequency
joule	J	energy
meter	m	distance

Unit	Abbreviation	Quantifies:
newton	N	force
ohm	Ω	resistance
second	s	time
tesla	T	magnetic flux density
volt	V	electric potential
watt	W	power
weber	Wb	magnetic flux

To avoid ambiguity, it is important to always indicate the units of a quantity, e.g., writing “6371 km” instead of “6371.” Failure to do so is a common source of error and misunderstanding. An example is the expression:

$$l = 3t$$

where l is length and t is time. It could be that l is in meters and t is in seconds, in which case “3” really means “3 m/s”. However, if it is intended that l is in kilometers and t is in hours, then “3” really means “3 km/h” and the equation is literally different. To patch this up, one might write “$l = 3t$ m/s”; however, note that this does not resolve the ambiguity we just identified – i.e., we still don’t know the units of the constant “3.” Alternatively, one might write “$l = 3t$ where l is in meters and t is in seconds,” which is unambiguous but becomes quite awkward for more complicated expressions. A better solution is to write instead:

$$l = (3\text{ m/s})t$$

or even better:

$$l = at$$

where $a = 3\text{ m/s}$ since this separates this issue of units from the perhaps more important fact that l is proportional to t and the constant of proportionality (a) is known.

The meter is the fundamental unit of length in the International System of Units, known by its French acronym “SI” and sometimes informally referred to as the “metric system.”

> In this work, we will use SI units exclusively.

Although SI is probably the most popular for engineering use overall, other systems remain in common use. For example, the English system, where the radius of the Earth might alternatively be said to be about 3959 miles, continues to be used in various applications and, to a lesser or greater extent, in various regions of the world. An alternative system commonly used in physics and material science applications is the CGS (“centimeter-gram-second”) system. The CGS system is similar to SI but with some significant differences. For example, the base unit of energy in the CGS system is not the “joule,” but rather the “erg,” and the values of some physical constants become unitless. Therefore – once again – it is very important to include units whenever values are stated.

SI defines seven fundamental units from which all other units can be derived. These fundamental units are distance in meters (m), time in seconds (s), current in amperes (A), mass in kilograms (kg), temperature in kelvin (K), particle count in moles (mol), and luminosity in candela (cd). SI units for electromagnetic quantities, such as coulombs (C) for charge and volts (V) for electric potential, are derived from these fundamental units.

A frequently overlooked feature of units is their ability to assist in error-checking mathematical expressions. For example, the electric field intensity may be specified in volts per meter (V/m), so an expression for the electric field intensity that yields units of V/m is said to be “dimensionally correct” (but not necessarily correct), whereas an expression that cannot be reduced to units of V/m *cannot* be correct.

Additional Reading

- Wikipedia contributors. International System of Units [Internet]. Wikipedia, The Free Encyclopedia.
- Wikipedia contributors. Centimetre–gram–second system of units [Internet]. Wikipedia, The Free Encyclopedia.

Technician Exam Questions

To become an Amateur Radio operator at the introductory level in the United States, a person must pass the Technician licensing exam. This exam consists of 35 multiple-choice questions. The questions are drawn at random from a pool of questions over a distribution of topics. Selected questions from the current pool (valid for 2022–26) are shown throughout the text.

Metric Prefixes

Relevant exam questions include: T5B01–08, T5B12–13, T5C07, T5C13.

Exercise 1.3.1

T5B01. How many milliamperes is 1.5 amperes?

A. 15 milliamperes

B. 150 milliamperes

C. 1500 milliamperes

D. 15,000 milliamperes

Answer

C. 1500 milliamperes. A current of 1 A is 1000 milliamperes, so 1.5 A is 1500 milliamperes.

[Beta Testing: Single Technician exam question from ADAPT (T5B01). It appears that you need to have an ADAPT account to access the question in the formative assignment.]

[Beta Testing: Multiple Technician exam questions from ADAPT (T5B01-08, T5B12-13, T5C07, T5C13). It appears that you need have an ADAPT account to access these questions in a formative assignment.]

Metric Units Formative Test

1.4: Electromagnetic Spectrum

Learning Objectives

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

- Define electromagnetic spectrum
- Identify ranges in the electromagnetic spectrum corresponding to different physical phenomena.
- Identify ranges in the radio-frequency spectrum and their applications.

Electromagnetic waves are oscillating disturbances in the electromagnetic field. The **frequency** of the wave is the number of oscillations per second and is given in units of Hertz (1 Hz = 1 oscillation per second). Electromagnetic fields exist at frequencies from 0 Hz to at least 10^{20} Hz – that's at least 20 orders of magnitude! At 0 Hz, electromagnetics consists of two distinct disciplines: **electrostatics**, concerned with electric fields, and **magnetostatics**, concerned with magnetic fields. At higher frequencies, electric and magnetic fields interact to form propagating waves. Waves having frequencies within certain ranges are given names based on how they manifest as physical phenomena. These names are (in order of increasing frequency) radio, infrared (IR), optical (also known as "light"), ultraviolet (UV), X-rays, and gamma rays (γ-rays). See Table 1.4.1 for the corresponding frequency ranges.

Definition: Electromagnetic Spectrum

The term **electromagnetic spectrum** refers to the various forms of electromagnetic phenomena that exist over the continuum of frequencies [1].

The speed (properly known as "phase velocity") at which electromagnetic fields propagate in free space is given the symbol c, and has the value of 299,792,458 m/s (exactly) or approximately 3.00×10^8 m/s. This value is often referred to as the "speed of light." While it is certainly the speed of light in free space, it is also the speed of *any* electromagnetic wave in free space. Given frequency f, wavelength is given by the expression

$$\underbrace{\lambda = \frac{c}{f}}_{\text{in free space}} \tag{1.4.1}$$

A sinusoidal wave repeats itself in space over a distance of the wavelength, as illustrated in Figure 1.4.1. According to Eq. 1.4.1, the wavelength and frequency are inversely related so that as one increases, the other decreases. To a good approximation, the wavelength in meters can then be calculated as 300 (in Mm) divided by the frequency in MHz.

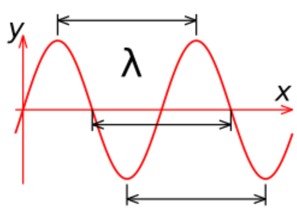


Figure 1.4.1: Wavelength λ for a sinusoidal wave. [2]

Table 1.4.1 shows the free space wavelengths associated with each of the regions of the electromagnetic spectrum.

Table 1.4.1: The electromagnetic spectrum. Note that the indicated ranges are arbitrary but consistent with common usage.

Regime	Frequency Range	Wavelength Range
γ-Ray	$> 3 \times 10^{19}$ Hz	< 0.01 nm
X-Ray	3×10^{16} Hz $-$ 3×10^{19} Hz	10–0.01 nm
Ultraviolet (UV)	2.5×10^{15} $-$ 3×10^{16} Hz	120–10 nm
Optical	4.3×10^{14} $-$ 2.5×10^{15} Hz	700–120 nm
Infrared (IR)	300 GHz $-$ 4.3×10^{14} Hz	1 mm – 700 nm

Regime	Frequency Range	Wavelength Range
Radio	$3\ \mathrm{kHz}-300\ \mathrm{GHz}$	100 km – 1 mm

This book presents a version of electromagnetic theory that is based on classical physics. This approach works well for most practical problems. However, at very high frequencies, wavelengths become small enough that quantum mechanical effects may be important. This effect usually happens in the X-ray band and above. In some applications, these effects become important at frequencies as low as the optical, IR, or radio bands. (A prime example is the **photoelectric effect**. [3]) Thus, caution is required when applying the classical version of the electromagnetic theory presented here, especially at these higher frequencies.

Caution

The theory presented in this book applies to static fields (0 Hz) along with radio, infrared, and light waves and, to a lesser extent, to ultraviolet waves, X-rays, and gamma rays. Certain phenomena in these frequency ranges – in particular, quantum mechanical effects – are not addressed in this book.

The **radio-frequency (RF)** portion of the electromagnetic spectrum alone spans 12 orders of magnitude in frequency (and wavelength), and so, not surprisingly, exhibits a broad range of phenomena (Figure $1.4.2$). For this reason, the radio spectrum is further subdivided into bands as shown in Table $1.4.2$ [4]. Also shown in Table $1.4.2$ are commonly used band identification acronyms and some typical applications. Amateur Radio operators can use portions of the radio bands starting as low as the Low Frequency (LF) band through the Extremely High Frequency (EHF) band and beyond, depending on their license class. However, in practice, activity for Amateur Radio tends to occur primarily on the High Frequency (HF), Very High Frequency (VHF), and Ultra High Frequency (UHF) bands.

Similarly, the optical band is partitioned into the familiar "rainbow" of red through violet, as shown in Figure $1.4.2$ and Table $1.4.3$. Other portions of the spectrum are sometimes similarly subdivided in certain applications.

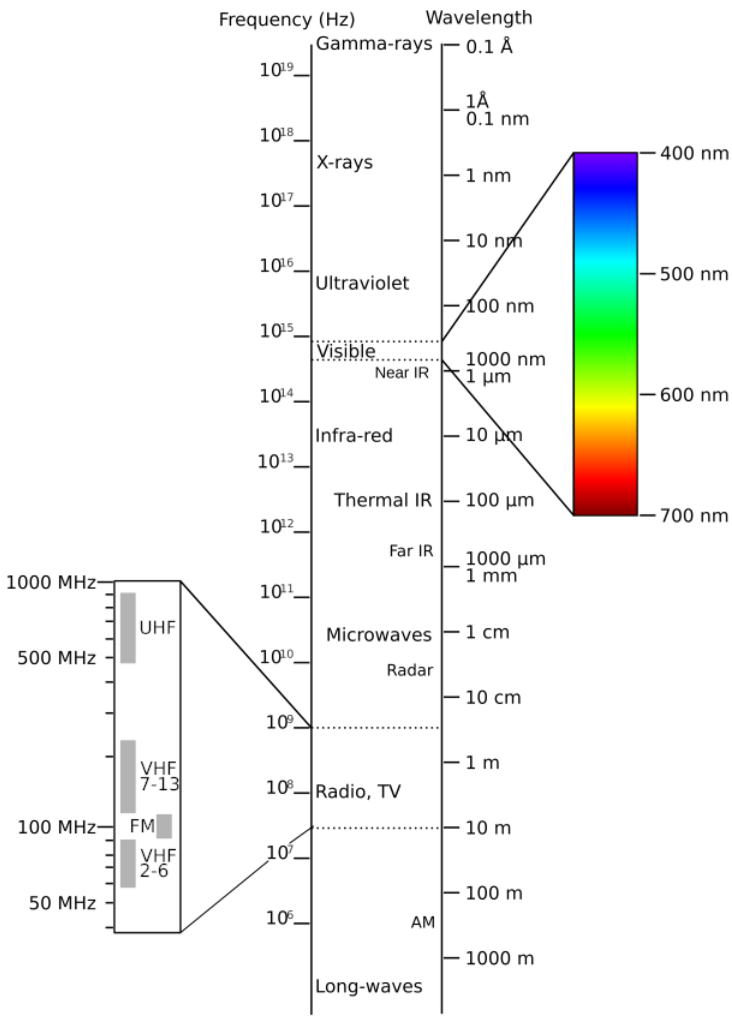


Figure 1.4.2: Electromagnetic Spectrum.

Table 1.4.2: The radio-frequency (RF) portion of the electromagnetic spectrum, according to a common scheme for naming ranges of radio frequencies from the International Telecommunication Union. WLAN: Wireless local area network, LMR: Land mobile radio, RFID: Radio frequency identification [5].

Band	Frequencies	Wavelengths	Typical Applications
THF (Tremendously High Frequency)	300–3000 GHz (0.3–3 THz)	1–0.1 mm	Short-range communication, Imaging, Spectroscopy
EHF (Extremely High Frequency)	30–300 GHz	10–1 mm	60 GHz WLAN, Point-to-point data links
SHF (Super High Frequency)	3–30 GHz	10–1 cm	Terrestrial & Satellite data links, Radar
UHF (Ultra High Frequency)	300–3000 MHz	1–0.1 m	TV broadcasting, Cellular, WLAN
VHF (Very High Frequency)	30–300 MHz	10–1 m	FM & TV broadcasting, LMR
HF (High Frequency)	3–30 MHz	100–10 m	Global terrestrial comm., CB Radio

Band	Frequencies	Wavelengths	Typical Applications
MF (Medium Frequency)	300–3000 kHz	1000–100 m	AM broadcasting
LF (Low Frequency)	30–300 kHz	10–1 km	Navigation, RFID
VLF (Very Low Frequency)	3–30 kHz	100–10 km	Navigation
ULF (Ultra Low Frequency)	300–3000 Hz	1000–100 km	Underground communication
SLF (Super Low Frequency)	30–300 Hz	10000–1000 km	Underwater communication
ELF (Extremely Low Frequency)	3–30 Hz	100000–10000 km	Underwater communication, Lightning

Table 1.4.3: The optical portion of the electromagnetic spectrum.

Band	Frequencies	Wavelengths
Violet	668–789 THz	450–380 nm
Blue	606–668 THz	495–450 nm
Green	526–606 THz	570–495 nm
Yellow	508–526 THz	590–570 nm
Orange	484–508 THz	620–590 nm
Red	400–484 THz	750–620 nm

References

1. Wikipedia contributors. Electromagnetic spectrum [Internet]. Wikipedia, The Free Encyclopedia.
2. Wikimedia Commons contributors. File:Sine wavelength.svg [Internet]. Wikimedia Commons. (CC BY-SA 3.0, Richard F. Lyon)
3. Wikipedia contributors. Photoelectric effect [Internet]. Wikipedia, The Free Encyclopedia.
4. Wikipedia contributors. Radio spectrum [Internet]. Wikipedia, The Free Encyclopedia.
5. Wikimedia Commons contributors. File:Electromagnetic-Spectrum.svg [Internet]. Wikimedia Commons. (CC BY-SA 3.0, V. Blacus)

Technician Exam Questions

The Radio Spectrum

Relevant exam questions include: T3B08–10, T5A06, T5A12, T5C06.

Wavelength

Relevant exam questions include: T3B04–07, T3B11.

1.5: Amateur Radio Equipment Basics

Learning Objectives

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

- Define transceiver, and its basic operation in an radio system.
- Define repeater, and explain why repeaters are used.

We will begin our discussion of radio equipment with a basic overview of the topic. The AM or FM broadcast radios that are common in everyday life are **receivers** only. In other words, they only receive the electromagnetic waves that are broadcast from the remote AM or FM radio station and then convert the received signal to sound. While this setup is reasonable for music or news, it lacks the interactivity that most Amateur Radio operators desire to make two-way contacts.

Transceivers

In contrast, radios that are able to both receive and transmit are called **transceivers**. The basic components of a transceiver are illustrated in Fig. 1.5.1. The operator communicating by voice can switch between receiving and transmitting, usually through a "push-to-talk" (PTT) button on a microphone. This button enables a transmit/receive switch so that the **transmitter** sends its electrical signals through a **feed line** out to the antenna. The **antenna** then converts the electrical signals in the radio to the electromagnetic waves that travel outwards into the air. When the button is released, the transmit-receive switch connects the antenna to the receiver, and any electromagnetic waves detected by the antenna are then converted into an electrical signal that can be turned into audio. The transceiver operates using **simplex** (also **half-duplex**) communication as it can simply receive or transmit using the same frequency but not both simultaneously.

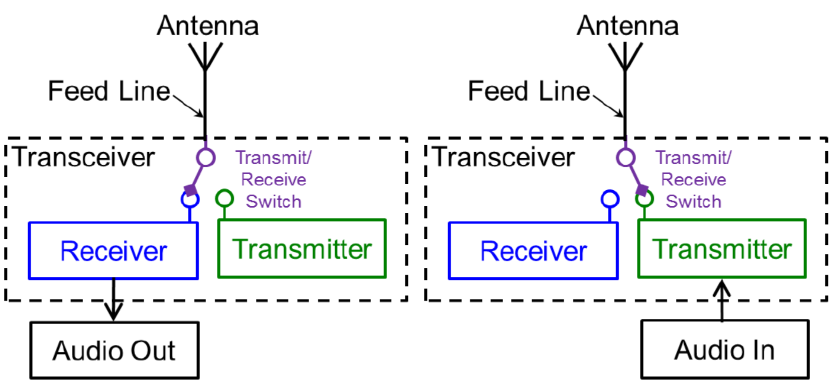


Figure 1.5.1: Block diagrams of a radio system using a transceiver. (left) The transceiver is set to receive. (right) The transceiver is set to transmit by depressing the "push to talk" (PTT) button. [1]

Figure 1.5.2 shows photos of actual transceivers useable for Amateur Radio. In both models, a "push-to-talk" button enables the operator to speak into the microphone and transmit the signal. When the button is released, the radios will return to receiving the signals from the connected antenna.

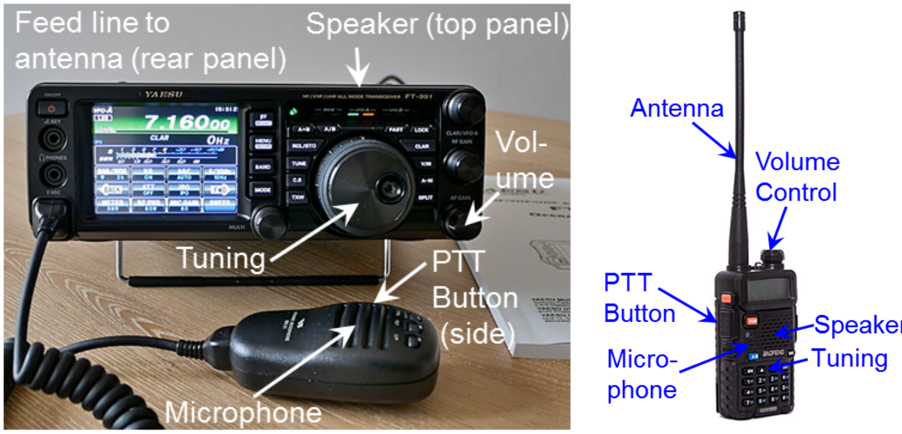


Figure 1.5.2: (left to right) Examples of an Amateur Radio base-station transceiver [2] and handheld transceiver [3]. For the base-station transceiver, the feed line connects to the back of the unit and is not shown. For the handheld transceiver, the feed line is just the connector from the antenna to the unit.

Repeaters

In the VHF and UHF range (above 30 MHz), the signals from the radio's antenna travel will usually travel primarily along their **line of sight** [4], which is as far as the wave can travel in a straight line away from the antenna and still directly interact with another antenna, as illustrated in Figure 1.5.3.

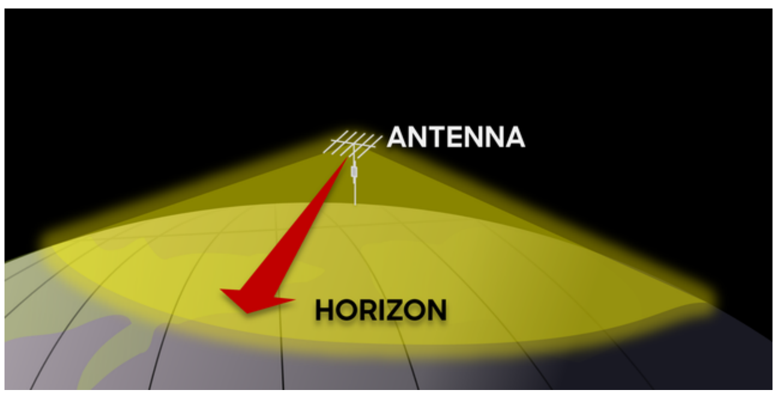


Figure 1.5.3: Illustration of the idealized range of "line of sight" propagation of electromagnetic waves. (Note that the antenna size is exaggerated to show the effect.) [5]

To extend the ability of a station to communicate beyond its "line of sight" range, **repeaters** are employed. As the name suggests, repeaters are special radios which have antennas that receive signals transmitted on a **channel** of a particular frequency, and then re-transmit (repeat) the incoming signal on a different channel of a second frequency that is offset slightly from the receiving frequency. Figure 1.5.4 shows a block diagram of the operation of a repeater.

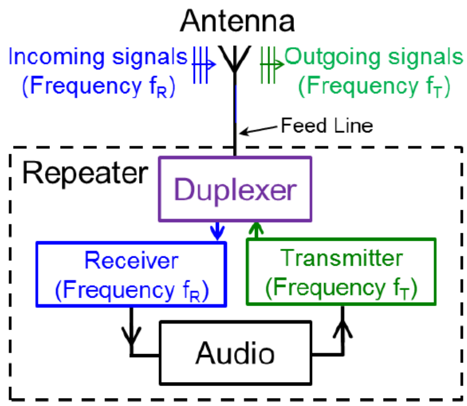


Figure 1.5.4: Block diagram of a repeater. Signals received by the antenna on a channel of one frequency are re-transmited on a second channel of a different frequency. [6]

A second station that is not within the range of the first station can then communicate with the first station using the repeater as an intermediary, as illustrated in Figure 1.5.5(B). In some cases, the communication may even be further extended using a network of multiple repeaters. As suggested by Fig. 1.5.5., it is advantageous to place repeaters in high locations like towers, tall building, hills, or mountains to maximize their range.

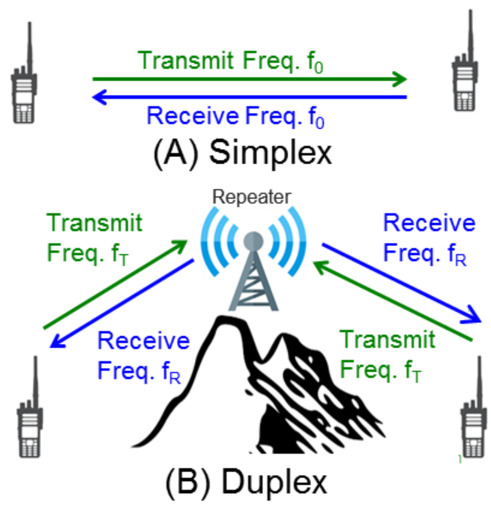


Figure 1.5.5: (A) Simplex communication occurs on the same frequency between the stations with each station either transmitting or receiving but not both at the same time. (B) Duplex communication uses a repeater as intermediate station with transmitting and receiving on different frequencies, possibly at the same time. [7]

Other examples of communication systems which use duplexing include wired and cellular telephones, although the method for the duplexing may be different than described above [8].

References

1. Kumon RE. Block diagram of a transceiver. (CC BY-SA 4.0)
2. Wikimedia Commons contributors. File:Yaesu FT991.jpg [Internet]. Wikimedia Commons. (Image: CC0 1.0; mick m43, All mode HF-6m-2m-70cm-AM-SSB-FM-C4FM Fusion Yaesu FT-991 transceiver, Annotation: CC BY-SA 4.0, Ronald E. Kumon)
3. Wikimedia Commons contributors. File:Baofeng UV-5R transceiver 5.jpg [Internet]. Wikimedia Commons. (Image: CC BY-SA 4.0; Barbara Kaminskaya, Dual band amateur VHF/UHF transceiver Baofeng UV-5R, Annotation: CC BY-SA 4.0, Ronald E. Kumon)
4. Wikipedia contributors. Line-of-sight propagation [Internet]. Wikipedia, The Free Encyclopedia.
5. Wikimedia Commons contributors. File:VHF propagation.png [Internet]. Wikimedia Commons. (CC BY-SA 4.0; Rudolf Davis Strazds, VHF radio wave propagation)
6. Kumon RE. Block diagram of a repeater. (CC BY-SA 4.0)

7. Wikimedia Commons contributors. File:Repeater Talk around channel.png [Internet]. Wikimedia Commons. (Original image: CC0 1.0; Goodtiming8871, Repeater Talk around channel diagram, Modifications: CC BY-SA 4.0; Ronald Kumon)
8. Wikipedia contributors. Duplex (telecommunications) [Internet]. Wikipedia, The Free Encyclopedia.

Technician Exam Questions

Relevant exam questions include: T7A02, T1F09.

1.6: Notation

Learning Objectives

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

- Identify the mathematical notation used in this book.

The list below describes the notation used in this book. If you have not seen some of this notation or terminology before, do not worry! The notation or terminology will be covered again in the body of the text.

- *Time*: The symbol t indicates time.
- *Position*: The symbols (x, y, z), (ρ, ϕ, z), and (r, θ, ϕ) indicate positions using the Cartesian, cylindrical, and polar coordinate systems, respectively. It is sometimes convenient to express the position coordinate in a manner that is independent of a coordinate system; in this case, we typically use the symbol $\vec{r}$. For example, $\vec{r} = x\hat{i} + y\hat{j} + z\hat{k}$ in the Cartesian coordinate system.
- *Vectors:* A symbol that is in boldface and/or with an arrow on top is used to indicate a vector; e.g., the electric field vector will typically appear as $\vec{E}$ or $\mathbf{E}$ or $\vec{\mathbf{E}}$. The scalar magnitudes are typically written in italics without an arrow on top, e.g., E can denote the magnitude of the electric field vector.
- *Unit vectors:* A circumflex ("hat" symbol) is used to indicate a unit vector; i.e., a vector having magnitude equal to one. For example, the unit vectors pointing in the $+x$, $+y$, $+z$ directions are indicated using the lowercase unit vectors $\hat{i}$, $\hat{j}$, $\hat{k}$, respectively. In discussion, the quantity "$\hat{i}$" is typically spoken "i hat." Note that the Cartesian unit vectors are sometimes also indicated by $\hat{x}$, $\hat{y}$, $\hat{z}$ for consistency of notation with unit vectors in other non-Cartesian coordinate systems.
- *Curves, surfaces, and volumes*: These geometrical entities will usually be indicated by single capital letter in italics or script; e.g., an open surface might be indicated as $\mathcal{S}$ and the curve bounding this surface might be indicated as $\mathcal{C}$. Similarly, the volume enclosed by a closed surface $\mathcal{S}$ may be indicated as $\mathcal{V}$.
- *Integrations over curves, surfaces, and volumes* will usually be indicated using a single integral sign with the appropriate subscript. For example:

$$\int_{\mathcal{C}} \cdots dl \text{ is an integral over the curve } \mathcal{C}$$

$$\int_{\mathcal{S}} \cdots ds \text{ is an integral over the surface } \mathcal{S}$$

$$\int_{\mathcal{V}} \cdots ds \text{ is an integral over the volume } \mathcal{V}.$$

- *Integrations over closed curves and surfaces* will be indicated using a circle superimposed on the integral sign. For example:

$$\oint_{\mathcal{C}} \ldots dl \text{ is an integral over the closed curve } \mathcal{C}$$

$$\oint_{\mathcal{S}} \ldots ds \text{ is an integral over the closed surface } \mathcal{S}$$

  A "closed curve" is one which forms an unbroken loop; e.g., a circle. A "closed surface" is one which encloses a volume with no openings; e.g., a sphere.
- *Phasors:* A tilde over a symbol is used to indicate a phasor quantity; e.g., a voltage phasor might be indicated as $\tilde{V}$, and the phasor representation of $\mathbf{E}$ will be indicated as $\tilde{\mathbf{E}}$.

- The symbol "$\cong$" means "approximately equal to." This symbol is used when equality exists, but is not being expressed with exact numerical precision. For example, the ratio of the circumference of a circle to its diameter is π, where $\pi \cong 3.14$.
- The symbol "$\approx$" also indicates "approximately equal to," but in this case the two quantities are unequal even if expressed with exact numerical precision. For example, $e^x = 1 + x + x^2/2 + \ldots$ as a infinite series, but $e^x \approx 1 + x$ for $x \ll 1$. Using this approximation $e^{0.1} \approx 1.1$, which is in good agreement with the actual value $e^{0.1} \cong 1.1052$.
- The symbol "$\sim$" indicates "on the order of," which is a relatively weak statement of equality indicating that the indicated quantity is within a factor of 10 or so the indicated value. For example, $\mu \sim 10^5$ for a class of iron alloys, with exact values being being larger or smaller by a factor of 5 or so.
- The symbol "$\triangleq$" means "is defined as" or "is equal as the result of a definition."
- Complex numbers: An italic $j \triangleq \sqrt{-1}$.

1.7: Coordinate Systems

Learning Objectives

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

- Identify the key differences between Cartesian, cylindrical, and spherical systems.

The coordinate systems most commonly used in scientific and engineering analysis are the **Cartesian** [1], **cylindrical** [2], and **spherical** [3] systems. These systems are illustrated in Figures $1.7.1$, $1.7.2$, and $1.7.3$, respectively.

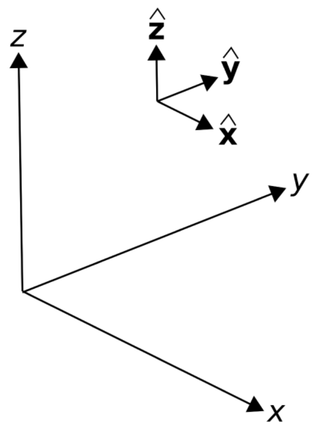


Figure $1.7.1$: Cartesian coordinate system [4]

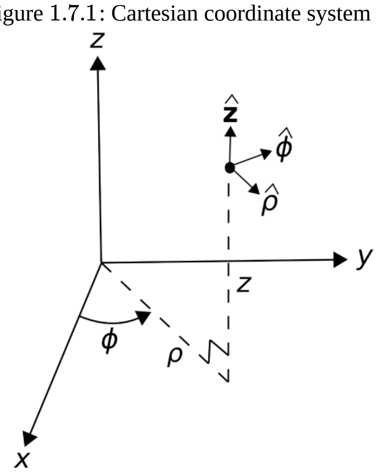


Figure $1.7.2$: Cylindrical coordinate system. [5]

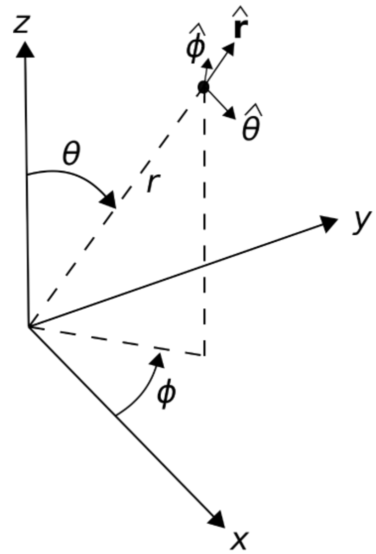


Figure 1.7.3: Spherical coordinate system. [6]

Note that the use of variables is not universal; in particular, it is common to encounter the use of r in lieu of ρ for the **radial coordinate** in the cylindrical system, and the use of R in lieu of r for the radial coordinate in the spherical system. The angle ϕ is typically called the **azimuthal coordinate** in both the cylindrical and spherical coordinates, while the angle θ is typically called the **polar coordinate** in spherical coordinates. In contrast to Cartesian coordinates, in which the coordinate unit vectors $\hat{x}$, $\hat{y}$, $\hat{z}$ always point in the same direction, the unit vectors in cylindrical and spherical coordinates will change direction based on the point in space that is being referenced, as indicated in the figures.

References

1. Wikipedia contributors. Cartesian coordinate system [Internet]. Wikipedia, The Free Encyclopedia.
2. Wikipedia contributors. Cylindrical coordinate system [Internet]. Wikipedia, The Free Encyclopedia.
3. Wikipedia contributors. Spherical coordinate system [Internet]. Wikipedia, The Free Encyclopedia.
4. Wikimedia Commons contributors. File:M0006 fCartesianBasis.svg [Internet]. Wikimedia Commons. (CC BY-SA 4.0; K. Kikkeri)
5. Wikimedia Commons contributors. File:M0096 fCylindricalCoordinates.svg [Internet]. Wikimedia Commons. (CC BY-SA 4.0; K. Kikkeri)
6. Wikimedia Commons contributors. File:Spherical Coordinate System.svg [Internet]. Wikimedia Commons. (CC BY-SA 4.0; K. Kikkeri)

1.8: Where Do We Go from Here?

Learning Objectives

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

- Recognize the organization and major topics of the chapters in the book
- Know where to find more detailed or advanced material in the book

Having placed the study of electricity & magnetism into the context of wireless technology, particularly Amateur Radio, we would now like to understand in more detail how the basic principles of electromagnetism can be used to develop such useful technology. This section outlines the rest of the book's chapters and provides some motivation for their content and arrangement. However, like the first chapter, the remainder of the book will continue to serve two purposes: (1) provide a conceptual and quantitative understanding of the physics and (2) explain how this physics can be used at a basic level to explain how radios and radio waves can be used for wireless applications. Along the way, we hope to cover much of the technical material that is included on the Amateur Radio Technician exam in a way that you not only learn the answers to the test questions but actually understand why they are correct.

The book is divided into two parts. If we liken the book's content to a good meal, then Chapter 1 is the appetizer, and Part I is the main course, including entrées and dessert! Part II then provides learners who are still hungry for more to indulge in some of the finer tastes of the theory and provides some generous second helpings of quantitative details.

Part I: Main Course

Part I consists of Chapters 1–13. These chapters introduce the fundamental principles and concepts of electricity and magnetism and some of their practical applications. They are best read in order. To begin the entrées, Chapter 2 formally introduces the idea of electric charge, electric field, and electric force. Chapter 3 takes a complementary perspective using an energy-based approach, including the concepts of electric potential energy and electric potential (voltage). Chapters 2 and 3 are a basic introduction to electrostatics, or in other words, electrical systems that are constant in time. Chapter 4 describes the relationship between the electric field and electric potential and also discusses some practical applications of the associated physical principles. Chapter 5 introduces the concept of moving charge, commonly called electric current, and the related concept of electric resistance in a simple electrical conduction model. Chapter 6 combines the ideas of potential, resistance, and voltage to analyze some basic direct-current (DC) circuits and discusses their applications. Resistors dissipate electrical energy, but other devices can store electrical energy. Chapter 7 introduces such a device, the capacitor, describes how it works, and discusses how to analyze some simple capacitor circuits and resistor-capacitor (R-C) circuits and their applications.

Chapter 8 then introduces magnetism, including the concept of the magnetic field, and describes how these fields are generated by moving charges or currents. In this chapter, all the magnetic fields are constant in time, providing a basic introduction to magnetostatics. Chapter 9 introduces the concept of electromagnetic induction. It starts to show the connection between magnetic and electric fields, describing how a changing magnetic field can induce an electric field and, therefore, when charges are present, an electric current. Chapter 10 introduces a new device, the inductor, which can store energy in a magnetic field. It also discusses how to analyze some simple resistor-inductor (R-L) circuits. It concludes with a discussion of inductor-capacitor (L-C) circuits and resistor-inductor-capacitor (R-L-C) circuits, which are examples of the simplest kinds of oscillatory circuits that can vary their electrical quantities regularly in time. Creating time-varying electrical circuits is essential for generating the signals at different frequencies needed for transceivers and repeaters.

We're now getting closer to our final course of the meal! Chapter 11 describes electromagnetic waves and their basic properties. But how are such waves generated in practice? Chapter 12 provides a basic answer to this question by describing how some basic antennas work, including how electrical signals are transmitted from radios to antennas using feed lines. Chapter 13 introduces concepts related to the propagation of electromagnetic waves, and discusses how they propagate in the Earth's environment.

Finally, we are reaching dessert. Chapter 14 introduces the more realistic conduction model known as band theory. This theory can explain the behavior of semiconductor materials, which are critical for the operation of modern solid-state electronic devices like electrical diodes, light-emitting diodes (LEDs), photovoltaic (solar) cells, and transistors. These devices are commonly used to create solid-state radios and supporting equipment.

Part II: Finer Tastes & Second Helpings

Part II consists of material that provides more detail or advanced approaches to topics introduced in Part I. Chapter 16 describes quantitative methods for calculating the electric field from a distribution of electric charges, thereby justifying some of the results introduced in Chapter 2 and should be accessible after covering that chapter. Chapter 17 describes how to use Gauss's Law to calculate the electric fields for charge distributions, which may be easier to understand after the discussion of flux in Chapter 11. Gauss's Law is a more mathematically sophisticated technique but turns out to be easier to deploy for some symmetric charge distributions. Chapter 18 describes quantitative methods for calculating the magnetic field from a distribution of currents, thereby justifying some of the results introduced in Chapter 7, which should be accessible after covering that chapter. It also discusses the use of Ampère's Law to more easily calculate magnetic fields for current distributions with certain symmetries. Lastly, the chapter discusses additional practical examples of magnetic fields, including electric motors, the Hall effect, and other applications. Chapter 19 introduces tools to analyze circuits with alternating-current (AC) circuits and discusses how electrical resonance can occur in these circuits. Chapter 20 provides a more detailed description of Maxwell's equations and their implications than described in Part I.

Chapter 21 provides a formal theory for the propagation of electromagnetic signals through transmission (feed) lines, extending the introduction provided in Chapter 13, which should be accessible after covering that chapter. It describes how coaxial cables operate, how to calculate the standing wave ratio, and how to minimize losses in feed lines. Chapter 22 provides a more formal and rigorous theory for generating and detecting electromagnetic waves and predicting the radiation patterns from simple antennas. It should also be accessible after covering Chapter 13. Finally, Chapter 23 goes beyond the physics into the important topic of how information is transmitted using radio waves through signal modulation, including amplitude modulation (AM), frequency modulation (FM), and other methods. This chapter is mostly independent of the other chapters in the book but may also be more complicated in some parts.

Consult with your instructor regarding your need to know the content of Part II, but also feel free to explore it yourself. Your banquet of electromagnetism awaits!

CHAPTER OVERVIEW

2: The Electric Field

2.1: Introduction

This chapter introduces a new physical property of matter: the **electric charge**. As you will see, charged objects can exert forces on other charged objects and even on objects that have no net charge. However, unlike many contact forces often studied in Newtonian mechanics (for example, pushing, pulling, friction, and the normal force), the forces exerted by electric charges can act at a distance from an object with no contact whatsoever! Moreover, the **electric force** can be attractive, like gravity, but also repulsive (Figure $2.1.1$). How is this possible?

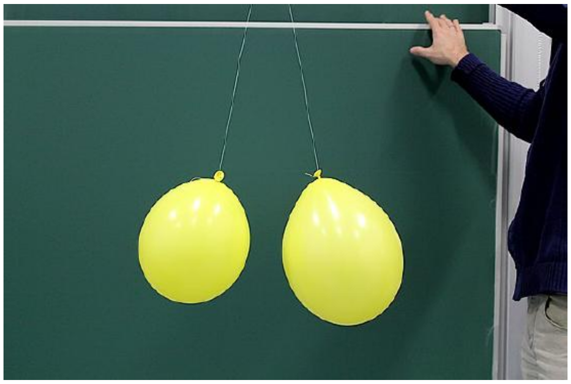


Figure $2.1.1$: Repulsive electric forces between two charged balloons. [1]

To explain this effect, we will introduce the concept of an **electric field**. The electric field surrounds the space around an electric charge and enables that charge to interact with other charges throughout space. We will learn how to calculate the electric field and its electric force on the surrounding charge. We will also learn ways to visualize and calculate the electric field for various common distributions of stationary electric charges. Some of these results will prove to be useful in later chapters (for example, the electric field in a parallel-plate capacitor).

References

1. Wikimedia Commons contributors. File:Repulsive-electric-force-between-balloons.jpg [Internet]. Wikimedia Commons. (CC BY-SA 4.0, MikeRun)

2.2: Electric Charge Model

Learning Objectives

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

- Describe the concept of electric charge
- Explain qualitatively the force electric charge creates

You are certainly familiar with electronic devices that you activate with the click of a switch, from computers to cell phones to television. And you have certainly seen electricity in a flash of lightning during a heavy thunderstorm. But you have also most likely experienced electrical effects in other ways, maybe without realizing that an electric force was involved. Let's take a look at some of these activities and see what we can learn from them about electric charges and forces.

Discoveries

You have probably experienced the phenomenon of **static electricity**: When you first take clothes out of a dryer, many (not all) of them tend to stick together; for some fabrics, they can be very difficult to separate. Another example occurs if you take a woolen sweater off quickly—you can feel (and hear) the static electricity pulling on your clothes and perhaps even your hair. If you comb your hair on a dry day and then put the comb close to a thin stream of water coming out of a faucet, you will find that the water stream bends toward (is attracted to) the comb (Figure $2.2.1$).

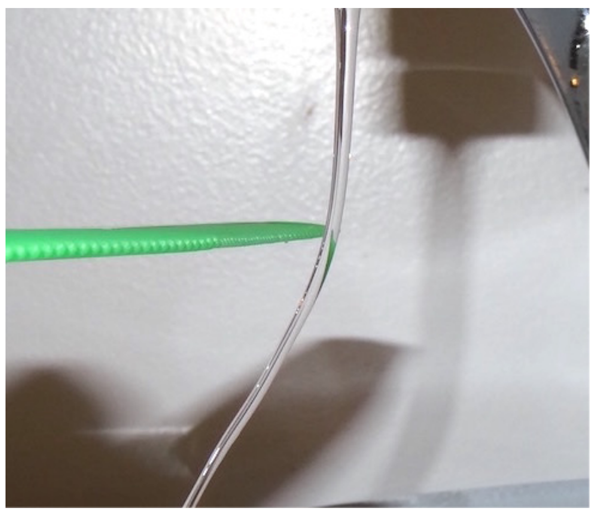


Figure $2.2.1$: An electrically charged comb attracts a stream of water from a distance. Note that the water is not touching the comb. (credit: Jane Whitney)

Suppose you bring the comb close to some small strips of paper; the strips of paper are attracted to the comb and even cling to it (Figure $2.2.2$). In the kitchen, if you quickly pull a length of plastic cling wrap off the roll, it tends to cling to nonmetallic materials such as plastic, glass, or food. If you rub a balloon on a wall for a few seconds, it sticks to the wall. Probably the most annoying effect of static electricity is getting shocked by a doorknob (or a friend) after shuffling your feet on some types of carpeting.

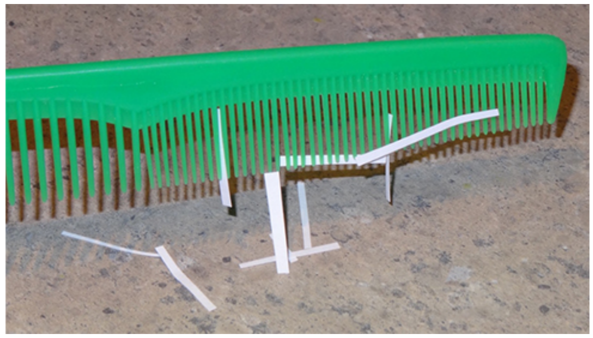


Figure 2.2.2: After being used to comb hair, this comb attracts small strips of paper from a distance, without physical contact. Investigation of this behavior helped lead to the concept of the electric force.

Many of these phenomena have been known for centuries. The ancient Greek philosopher Thales of Miletus (624–546 BCE) recorded that when amber (a hard, translucent, fossilized resin from extinct trees) was vigorously rubbed with a piece of fur, a force was created that caused the fur and the amber to be attracted to each other (Figure 2.2.3). Additionally, he found that the rubbed amber would not only attract the fur, and the fur attract the amber, but they both could affect other (nonmetallic) objects, even if not in contact with those objects (Figure 2.2.4).

Figure 2.2.3: Borneo amber is mined in Sabah, Malaysia, from shale-sandstone-mudstone veins. When a piece of amber is rubbed with a piece of fur, the amber gains more electrons, giving it a net negative charge. At the same time, the fur, having lost electrons, becomes positively charged. (credit: "Sebakoamber"/Wikimedia Commons)

The English physicist William Gilbert (1544–1603) also studied this attractive force, using various substances. He worked with amber, and, in addition, he experimented with rock crystal and various precious and semi-precious gemstones. He also experimented with several metals. He found that the metals never exhibited this force, whereas the minerals did. Moreover, although an electrified amber rod would attract a piece of fur, it would repel another electrified amber rod; similarly, two electrified pieces of fur would repel each other.

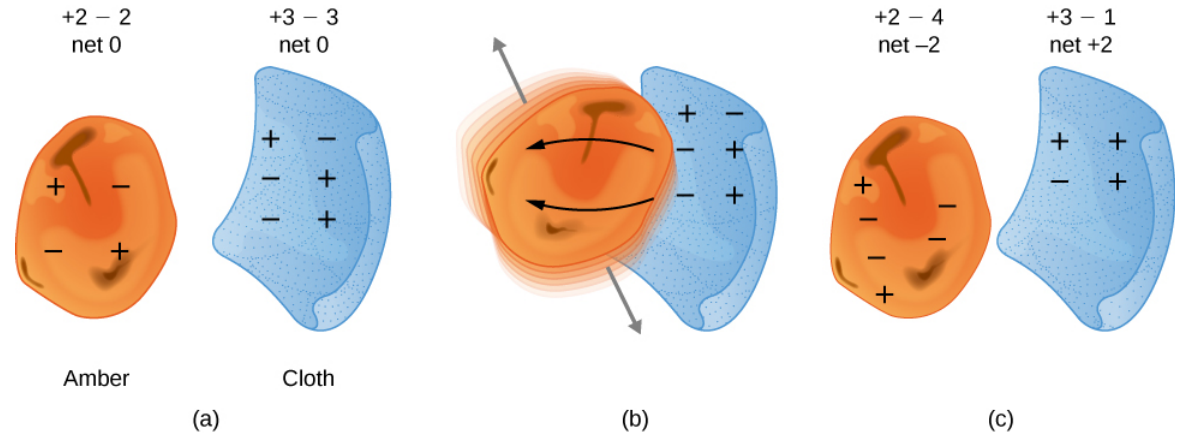


Figure 2.2.4: When materials are rubbed together, charges can be separated, particularly if one material has a greater affinity for electrons than another. (a) Both the amber and cloth are originally neutral, with equal positive and negative charges. Only a tiny fraction of the charges are involved, and only a few of them are shown here. (b) When rubbed together, some negative charge is transferred to the amber, leaving the cloth with a net positive charge. (c) When separated, the amber and cloth now have net charges, but the absolute value of the net positive and negative charges will be equal.

This suggested there were two types of an electric property; this property eventually came to be called **electric charge**. The difference between the two types of electric charge is in the directions of the electric forces that each type of charge causes: These forces are repulsive when the same type of charge exists on two interacting objects and attractive when the charges are of opposite types. The SI unit of electric charge is the **coulomb** (C), after the French physicist Charles Augustine de Coulomb (1736–1806).

The most peculiar aspect of this new force is that it does not require physical contact between the two objects in order to cause an acceleration. This is an example of a so-called "long-range" force. (Or, as James Clerk Maxwell later phrased it, "action at a distance.") With the exception of gravity, all other forces we have discussed so far act only when the two interacting objects actually touch.

The American physicist and statesman Benjamin Franklin found that he could concentrate charge in a "Leyden jar," which was essentially a glass jar with two sheets of metal foil, one inside and one outside, with the glass between them (Figure 2.2.4). This created a large electric force between the two foil sheets.

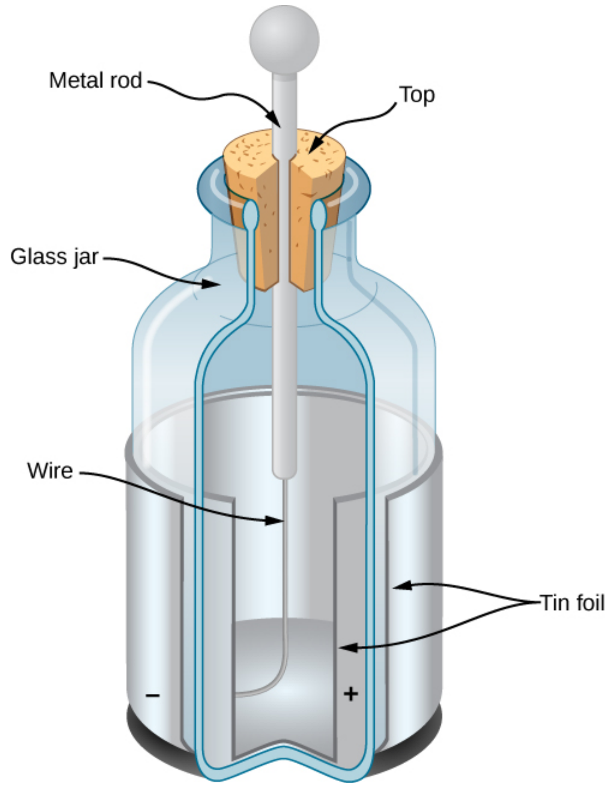


Figure 2.2.4: A Leyden jar (an early version of what is now called a capacitor) allowed experimenters to store large amounts of electric charge. Benjamin Franklin used such a jar to demonstrate that lightning behaved exactly like the electricity he got from the equipment in his laboratory.

Franklin pointed out that the observed behavior could be explained by supposing that one of the two types of charge remained motionless while the other type of charge flowed from one piece of foil to the other. He further suggested that an excess of what he called this "electrical fluid" be called "positive electricity" and its deficiency be called "negative electricity." With some minor modifications, his suggestion is the model we use today. (With the experiments that he could do, this was a pure guess; he had no way of actually determining the sign of the moving charge. Unfortunately, he guessed wrong; we now know that the charges that flow are the ones Franklin labeled negative, and the positive charges remain largely motionless. Fortunately, as we will see, it makes no practical or theoretical difference which choice we make, as long as we stay consistent with our choice.)

Let's list the specific observations that we have of this **electric force**:

- The force acts without physical contact between the two objects.
- The force can be either attractive or repulsive: If two interacting objects carry the same sign of charge, the force is repulsive; if the charges are of opposite sign, the force is attractive. These interactions are referred to as **electrostatic repulsion** and **electrostatic attraction**, respectively.
- Not all objects are affected by this force.
- The magnitude of the force decreases (rapidly) with increasing separation distance between the objects.

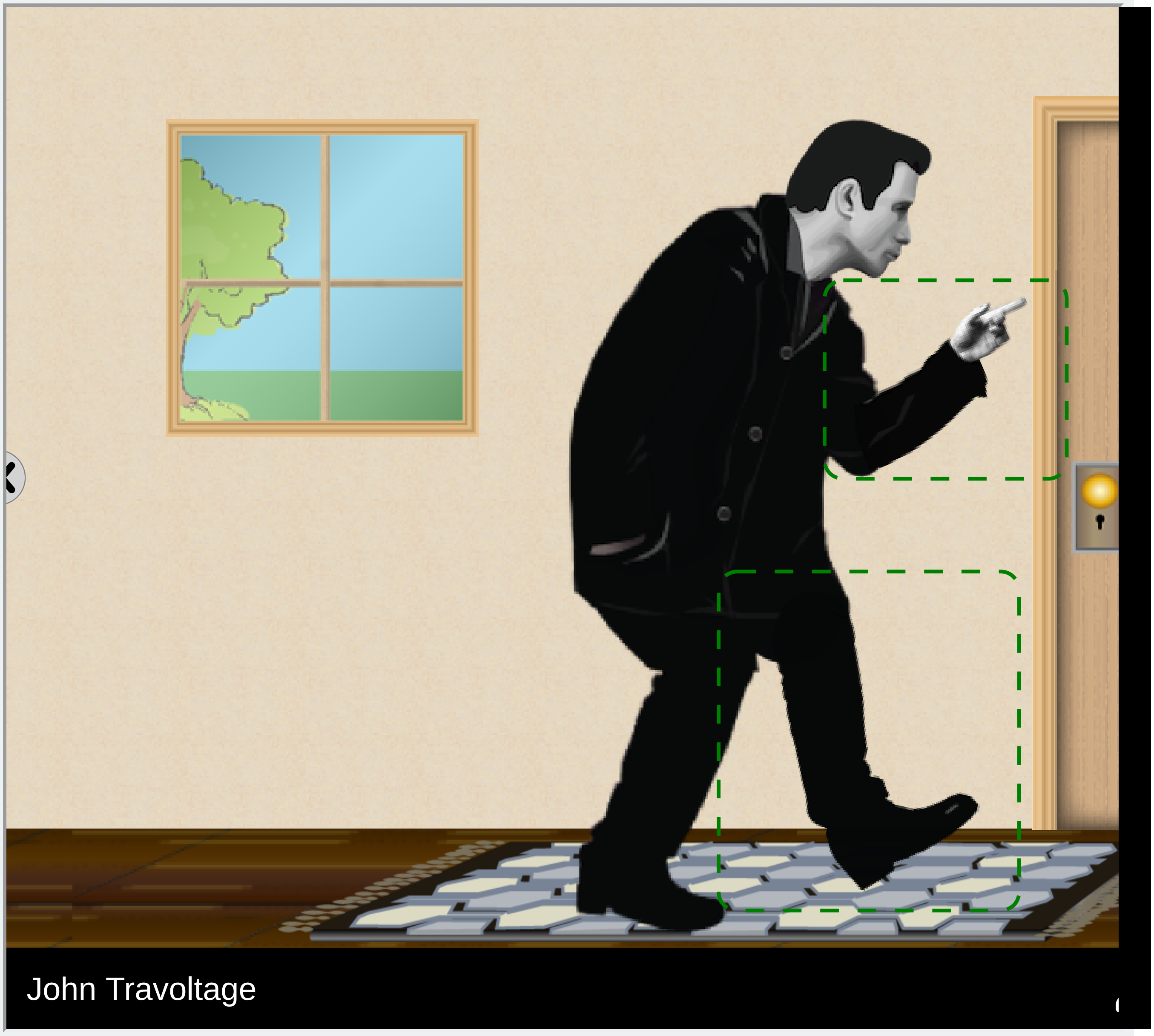
To be more precise, we find experimentally that the magnitude of the force decreases as the square of the distance between the two interacting objects increases. Thus, for example, when the distance between two interacting objects is doubled, the force between them decreases to one fourth what it was in the original system. We can also observe that the surroundings of the charged objects affect the magnitude of the force. However, we will explore this issue in a later chapter.

PhET Simulation: John Travoltage

Instructions:

1. Rub John's foot on the carpet a small amount. What happens? What is this effect called?
2. Rub John's foot on the carpet more, and then wait. What happens to the charge? Why is this happening?
3. Rub John's foot vigorously until something changes. What changed? Why did that happen?

4. Adjust the position of John's finger to be closer to the door knob, and then rub his foot against the carpet again. How did this adjustment change the results?



Source: John Travoltage

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

Properties of Electric Charge

In addition to the existence of two types of charge, several other properties of charge have been discovered.

- **Charge is quantized.** This means that electric charge comes in discrete amounts, and there is a smallest possible amount of charge that an object can have. In the SI system, this smallest amount is $e \equiv 1.602 \times 10^{-19}\ \mathrm{C}$. No free particle can have less charge than this, and, therefore, the charge on any object—the charge on all objects—must be an integer multiple of this amount. All macroscopic, charged objects have charge because electrons have either been added or taken away from them, resulting in a net charge.

- **The magnitude of the charge is independent of the type.** Phrased another way, the smallest possible positive charge (to four significant figures) is $+1.602 \times 10^{-19}$ C, and the smallest possible negative charge is -1.602×10^{-19} C; these values are exactly equal. This is simply how the laws of physics in our universe turned out.
- **Charge is conserved.** Charge can neither be created nor destroyed; it can only be transferred from place to place, from one object to another. Frequently, we speak of two charges "canceling"; this is verbal shorthand. It means that if two objects that have equal and opposite charges are physically close to each other, then the (oppositely directed) forces they apply on some other charged object cancel, for a net force of zero. It is important that you understand that the charges on the objects by no means disappear, however. The net charge of the universe is constant.
- **Charge is conserved in closed systems.** In principle, if a negative charge disappeared from your lab bench and reappeared on the Moon, conservation of charge would still hold. However, this never happens. If the total charge you have in your local system on your lab bench is changing, there will be a measurable flow of charge into or out of the system. Again, charges can and do move around, and their effects can and do cancel, but the net charge in your local environment (if closed) is conserved. The last two items are both referred to as the **law of conservation of charge**.

The Source of Charges: The Structure of the Atom

Once it became clear that all matter was composed of particles that came to be called atoms, it also quickly became clear that the constituents of the atom included both positively charged particles and negatively charged particles. The next question was, what are the physical properties of those electrically charged particles?

The negatively charged particle was the first one to be discovered. In 1897, the English physicist J. J. Thomson was studying what was then known as *cathode rays*. Some years before, the English physicist William Crookes had shown that these "rays" were negatively charged, but his experiments were unable to tell any more than that. (The fact that they carried a negative electric charge was strong evidence that these were not rays at all, but particles.) Thomson prepared a pure beam of these particles and sent them through crossed electric and magnetic fields, and adjusted the various field strengths until the net deflection of the beam was zero. With this experiment, he was able to determine the charge-to-mass ratio of the particle. This ratio showed that the mass of the particle was much smaller than that of any other previously known particle—1837 times smaller, in fact. Eventually, this particle came to be called the **electron**.

Since the atom as a whole is electrically neutral, the next question was to determine how the positive and negative charges are distributed within the atom. Thomson himself imagined that his electrons were embedded within a sort of positively charged paste, smeared out throughout the volume of the atom. However, in 1908, the New Zealand physicist Ernest Rutherford showed that the positive charges of the atom existed within a tiny core—called a nucleus—that took up only a very tiny fraction of the overall volume of the atom, but held over 99% of the mass (see the OpenStax chapter on Linear Momentum and Collisions.) In addition, he showed that the negatively charged electrons perpetually orbited about this nucleus, forming a sort of electrically charged cloud that surrounds the nucleus (Figure $2.2.5$). Rutherford concluded that the nucleus was constructed of small, massive particles that he named **protons**.

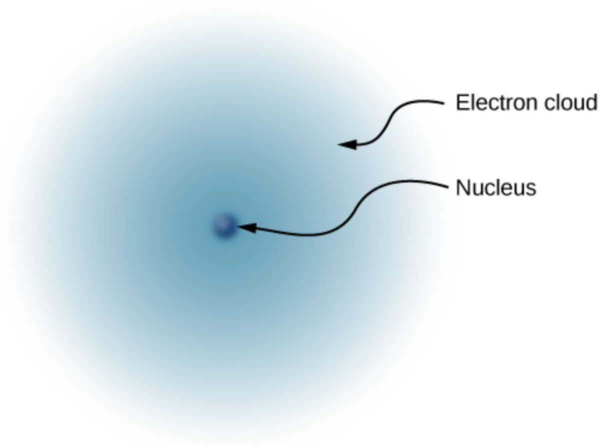


Figure 2.2.5: This simplified model of a hydrogen atom shows a positively charged nucleus (consisting, in the case of hydrogen, of a single proton), surrounded by an electron "cloud." The charge of the electron cloud is equal (and opposite in sign) to the charge of the nucleus, but the electron does not have a definite location in space; hence, its representation here is as a cloud. Normal macroscopic amounts of matter contain immense numbers of atoms and molecules, and, hence, even greater numbers of individual negative and positive charges.

Since it was known that different atoms have different masses and that ordinarily atoms are electrically neutral, it was natural to suppose that different atoms have different numbers of protons in their nucleus, with an equal number of negatively charged electrons orbiting about the positively charged nucleus, thus making the atoms overall electrically neutral. However, it was soon discovered that although the lightest atom, hydrogen, did indeed have a single proton as its nucleus, the next heaviest atom—helium—has twice the number of protons (two) but *four* times the mass of hydrogen.

This mystery was resolved in 1932 by the English physicist James Chadwick with the discovery of the **neutron**. The neutron is an electrically neutral twin of the proton, with no electric charge but (nearly) identical mass to the proton. The helium nucleus, therefore, has two neutrons along with its two protons. (Later experiments were to show that although the neutron is electrically neutral overall, it does have an internal charge structure. Furthermore, although the masses of the neutron and the proton are *nearly* equal, they aren't exactly equal: The neutron's mass is slightly larger than the proton's mass. That slight mass excess turned out to be of great importance. (For more information about that story, see the OpenStax chapter on Nuclear Physics.)

Thus, in 1932, the picture of the atom was of a small, massive nucleus constructed of a combination of protons and neutrons, surrounded by a collection of electrons whose combined motion formed a sort of negatively charged "cloud" around the nucleus (Figure 2.2.6). In an electrically neutral atom, the total negative charge of the collection of electrons is equal to the total positive charge in the nucleus. The very low-mass electrons can be more or less easily removed or added to an atom, changing the net charge on the atom (though without changing its type). An atom that has had the charge altered in this way is called an **ion**. Positive ions have had electrons removed, whereas negative ions have had excess electrons added. We also use this term to describe molecules that are not electrically neutral.

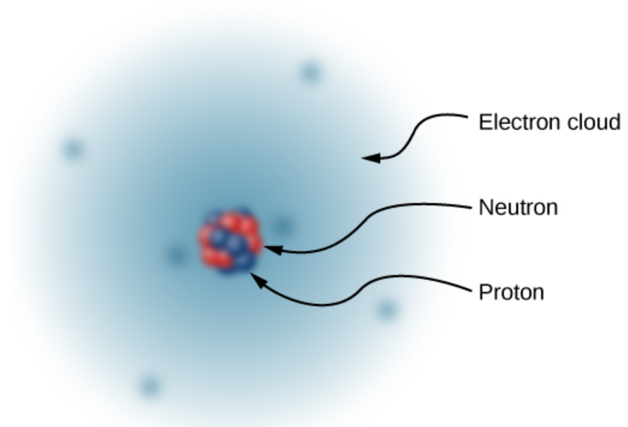


Figure 2.2.6: The nucleus of a carbon atom is composed of six protons and six neutrons. As in hydrogen, the surrounding six electrons do not have definite locations and so can be considered to be a sort of cloud surrounding the nucleus.

The story of the atom does not stop there, however. In the latter part of the twentieth century, many more subatomic particles were discovered in the nucleus of the atom: pions, neutrinos, and quarks, among others. With the exception of the photon, none of these particles are directly relevant to the study of electromagnetism, so they are omitted here. (For more information about the other particles, see the OpenStax chapter on particle physics (Particle physics and Cosmology).

A Note on Terminology

As noted previously, electric charge is a property that an object can have. This is similar to how an object can have a property that we call mass, a property that we call density, a property that we call temperature, and so on. Technically, we should always say something like, “Suppose we have a particle that carries a charge of $\mu\mathrm{C}$.” However, it is very common to say instead, “Suppose we have a $\mu\mathrm{C}$ charge.” Similarly, we often say something like, “Six charges are located at the vertices of a regular hexagon.” A charge is not a particle; rather, it is a **property** of a particle. Nevertheless, this terminology is extremely common (and is frequently used in this book, as it is everywhere else). So, keep in the back of your mind what we really mean when we refer to a “charge.”

2.3: Conduction and Charging

Learning Objectives

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

- Explain what a conductor is
- Explain what an insulator is
- List the differences and similarities between conductors and insulators
- Describe the process of charging by induction

In the preceding section, we said that scientists could create electric charge only on nonmetallic materials and never on metals. To understand why this is the case, you have to understand more about the nature and structure of atoms. In this section, we discuss how and why electric charges do—or do not—move through materials (Figure $2.3.1$). A more complete description is given in a later chapter.

Figure $2.3.1$: This power adapter uses metal wires and connectors to conduct electricity from the wall socket to a laptop computer. The conducting wires allow electrons to move freely through the cables, which are shielded by rubber and plastic. These materials act as insulators that don't allow electric charge to escape outward. (credit: modification of work by "Evan-Amos"/Wikimedia Commons)

Conductors and Insulators

As discussed in the previous section, electrons surround the tiny nucleus in the form of a (comparatively) vast cloud of negative charge. However, this cloud does have a definite structure to it. Let's consider an atom of the most commonly used conductor, copper.

In some atoms, an outermost electron is only loosely bound to the atom's nucleus. It can be easily dislodged; it then moves to a neighboring atom. In a large mass of copper atoms (such as a copper wire or a sheet of copper), these vast numbers of outermost electrons (one per atom) wander from atom to atom, and are the electrons that do the moving when electricity flows. These wandering, or "free," electrons are called **conduction electrons**, and copper is therefore an excellent **conductor** (of electric charge). All conducting elements have a similar arrangement of their electrons, with one or two conduction electrons. This includes most metals. (For a deeper understanding about how electrons are arranged in atoms, see the OpenStax chapter on Atomic Structure.)

Insulators, in contrast, are made from materials that lack conduction electrons; charge flows only with great difficulty, if at all. Even if excess charge is added to an insulating material, it cannot move, remaining indefinitely in place. This is why insulating materials exhibit the electrical attraction and repulsion forces described earlier, whereas conductors do not; any excess charge placed on a conductor would instantly flow away (due to mutual repulsion from existing charges), leaving no excess charge around to create forces. Charge cannot flow along or through an **insulator**, so its electric forces remain for long periods of time. (Charge will dissipate from an insulator, given enough time.) As it happens, amber, fur, and most semi-precious gems are insulators, as are materials like wood, glass, and plastic.

Charging by Induction

Let's examine in more detail what happens in a conductor when an electrically charged object is brought close to it. As mentioned, the conduction electrons in the conductor are able to move with nearly complete freedom. As a result, when a charged insulator (such as a positively charged glass rod) is brought close to the conductor, the (total) charge on the insulator exerts an electric force on the conduction electrons. Since the rod is positively charged, the conduction electrons (which themselves are negatively charged) are attracted, flowing toward the insulator to the near side of the conductor (Figure $2.3.2$).

Now, the conductor is still overall electrically neutral; the conduction electrons have changed position, but they are still in the conducting material. However, the conductor now has a charge *distribution*; the near end (the portion of the conductor closest to the insulator) now has more negative charge than positive charge, and the reverse is true of the end farthest from the insulator. The relocation of negative charges to the near side of the conductor results in an overall positive charge in the part of the conductor farthest from the insulator. We have thus created an electric charge distribution where one did not exist before. This process is referred to as *inducing polarization*—in this case, polarizing the conductor. The resulting separation of positive and negative charge is called **polarization**, and a material, or even a molecule, that exhibits polarization is said to be polarized. A similar situation occurs with a negatively charged insulator, but the resulting polarization is in the opposite direction.

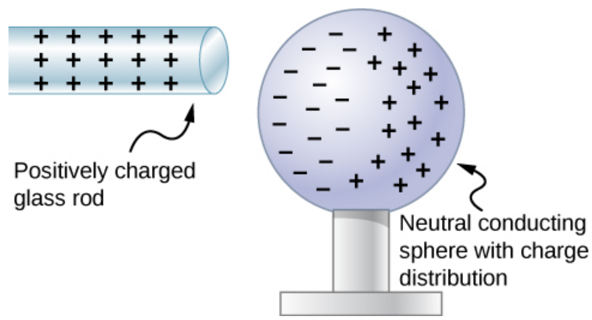


Figure $2.3.2$: Induced polarization. A positively charged glass rod is brought near the left side of the conducting sphere, attracting negative charge and leaving the other side of the sphere positively charged. Although the sphere is overall still electrically neutral, it now has a charge distribution, so it can exert an electric force on other nearby charges. Furthermore, the distribution is such that it will be attracted to the glass rod.

The result is the formation of what is called an **electric dipole**, from a Latin phrase meaning "two ends." The presence of electric charges on the insulator—and the electric forces they apply to the conduction electrons—creates, or "induces," the dipole in the conductor. (See the section on Electric Dipoles for more information.)

Neutral objects can be attracted to any charged object. The pieces of straw attracted to polished amber are neutral, for example. If you run a plastic comb through your hair, the charged comb can pick up neutral pieces of paper. Figure $2.3.3$ shows how the polarization of atoms and molecules in neutral objects results in their attraction to a charged object.

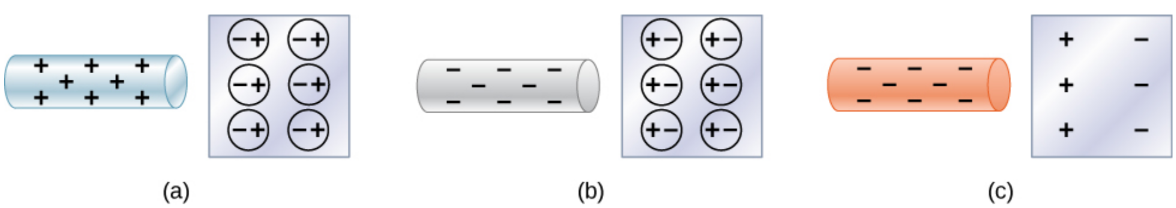


Figure $2.3.3$: Both positive and negative objects attract a neutral object by polarizing its molecules. (a) A positive object brought near a neutral insulator polarizes its molecules. There is a slight shift in the distribution of the electrons orbiting the molecule, with unlike charges being brought nearer and like charges moved away. Since the electrostatic force decreases with distance, there is a net attraction. (b) A negative object produces the opposite polarization, but again attracts the neutral object. (c) The same effect occurs for a conductor; since the unlike charges are closer, there is a net attraction.

When a charged rod is brought near a neutral substance, an insulator in this case, the distribution of charge in atoms and molecules is shifted slightly. Opposite charge is attracted nearer the external charged rod, while like charge is repelled. Since the electrostatic force decreases with distance, the repulsion of like charges is weaker than the attraction of unlike charges, and so there is a net attraction. Thus, a positively charged glass rod attracts neutral pieces of paper, as will a negatively charged rubber rod. Some molecules, like water, are polar molecules. Polar molecules have a natural or inherent separation of charge, although they are neutral overall. Polar molecules are particularly affected by other charged objects and show greater polarization effects than molecules with naturally uniform charge distributions.

When the two ends of a dipole can be separated, this method of **charging by induction** may be used to create charged objects without transferring charge. In Figure $2.3.4$, we see two neutral metal spheres in contact with one another but insulated from the rest of the world. A positively charged rod is brought near one of them, attracting negative charge to that side, leaving the other sphere positively charged.

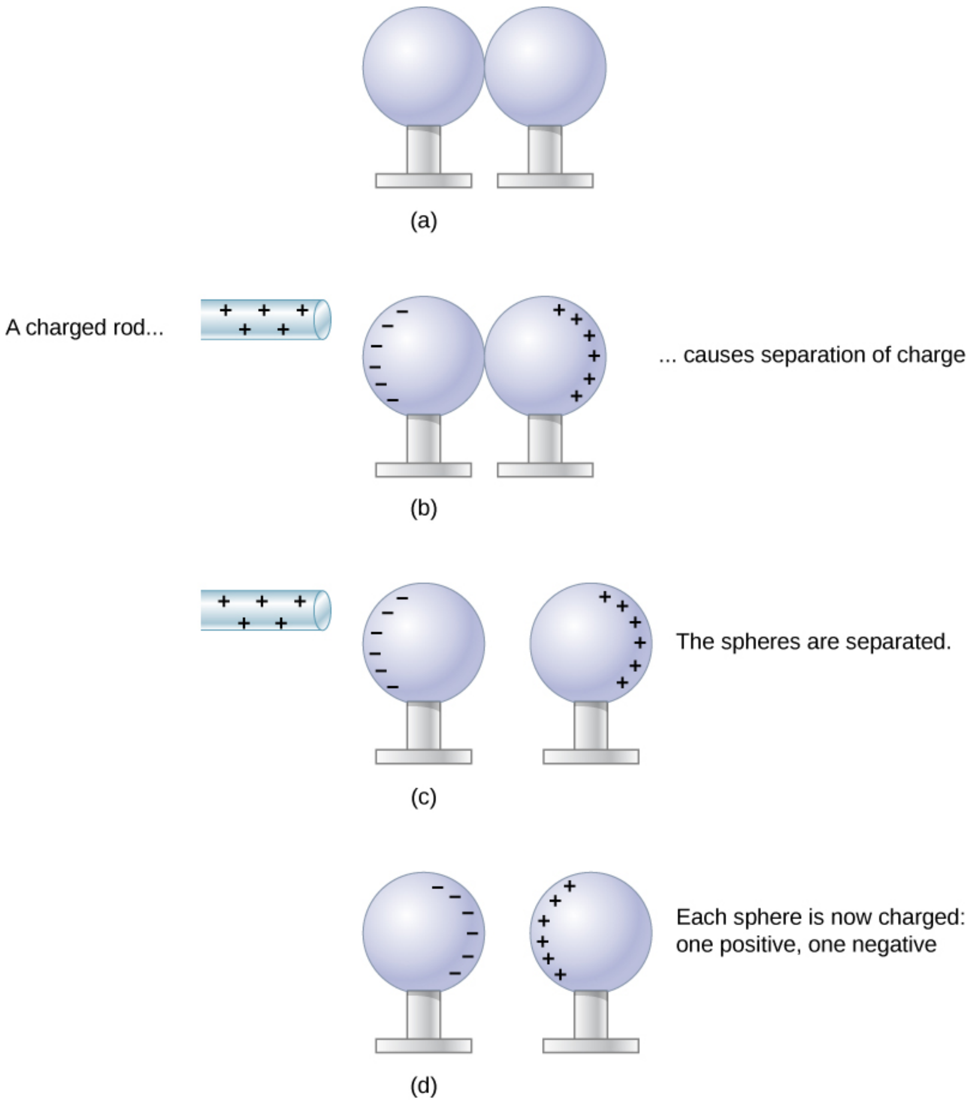


Figure $2.3.4$: Charging by induction. (a) Two uncharged or neutral metal spheres are in contact with each other but insulated from the rest of the world. (b) A positively charged glass rod is brought near the sphere on the left, attracting negative charge and leaving the other sphere positively charged. (c) The spheres are separated before the rod is removed, thus separating negative and positive charges. (d) The spheres retain net charges after the inducing rod is removed—without ever having been touched by a charged object.

Another method of charging by induction is shown in Figure $2.3.5$. The neutral metal sphere is polarized when a charged rod is brought near it. The sphere is then grounded, meaning that a conducting wire is run from the sphere to the ground. Since Earth is large and most of the ground is a good conductor, it can supply or accept excess charge easily. In this case, electrons are attracted to the sphere through a wire called the ground wire, because it supplies a conducting path to the ground. The ground connection is broken before the charged rod is removed, leaving the sphere with an excess charge opposite to that of the rod. Again, an opposite charge is achieved when charging by induction, and the charged rod loses none of its excess charge.

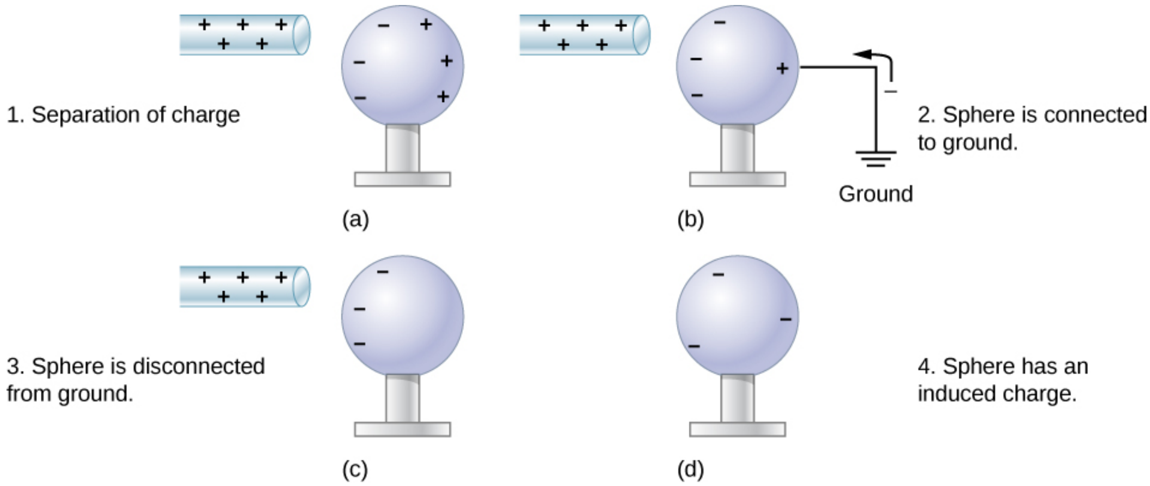


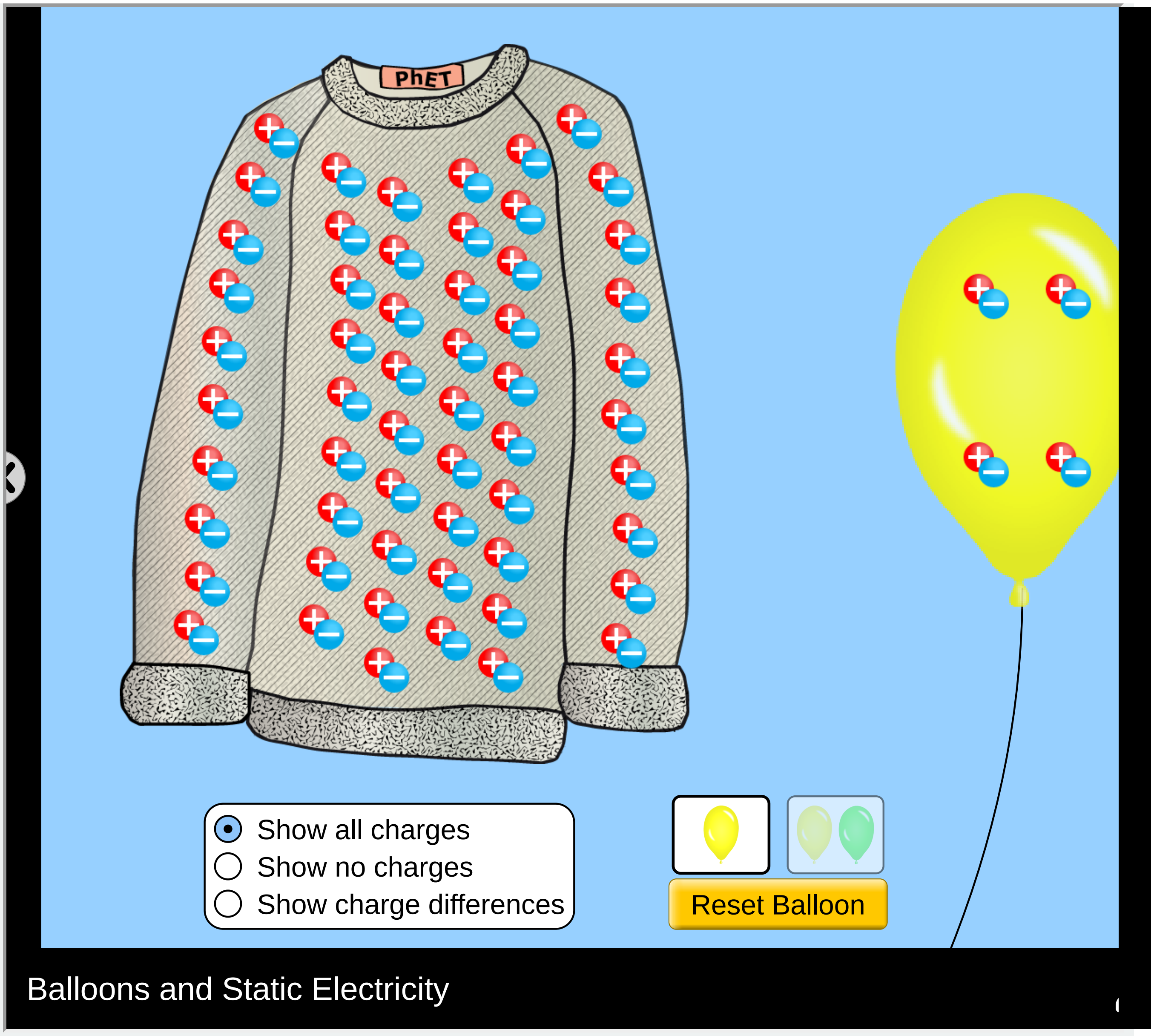
Figure 2.3.5: Charging by induction using a ground connection. (a) A positively charged rod is brought near a neutral metal sphere, polarizing it. (b) The sphere is grounded, allowing electrons to be attracted from Earth's ample supply. (c) The ground connection is broken. (d) The positive rod is removed, leaving the sphere with an induced negative charge.

The following computer simulation will allow you to explore charging, discharging, and charging by induction.

PhET Simulation: Balloons and Static Electricity

Instructions:

1. Rub the balloon on the sweater. What happens? What is this effect called?
2. Pull the balloon a short distance from the sweater, and then let it go. What happens? Why does this happen?
3. Move the balloon close to the wall but not touching, and then let it go. What happens? Why does this happen? What is this effect called?
4. Click on the button to add a second balloon to the system. Explore the various situations of (1) charged and uncharged balloons and (2) with and without the wall.
5. Feel free to adjust the settings of the simulation to "Show no charges" or "Show charge differences" to get a different (and perhaps more realistic view) of the system.



Source: Baloons and Static Electricity

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

2.4: Electric Fields and Forces

LEARNING OBJECTIVES

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

- Explain the purpose of the electric field concept
- Describe the properties of the electric field
- Calculate the electric force that charges exert on each other
- Determine the direction of the electric force for different source charges

Having seen the qualitative effects of charged objects on other charged objects, we would also like to be able to calculate quantities related to these effects. In this section, we will begin to define our approach to these calculations.

Defining a Field

We have seen experimentally that electric charges can have effects over a distance without contact. It is then natural to hypothesize that a physical field around a charged particle exists that can exert force on other charged particles. Mathematically, a physical field is a function that gives a value of a physical quantity at each point in space and each moment in time (see Example $2.4.1$ and Example $2.4.2$).

Example $2.4.1$

In weather reports, a temperature field can be visualized on a map showing the temperature's value at each point on the map at a specific moment in time. Because each temperature is a number, this field is known as a **scalar field**.

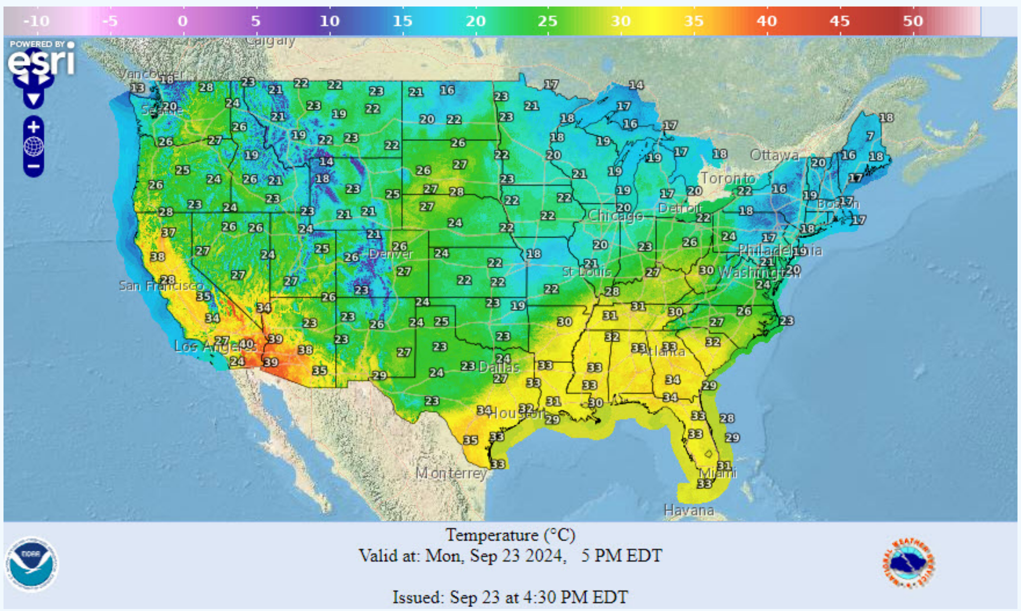


Figure $2.4.1$: Temperature field visualized in a map format. In the image, the values of temperature are given in degrees Celsius and are also color-coded. From the map, you can see that it is warmest in the southwestern states and coolest in the northern states and western mountainous regions. [1]

Example $2.4.2$

Weather reports will also sometimes give a wind velocity field. In this case, the map shows the magnitude and direction of the wind at each point on the map at a specific moment in time. Because the wind velocity is a vector, this field is known as a vector field.

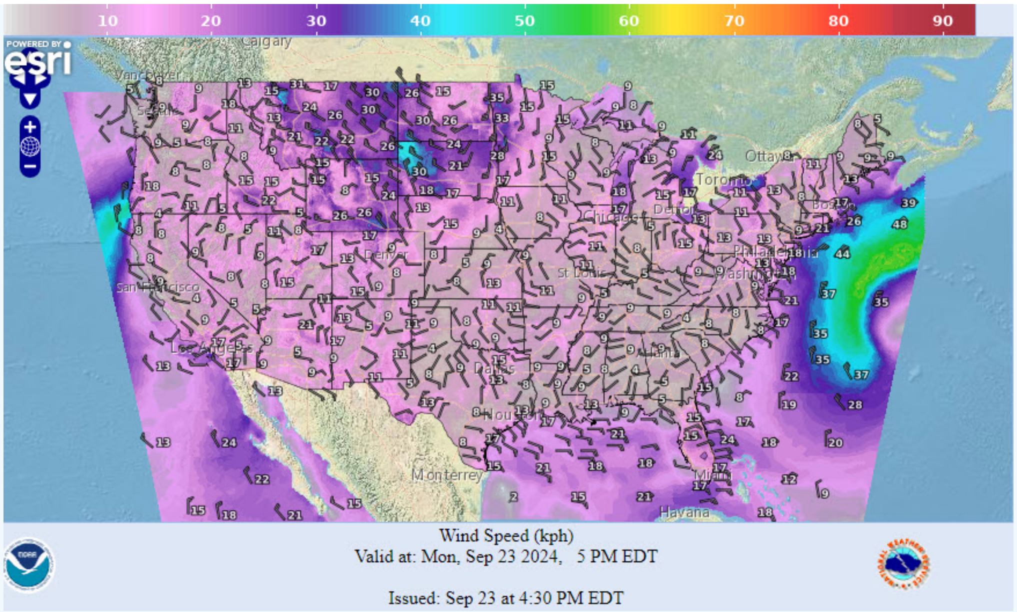


Figure 2.4.2: Wind velocity field visualized in a map format. In this map, the wind direction is given by wind barbs that point to the direction from which the wind is coming, while the speed is given by the color coding in kilometers per hour (1 kph = 1 km/h). The map shows it is windiest in the north-central United States and the Atlantic Ocean off of the East Coast. [2]

Defining the Electric Field

Because we expect the field around a charge to exert a force, which is a vector quantity, we will hypothesize that the electric field is a vector field. For simplicity, we will start by considering the source of the field to be a single **point charge**, which we define as an idealized particle at a single point in space at coordinates (x_s, y_s, z_s) carrying a specified **source charge** q. Based on experimental evidence, we define the electric field at **test location** (x_t, y_t, z_t) away from the source charge to be

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r}, \tag{2.4.1}$$

where ϵ_0 is the **permittivity of free space** (vacuum), r is the **separation distance** between the source charge and test location, and $\hat{r}$ is the **radial unit vector** that points in the direction from the source charge to the test location. From geometry, the distance between the source charge and test location is given by

$$r = [(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2]^{1/2}, \tag{2.4.2}$$

where $\Delta x = x_t - x_s$, $\Delta y = y_t - t = y_s$, $\Delta z = z_t - z_s$ are the differences in coordinates between the test and source charges. Note that the Also, from geometry, the radial unit vector is

$$\hat{r} = \frac{(\Delta x)\hat{i} + (\Delta y)\hat{j} + (\Delta z)\hat{k}}{[(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2]^{1/2}} \tag{2.4.3}$$

The radial unit vector in Equation 2.4.1 ensures that the electric field is always directed along the line between the source charge and the test location. The numerator sets the proper direction, while the denominator ensures that the length of the unit vector is one.

Direction of the Radial Unit Vector

The radial unit vector always points away from the source charge and toward the test location. You can remember this relation from the "**rst**" memory aid:

r-hat points from
source to
test

This is true regardless of the sign of the source charge, as Equation 2.4.3 does not contain any charge terms.

A consequence of Equation 2.4.1 is that the direction of the electric field is determined by the sign of the source charge. If the source charge is positive, the electric field points in the same direction as the radial unit vector (outward). If the source charge is negative, the electric field points in the opposite direction as the radial unit vector (inward).

Direction of the Electric Field

By convention, all electric fields $\vec{E}$ point away from positive source charges and point toward negative source charges.

It is important to emphasize that the electric field is not constant around a point charge but instead falls off as the reciprocal of the square of the separation distance. In addition, if either the source charge's location or the test location moves, then r and $\hat{r}$ change, and therefore, so does the electric field.

The permittivity of free space has a very important physical meaning that we will discuss in a later chapter; for now, it is simply an empirical proportionality constant. Its numerical value (to three significant figures) turns out to be

$$\epsilon_0 = 8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N} \cdot \mathrm{m}^2}. \tag{2.4.4}$$

These units are required to give the electric field in the correct units of newtons per coulomb. For convenience, we often define the **electrostatic constant** (also known as Coulomb's constant) to be

$$k = \frac{1}{4\pi\epsilon_0} = 8.99 \times 10^9 \frac{\mathrm{N} \cdot \mathrm{m}^2}{\mathrm{C}^2}. \tag{2.4.5}$$

A consequence of these constants being given in units of meters and coulombs is that the charge and separation distance in Equation 2.4.1 must also be given those same units to ensure that the force is in units of newtons.

Strictly speaking, Equation 2.4.1 is only valid for a point charge. However, it will still usually be a good approximation for the electric field of a uniformly charged particle at distances that are large compared to the size of the particle. Example 2.4.3 shows how you can calculate electric fields and forces for a problem with two point charges.

Example 2.4.3

Electric Field Due to Point Charges

(a) Suppose a point charge of +5.00nC is located at Point P_1 with coordinates $(-2.00\text{ m}, +2.00\text{ m}, 0.00\text{ m})$. What is the electric field at Point P_2 $(+2.00\text{ m}, +1.00\text{ m}, 0.00\text{ m})$?

(b) What is the strength (magnitude) of the electric field at $(+2.00\text{ m}, +1.00\text{ m}, 0.00\text{ m})$?

Solution

(a) We will follow a four-step problem-solving strategy:

PLAN We have a point charge, so we plan to use Eq. 2.4.1 to find the electric field.

SKETCH However, before we start to calculate, it is important to make a drawing of the geometry of the problem, marking the source charge at Point P_1 and the test location at Point P_2. We can also sketch in the direction of the radial unit vector because we know it lines along the line connecting Points P_1 and P_2 and points away from the source charge and toward the test location. We will later use the sketch to check our calculations.

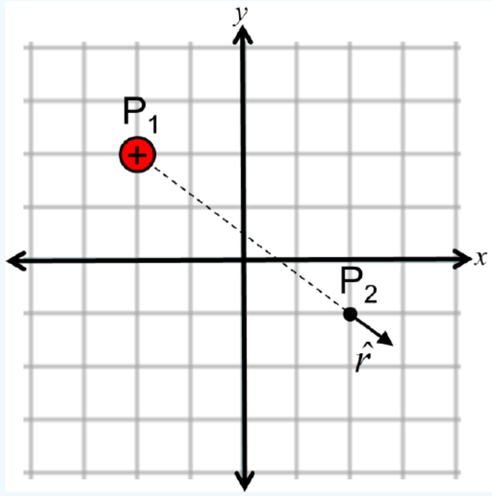


Figure 2.4.3: Sketch of the source charge and test location for part (a) of Example 2.4.3. [4]

CALCULATE To use Eq. 2.4.1, we must first know the charge, separation distance, and radial unit vector. We first convert the charge to coulombs by writing $q = 5\text{ nC} = 5\times10^{-9}\text{ C}$. Next, we identify the location of the source charge as $(x_s, y_s, z_s) = (-2.00\text{ m}, +2.00\text{ m}, 0.00\text{ m})$ and the test location as $(x_t, y_t, z_t) = (+2.00\text{ m}, -1.00\text{ m}, 0.00\text{ m})$. As a result,

$$\begin{aligned}\Delta x &= x_t - x_s = +2.00\text{ m} - -2.00\text{ m} = +4.00\text{ m},\\ \Delta y &= y_t - y_s = -1.00\text{ m} - 2.00\text{ m} = -3.00\text{ m},\\ \Delta z &= z_t - z_s = 0.00\text{ m} - 0.00\text{ m} = 0.00\text{ m},\end{aligned}$$

and therefore the separation distance is given by Equation 2.4.2

$$r = [(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2]^{1/2} = [(+4.00\text{ m})^2 + (-3.00\text{ m})^2 + (0.00\text{ m})^2]^{1/2} = 5.00\text{ m}.$$

The radial unit vector is then given by Equation 2.4.3

$$\hat{r} = \frac{(\Delta x)\hat{\imath} + (\Delta y)\hat{\jmath} + (\Delta z)\hat{k}}{[(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2]^{1/2}} = \frac{(+4.00\text{ m})\hat{\imath} + (-3.00\text{ m})\hat{\jmath} + (0.00\text{ m})\hat{k}}{5.00\text{ m}} = +0.800\hat{\imath} - 0.600\hat{\jmath}.$$

Observe that the radial unit vector is always dimensionless; the dimensions come from the other quantities in the problem. The electric field at the test location is then given by Eq. 2.4.1

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r} = \left(8.99\times10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{5.00\times10^{-9}\text{ C}}{(5.00\text{ m})^2}(0.800\hat{\imath} - 0.600\hat{\jmath}) = (+1.44\text{ N/C})\hat{\imath} + (-1.08\text{ N/C})\hat{\jmath} \tag{2.4.6}$$

CHECK The calculated electric field vector indicates that the vector points down and to the right, a result which is consistent with Fig. 2.4.3. It also indicates that the x-component is larger than the y-component, which is also consistent with the diagram. Finally, the units are in newtons per coulomb, as expected.

(b) The strength of the electric field is the magnitude of the electric field vector

$$|\vec{E}| = \frac{1}{4\pi\epsilon_0}\frac{|q|}{r^2} = [(+1.44\text{ N/C})^2 + (-1.08\text{ N/C})^2]^{1/2} = 1.80\text{ N/C} \tag{2.4.7}$$

Defining Electric Force

Experiments with electric charges have shown that if two objects each have electric charge, then they exert an electric force on each other. The electric field is useful because once we know the electric field at a test location, we can then calculate the electric force $\vec{F}$ on a point charge Q ("test charge") placed at that test location simply by multiplying the charge and electric field together

$$\overrightarrow{F_E} = Q\vec{E}. \tag{2.4.8}$$

With charge in coulombs (C) and electric field in newtons per coulombs (N/C) , their product will be in newtons, as expected for force. Example 2.4.4 show how you can calculate electric force between two point charges.

Example 2.4.4

Continuing with the scenario of Example 2.4.3,

(a) Suppose a second point charge of −3.00 nC is now placed at Point P_2. What is the electric force exerted on this new second charge of −3.00 nC due to the original charge of +5.00 nC?

(b) What is the electric force exerted on the original charge of +5.00 nC due to the second charge of −3.00 nC?

Solution

(a) We will follow a four-step problem-solving strategy:

PLAN Because we already know the electric field at Point P_2 and are given a second point charge, we can use Equation 2.4.8 to find the electric force.

SKETCH We can modify our previous diagram to show the geometry of both charges relative to each other. It is also a good idea to draw the radial unit vector and vectors in the expected directions of the electric field and electric force. Note that the radial unit vector still points in the same direction as in Fig. 2.4.3 because the +5 nC charge is still the source charge creating the electric field that is exerting the electric force on the −3 nC charge. The electric field vector is in the same direction as the radial unit vector because the source charge is positive. However, the direction of the electric force is in the opposite direction of the radial unit vector because we know that positive and negative charges attract.

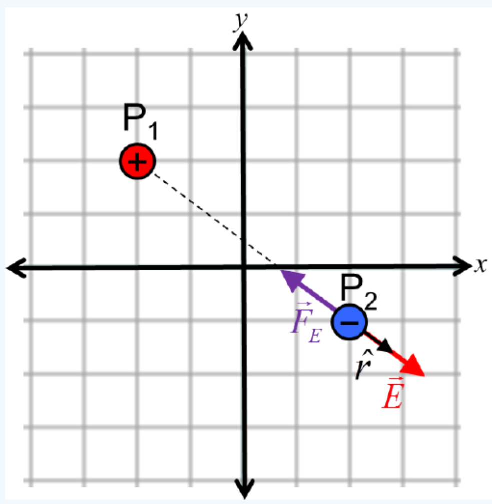


Figure 2.4.4: Sketch of charges, radial unit vector, electric field vector, and electric force vector for part (a) of Example 2.4.4. [4]

CALCULATE The electric force is given by Equation 2.4.8

$$\vec{F_E} = Q\vec{E} = (-3.00\times 10^{-9}\ \mathrm{C})[(+1.44\ \mathrm{N/C})\hat{\imath} + (-1.08\ \mathrm{N/C})\hat{\jmath}] = (-4.31\times 10^{-7}\ \mathrm{N})\hat{\imath} + (3.24\times 10^{-7}\ \mathrm{N})\hat{\jmath}$$

Note that the test charge must be written in coulombs in Equation 2.4.8 to ensure the force is in newtons.

CHECK The calculated force is directed left and up, a result which is consistent with the expected attractive force.

(b) In this question, we are asked to find the electric force exerted by the second charge of −3.00 nC on the +5.00 nC charge. In other words, the charges have switched roles, so the −3.00 nC charge is now the source charge and the +5.00 nC charge is the test charge. The electric field at Point P_1 due to the −3.00 nC charge can then be calculated using the same approach as in Example 2.4.3.

The location of the source charge is $(x_s, y_s, z_s) = (+2.00\ \mathrm{m}, -1.00\ \mathrm{m}, 0.00\ \mathrm{m})$, while the test location is $(x_t, y_t, z_t) = (-2.00\ \mathrm{m}, +2.00\ \mathrm{m}, 0.00\ \mathrm{m})$. As a result,

$$\begin{aligned}\Delta x &= x_t - x_s = -2.00\ \mathrm{m} - 2.00\ \mathrm{m} = -4.00\ \mathrm{m},\\ \Delta y &= y_t - y_s = 2.00\ \mathrm{m} - -1.00\ \mathrm{m} = +3.00\ \mathrm{m},\\ \Delta z &= z_t - z_s = 0.00\ \mathrm{m} - 0.00\ \mathrm{m} = 0.00\ \mathrm{m},\end{aligned}$$

and therefore the separation distance is given by Equation 2.4.2

$$r=[(\Delta x)^2+(\Delta y)^2+(\Delta z)^2]^{1/2}=[(-4.00\text{ m})^2+(+3.00\text{ m})^2+(0.00\text{ m})^2]^{1/2}=5.00\text{ m}.$$

The radial unit vector is then given by Equation 2.4.3

$$\hat{r}=\frac{(\Delta x)\hat{\imath}+(\Delta y)\hat{\jmath}+(\Delta z)\hat{k}}{[(\Delta x)^2+(\Delta y)^2+(\Delta z)^2]^{1/2}}=\frac{(-4.00\text{ m})\hat{\imath}+(3.00+m)\hat{\jmath}+(0.00\text{ m})\hat{k}}{5.00\text{ m}}=-0.800\hat{\imath}+0.600\hat{\jmath}.$$

The electric field at the test location of Point P_1 is given by Eq. 2.4.1

$$\vec{E}=\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r}=\left(8.99\times10^9\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{-3.00\times10^{-9}\text{ C}}{(5.00\text{ m})^2}(-0.800\hat{\imath}+0.600\hat{\jmath})=(+0.863\text{ N/C})\hat{\imath}+(-0.647\text{ N/C})\hat{\jmath}$$

According to Equation 2.4.8, the electric force on the +5 nC charge is then

$$\overrightarrow{F_E}=Q\vec{E}=(+5.00\times10^{-9}\text{ C})[(0.863\text{ N/C})\hat{\imath}+(-0.647\text{ N/C})\hat{\jmath}]=(+4.31\times10^{-7}\text{ N})\hat{\imath}+(-3.24\times10^{-7}\text{ N})\hat{\jmath}$$

We see that the force has the same magnitude as in part (a) but opposite direction. This result should be expected because we know from experiments that the positive and negative charges attract.

PhET Simulation: Electric Field of Dreams

Open the Electric Field of Dreams applet. This applet allows you to construct an electric field and then insert one or more charged particles into the field.

Instructions:

1. Open the applet and choose the appropriate emulation for your system (usually CheepJ will work).
2. Click on the blue dot in the "External Field" box and then drag it in some direction to define the magnitude and direction of the electric field.
3. Press the Click on the "Properties" button, and then enter the desired values of charge and mass for a particle.
4. Click on the "Add" button to add the particle to the field. Is the motion of the particle make sense? Is it consistent with Eq. (2.4.8)?
5. Click on the "Properties" button again, and change the properties to have the opposite sign of charge. Click on the "Add" button to add this second particle.
6. Observe the motion of the particles together. Does the motion of the second particle make sense? Can you see the particles interacting with each other?

Source: Electric Field of Dreams

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

Coulomb's Law

Substituting Equation 2.4.1 into Equation 2.4.8 yields an equation that allows you to calculate the electric force between two point charges directly. This equation is usually called **Coulomb's Law**.

Coulomb's Law

The electric force exerted by a point charge q_1 on a point charge q_2 separated by a distance r_{12} is equal to

$$\vec{F}_{12}=\frac{1}{4\pi\epsilon_0}\frac{q_1q_2}{r_{12}^2}\hat{r}_{12},\tag{2.4.9}$$

where $\hat{r}_{12}$ is the radial unit vector directed from the location of charge q_1 to the location of charge q_2.

We see that the magnitude of electric force is proportional to the product of the charges and inversely proportional to the square of the separation distance between the charges. The form of Equation 2.4.9 shows more clearly why the electric forces calculated in parts (a) and (b) in Example 2.4.4 must be equal in magnitude, because the magnitude of the force does not change if the roles of q_1 and q_2 are reversed and must be opposite in direction because $\hat{r}_{12}=-\hat{r}_{21}$.

Observe that Coulomb's Law allows the electric forces between two charges to be repulsive or attractive. If the charges have the same sign, then their product is positive, the electric force is in the same direction as $\hat{r}$, and the force is repulsive. If the charges have opposite signs, then their product is negative, the force is in the opposite direction of $\hat{r}$, and the force is attractive. This idea is illustrated in Figure 2.4.5.

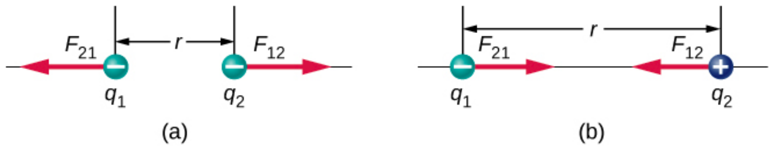


Figure 2.4.5: The electrostatic force $\vec{F}$ between point charges q_1 and q_2 separated by a distance r is given by Coulomb's law. Note that Newton's third law (every force exerted creates an equal and opposite force) applies as usual—the force on q_1 is equal in magnitude and opposite in direction to the force it exerts on q_2. (a) Like charges; (b) unlike charges.

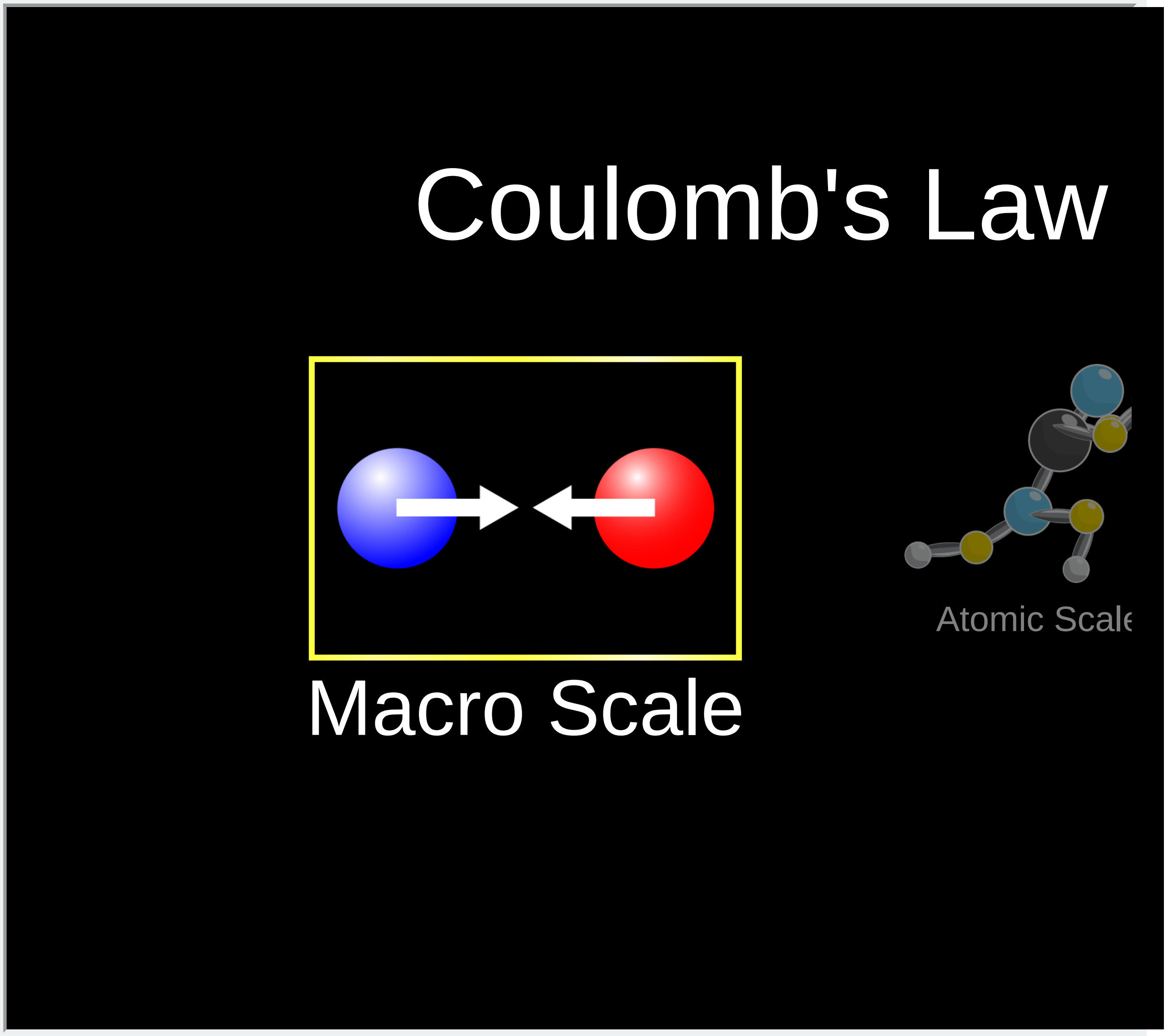
The following interactive simulation further illustrates the qualitative and quantitative properties of Coulomb's Law.

PhET Simulation: Coulomb's Law

Instructions:

1. Select "Macro Scale".
2. Move the slider back and forth for Charge 1 to adjust its charge from negative to zero to positive. How do the forces on the charges change?
3. Move the slider back and forth for Charge 2 to adjust its charge from negative to zero to positive. How do the forces on the charges change?
4. Is there a scenario where the forces are ever unequal in magnitude or in the same direction?
5. Move Charge 1 to the zero position, and then move the position of Charge 2. Does the inverse square law hold?

6. Bonus: Select the "Atomic Scale" button at the bottom of the screen. How do the results change at the smaller scale?



Source: Coulomb's Law

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

Example $2.4.5$ shows how Coulomb's Law can be applied to a familiar system of two charges, the hydrogen atom.

Example $2.4.5$: The Force on the Electron in Hydrogen

A hydrogen atom consists of a single proton and a single electron. The proton has a charge of $+e$, and the electron has $-e$. In the "ground state" of the atom, the electron orbits the proton at the most probable distance of $5.29 \times 10^{-11}\ \mathrm{m}$). Calculate the electric force on the electron due to the proton.

PLAN For the purposes of this example, we are treating the electron and proton as two point particles, each with an electric charge, and we are told the distance between them; we are asked to calculate the force on the electron. We thus use Coulomb's law (Equation 2.4.9).

SKETCH Figure $2.4.6$ shows a diagram the hydrogen atom with the force vector marked.

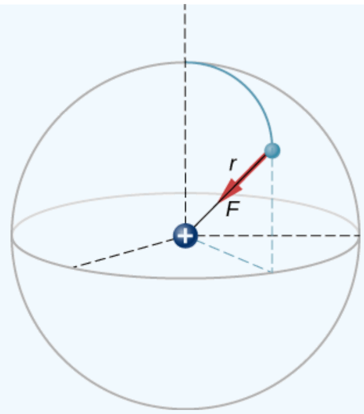


Figure 2.4.6: A schematic depiction of a hydrogen atom, showing the force on the electron. This depiction is only to enable us to calculate the force; the hydrogen atom does not really look like this.

CALCULATE Our two charges are $q_1 = +e = +1.602 \times 10^{-19}\ \mathrm{C}$ and $q_2 = -e = -1.602 \times 10^{-19}\ \mathrm{C}$, and the distance between them is given to be $r = 5.29 \times 10^{-11}\ \mathrm{m}$.

The force on the electron from Equation 2.4.9 is

$$\begin{aligned} \overrightarrow{F_E} &= \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{r} \\ &= \frac{1}{4\pi\left(8.85 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N}\cdot\mathrm{m}^2}\right)} \frac{(1.602 \times 10^{-19}\ \mathrm{C})(-1.602 \times 10^{-19}\ \mathrm{C})}{(5.29 \times 10^{-11}\ \mathrm{m})^2} \\ &= (8.25 \times 10^{-8}\ \mathrm{N})(-\hat{r}) \end{aligned}$$

Because the charges on the two particles are opposite, the force is negative (attractive); the force on the electron points radially directly toward the proton, everywhere in the electron's orbit.

CHECK This is a three-dimensional system, so the electron (and therefore the force on it) can be anywhere in an imaginary spherical shell around the proton. In this "classical" model of the hydrogen atom, the electrostatic force on the electron points in the inward centripetal direction, thus maintaining the electron's orbit. However, the more accurate quantum-mechanical model of hydrogen (discussed in Quantum Mechanics) is utterly different.

If you have studied Newtonian mechanics, you will be familiar with gravitational fields as another type of fundamental physical field. Interestingly, there are parallels between electric and gravitational fields and forces, as discussed in Example 2.4.6.

Example 2.4.6

Comparison: Gravitational Fields & Forces vs. Electric Fields & Forces

The mathematical equation for the electric field $\vec{E}$ of a point charge (Equation 2.4.1) is similar to the gravitational field $\vec{g}$ of Earth

$$\vec{g} = -G\frac{M_e}{r^2}\hat{r}, \tag{2.4.10}$$

where $G = 6.67 \times 10^{-11}\ \mathrm{N}\cdot\mathrm{m}^2/\mathrm{kg}^2$ is the gravitational constant, $M_e = 5.98 \times 10^{24}\ \mathrm{kg}$ is the mass of the Earth, and r is the distance away from the Earth's center, assuming the Earth is approximately spherical. You can see that the proportionality constant G plays the same role for the gravitational field $\vec{g}$ as the electrostatic constant $k = \frac{1}{4\pi\epsilon_0}$ does for the electric field $\vec{E}$.

On the surface of the Earth, the distance $r = R_e = 6.37 \times 10^6\ \mathrm{m}$ is the radius of the Earth, and Equation 2.4.10 reduces down to

$$\vec{g} = -G\frac{M_e}{R_e^2}\hat{r} = (9.81\ \mathrm{N/kg})(-\hat{r}) \quad (\text{On Earth's surface only})$$

The factor $9.81\ \mathrm{N/kg} = 9.81 \mathrm{m/s}^2$ is recognized as the acceleration of gravity, and the factor $-\hat{r}$ indicates that the force is attractive and directed toward the center of the Earth, as expected.

Once we have calculated the gravitational field at some point in space due to a source mass, we can use it any time we want to calculate the resulting force on any test mass we choose to place at that test location according to

$$\overrightarrow{F_G} = M\vec{g}. \tag{2.4.11}$$

In fact, this is exactly what we do when we calculate the force on different masses due to gravity (i.e., weight). You can see this equation is similar in form to Equation 2.4.8 for the electric force, with charge replacing mass. Combining Eqs. 2.4.10 and 2.4.11 yields the equation

$$\vec{F}_{12} = -G\frac{m_1 m_2}{r_{12}^2}\hat{r}_{12}, \tag{2.4.12}$$

which is Newton's Law of Universal Gravitation and, similar to Coulomb's Law, is an inverse-square law in the separation distance.

One significant difference between Coulomb's Law and Newton's Law of Universal Gravitation is that charge can be positive or negative, whereas mass is only positive. As a result, gravity is always an attractive force, whereas electric forces can be attractive or repulsive.

The Meaning of "Electric Field"

Like the gravitational field of an object with mass, you should picture the electric field of a charge-bearing object (the source charge) as a continuous, immaterial substance that surrounds the source charge, filling all of space—in principle, to $\pm\infty$ in all directions. The field exists at every physical point in space. To put it another way, the electric charge on an object alters the space around the charged object in such a way that all other electrically charged objects in space experience an electric force as a result of being in that field. The electric field, then, is the mechanism by which the electric properties of the source charge are transmitted to and through the rest of the universe. (Again, the range of the electric force is infinite.) We will see in subsequent chapters that the speed at which electrical phenomena travel is the same as the speed of light. There is a deep connection between the electric field and light.

In the next section, we describe how to determine the electric field and forces due to multiple individual source charges.

References

1. National Weather Service [Internet]. Graphical forecasts. 2024. National Weather Service. (Select "Temperature" and set the units to Celsius.) (Public domain)
2. National Weather Service [Internet]. Graphical forecasts. 2024. National Weather Service. (Select "Wind Speed".) (Public domain)
3. Wikipedia contributors. Euclidean distance [Internet]. Wikipedia, The Free Encyclopedia.
4. Wikimedia Commons contributors. File:Graph-paper.svg [Internet]. Wikimedia Commons. (CC-SA 3.0, Bobarino). Additional elements by Ronald Kumon (CC-BY-SA 4.0, Ronald Kumon).

2.5: Electric Fields and Forces with Multiple Charges

Learning Objectives

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

- Describe and apply the principle of superposition.
- Calculate the electric field of a set of source charges of either sign.

Principle of Superposition

Another experimental fact about the electric field is that it obeys the superposition principle. In this context, that means that we can (in principle) calculate the total electric field of many source charges by calculating the electric field of only q_1 at a test position P, then calculate the field of q_2 at P, while—and this is the crucial idea—ignoring the field of, and indeed even the existence of, q_1. We can repeat this process, calculating the field of each individual source charge, independently of the existence of any of the other charges. The total electric field, then, is the vector sum of all these fields.

The Electric Field of Multiple Point Charges

We next derive an equation that will allow us to perform the aforementioned calculation. Suppose we have N source charges $q_1, q_2, q_3, \ldots q_N$ located at positions $\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots \vec{r}_N$. By the principle of superposition, it follows from definition of the electric field that the net electric field is then

$$\vec{E} \equiv \frac{1}{4\pi\epsilon_0}\left(\frac{q_1}{r_1^2}\hat{r}_1 + \frac{q_2}{r_2^2}\hat{r}_2 + \frac{q_3}{r_3^2}\hat{r}_3 + \ldots + \frac{q_N}{r_1^2}\hat{r}_N\right) \tag{2.5.1}$$

or, more compactly,

$$\vec{E}(P) \equiv \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i. \tag{2.5.2}$$

This expression is called the electric field at position $P = P(x, y, z)$ of the N source charges. Here, P is the location of the point in space where you are calculating the field and is relative to the positions $\vec{r}_i$ of the source charges (Figure $2.5.1$). Note that we have to impose a coordinate system to solve actual problems.

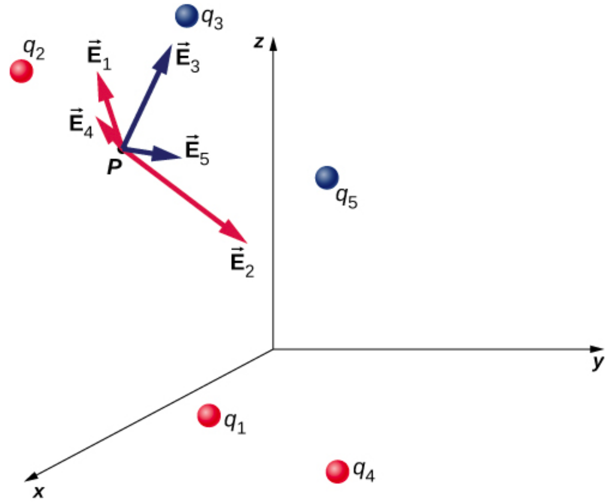


Figure $2.5.1$: Each of these five source charges creates its own electric field at every point in space; shown here are the field vectors at an arbitrary point P. Like the electric force, the net electric field obeys the superposition principle.

Notice that the calculation of the electric field does not use the test charge. Thus, the physically useful approach is to calculate the electric field and then use it to calculate the force on some test charge later, if needed. Different test charges may experience different forces, but the same electric field exists at the test location by Equation 2.5.2. That being said, recall that there is no fundamental difference between a test charge and a source charge; these are merely convenient labels for the system of interest. Any charge produces an electric field; however, just as Earth's orbit is not affected by Earth's own gravity, a charge is not subject to a force due to the electric field it generates. Charges are only subject to forces from the electric fields of other charges.

Examples $2.5.1A$ and $2.5.1B$ provide some examples of how to apply Equation 2.5.2

✓ Example $2.5.1A$: The E-field of an Atom

In an ionized helium atom, the most probable distance between the nucleus and the electron is $r = 26.5 \times 10^{-12}\ \text{m}$. What is the electric field due to the nucleus at the location of the electron?

PLAN Note that although the electron is mentioned, it is not used in any calculation. The problem asks for an electric field, not a force; hence, there is only one charge involved, and the problem specifically asks for the field due to the nucleus. Thus, only the distance of the electron matters. Also, since the distance between the two protons in the nucleus is much, much smaller than the distance of the electron from the nucleus, we can treat the two protons as a single charge $+2e$.

SKETCH

A diagram of the system is shown in Figure $2.5.2$.

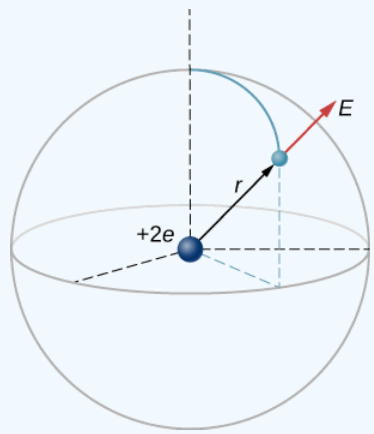


Figure $2.5.2$: A schematic representation of a helium atom. Again, helium physically looks nothing like this, but this sort of diagram is helpful for calculating the electric field of the nucleus.

CALCULATE The electric field is calculated by

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i.$$

Since there is only one source charge (the nucleus), this expression simplifies to

$$\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r}.$$

Here, $q = 2e = 2(1.6 \times 10^{-19}\ \text{C})$ (since there are two protons) and r is given; substituting gives

$$\begin{aligned}\vec{E} &= \frac{1}{4\pi\left(8.85 \times 10^{-12}\frac{\text{C}^2}{\text{N}\cdot\text{m}^2}\right)}\frac{2(1.6 \times 10^{-19}\ \text{C})}{26.5 \times 10^{-12}\ \text{m})^2}\hat{r} \\ &= 4.1 \times 10^{12}\frac{\text{N}}{\text{C}}\hat{r}.\end{aligned}$$

CHECK The direction of $\vec{E}$ is radially away from the nucleus in all directions. Why? Because a positive test charge placed in this field would accelerate radially away from the nucleus (since it is also positively charged), and again, the convention is that the direction of the electric field vector is defined in terms of the direction of the force it would apply to positive test charges.

Example 2.5.1B: The E-Field above Two Equal Charges

a. Find the electric field (magnitude and direction) a distance z above the midpoint between two equal charges $+q$ that are a distance d apart (Figure 2.5.3). Check that your result is consistent with what you would expect when $z \gg d$.

b. Next, change the right-hand charge to $-q$ instead of $+q$. What is the electric field at the same location with this new configuration of charges?

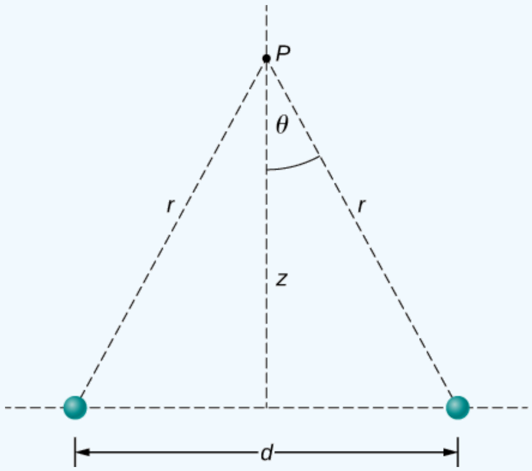


Figure 2.5.3: Finding the field of two identical source charges at the point P. The charges are separated by a distance d with perpendicular bisector passing through P at a distance z. Due to the symmetry, the net field at P is entirely vertical. (Notice that this is **not** true away from the midline between the charges.)

PLAN We add the two fields as vectors, per Equation 2.5.2. Notice that the system (and therefore the field) is symmetrical about the vertical axis; as a result, the horizontal components of the field vectors cancel. This simplifies the math. Also, we take care to express our final answer in terms of only quantities that are given in the original statement of the problem: q, z, d, and constants (π, ϵ_0).

SKETCH Figure 2.5.4 shows the charge configuration with the components of the electric field drawn in.

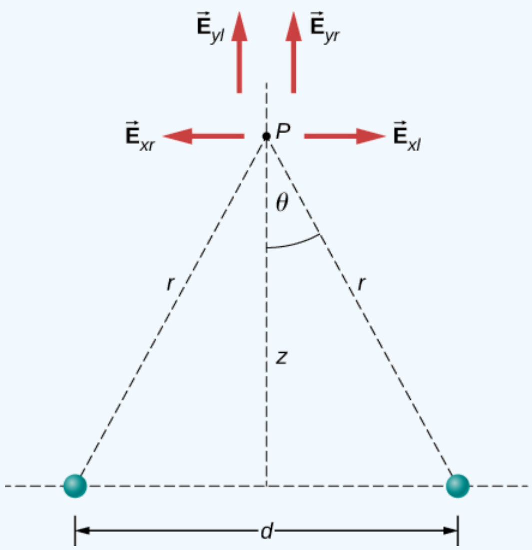


Figure 2.5.4. Note that the horizontal components of the electric fields from the two charges cancel each other out, while the vertical components add together.

CALCULATE

a. By symmetry, the horizontal (x)-components of $\vec{E}$ cancel (Figure 2.5.4);

$$\begin{aligned} E_x &= \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\sin\theta - \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\sin\theta \\ &= 0. \end{aligned}$$

The vertical (z)-component is given by

$$\begin{aligned} E_z &= \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\cos\,\theta + \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\cos\,\theta \\ &= \frac{1}{4\pi\epsilon_0}\frac{2q}{r^2}\cos\,\theta. \end{aligned}$$

Since none of the other components survive, this is the entire electric field, and it points in the $\hat{k}$ direction. Notice that this calculation uses the principle of **superposition**; we calculate the fields of the two charges independently and then add them together.

What we want to do now is replace the quantities in this expression that we don't know (such as r), or can't easily measure (such as $cos\,\theta$ with quantities that we do know, or can measure. In this case, by geometry,

$$r^2 = z^2 + \left(\frac{d}{2}\right)^2$$

and

$$\cos\,\theta = \frac{z}{R} = \frac{z}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{1/2}}.$$

Thus, substituting,

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{2q}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^2}\frac{z}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{1/2}}\hat{k}.$$

Simplifying, the desired answer is

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{2qz}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{3/2}}\hat{k}. \tag{2.5.3}$$

b. If the source charges are equal and opposite, the vertical components cancel because

$$E_z = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\,cos\,\theta - \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\cos\,\theta = 0$$

and we get, for the horizontal component of $\vec{E}$.

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{qd}{\left[z^2 + \left(\frac{d}{2}\right)^2\right]^{3/2}}\hat{i}. \tag{2.5.4}$$

CHECK

It is a very common and very useful technique in physics to check whether your answer is reasonable by evaluating it at extreme cases. In this example, we should evaluate the field expressions for the cases $d = 0$, $z \gg d$, and $z \to \infty$, and confirm that the resulting expressions match our physical expectations. Let's do so:

Let's start with Equation 2.5.3, the field of two identical charges. From far away (i.e., $z >> d$), the two source charges should "merge" and we should then "see" the field of just one charge, of size $2q$. So, let $z \gg d$; then we can neglect d^2 in Equation 2.5.3 to obtain

$$\begin{aligned}\lim_{d\to 0}\vec{E} &= \frac{1}{4\pi\epsilon_0}\frac{2qz}{[z^2]^{3/2}}\hat{k} \\ &= \frac{1}{4\pi\epsilon_0}\frac{2qz}{z^3}\hat{k} \\ &= \frac{1}{4\pi\epsilon_0}\frac{2q}{z^2}\hat{k},\end{aligned}$$

which is the correct expression for a field at a distance z away from a charge $2q$.

Next, we consider the field of equal and opposite charges, Equation 2.5.4. It can be shown (via a Taylor expansion) that for $d \ll z \ll \infty$, , this becomes

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{qd}{z^3}\hat{i},$$

which is the field of a dipole, a system that we will study in more detail later. (Note that the units of $\vec{E}$ are still correct in this expression, since the units of d in the numerator cancel the unit of the "extra" z in the denominator.) If z is *very* large $(z \to \infty)$, then $E \to 0$, as it should; the two charges "merge" and so cancel out.

? Exercise 2.5.1

Can you answer the questions in Example $2.5.1B$ without using symmetry arguments? (Hint: Use Equation 2.5.2 and directly calculate the values of \hat{r}.)

Superposition of Electric Forces

Electric forces exerted by multiple charges also exhibit superposition.

Proof

Suppose we have N source charges $q_1, q_2, q_3, \ldots q_N$ located at positions $\vec{r}_1, \vec{r}_2, \vec{r}_3, \ldots \vec{r}_N$ along with a test charge Q at separate location (Fig. $2.5.5$). (Note that there is no physical difference between Q and q_i; the difference in labels is merely to allow clear discussion, with Q being the charge we are determining the force on.)

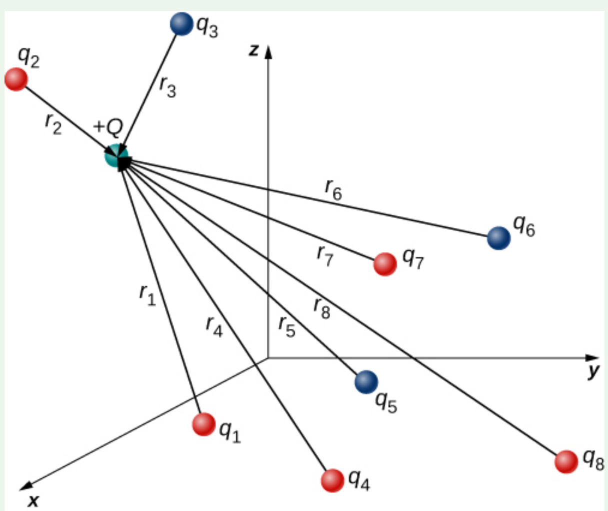


Figure $2.5.5$: In this example, the eight source ($N = 8$) charges each apply a force on the single test charge **Q**. Each force can be calculated independently of the other seven forces. This is the essence of the superposition principle.

By Equation 2.5.2, the net field at the test location of Q is

$$\vec{E_{net}}(P) = \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i.$$

From the definition of electric force,

$$\begin{aligned}\vec{F}_{net} &= Q\vec{E}_{net} \\ &= Q\left(\frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i\right) \\ &= \left(\frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\frac{Qq_i}{r_i^2}\hat{r}_i\right) \\ &= \left(\sum_{i=1}^{N}\vec{F}_i\right)\end{aligned}$$

Therefore, the electric force also obeys superposition.

There is a complication, however. Just as the source charges each exert a force on the test charge, so too (by Newton's third law) does the test charge exert an equal and opposite force on each of the source charges. As a consequence, each source charge would change position. However, by the definition of electric force, the force on the test charge is a function of position; thus, as the positions of the source charges change, the net force on the test charge necessarily changes, which changes the force, which again changes the positions. Thus, the entire mathematical analysis quickly becomes intractable. Later, we will learn techniques for handling this situation, but for now, we make the simplifying assumption that the source charges are fixed in place, so that their positions are constant in time. (The test charge is allowed to move.) With this restriction in place, the analysis of charges is known as **electrostatics**, where "statics" refers to the constant (that is, static) positions of the source charges and the force is referred to as an **electrostatic force**.

Example 2.5.2: The Net Force from Two Source Charges

Three different, small charged objects are placed as shown in Figure 2.5.2. The charges q_1 and q_3 are fixed in place; q_2 is free to move. Given $q_1 = 2e$, $q_2 = -3e$, and $q_3 = -5e$, and that $d = 2.0 \times 10^{-7}$ m, what is the net force on the middle charge q_2?

PLAN

There are two possible approaches to this problem. One way would be to use Coulomb's law twice: once to calculate the electric force of q_1 on q_2 and again to calculate the electric force of q_3 on q_2. In this approach, q_2 is our test charge, and q_1 and q_3 are source charges. Superposition of electric forces could then be used to add the two forces to find the net force on q_2.

The other way would be to calculate the net electric field at the test location of q_2 by using superposition of the electric fields due to q_1 and q_2. The net force can then be computed by using the definition of electric force to multiply the net electric field by the charge q_2.

We will show solutions by both approaches. In both approaches, we will model the small charged objects as point charges.

SKETCH Figure 2.5.2 plots the locations of the charges. We also draw the expected directions of the forces given the signs of the charges.

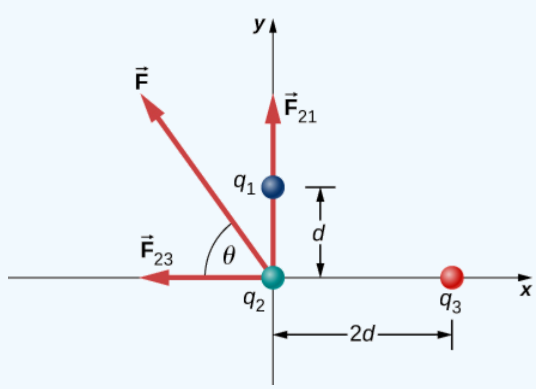


Figure 2.5.4: Source charges q_1 and q_3 each apply a force on q_2.

Solution #1: Force Superposition

CALCULATE We have two source charges q_1 and q_3 a test charge q_2, distances r_{21} and r_{23} and we are asked to find a force. This calls for Coulomb's law and superposition of forces. There are two forces:

$$\vec{F} = \vec{F}_{21} + \vec{F}_{23} = \frac{1}{4\pi\epsilon_0}\left[\frac{q_2 q_1}{r_{21}^2}\hat{j} + \left(-\frac{q_2 q_3}{r_{23}^2}\hat{i}\right)\right].$$

We cannot add these forces directly because they don't point in the same direction: $\vec{F}_{12}$ points only in the $-x$-direction, while $\vec{F}_{13}$ points only in the $+y$-direction. The net force is obtained from applying the Pythagorean theorem to its x- and y-components:

$$F = \sqrt{F_x^2 + F_y^2}$$

and

$$\begin{aligned} F_x &= -F_{23} = -\frac{1}{4\pi\epsilon_0}\frac{q_2 q_3}{r_{23}^2} \\ &= -\left(8.99 \times 10^9 \frac{\mathrm{N}\cdot\mathrm{m}^2}{\mathrm{C}^2}\right)\frac{(4.806 \times 10^{-19}\,\mathrm{C})(8.01 \times 10^{-19}\,\mathrm{C})}{(4.00 \times 10^{-7}\backslash.\mathrm{m})^2} \\ &= -2.16 \times 10^{-14}\,\mathrm{N} \end{aligned}$$

and

$$\begin{aligned} F_y &= F_{21} = \frac{1}{4\pi\epsilon_0}\frac{q_2 q_1}{r_{21}} \\ &= \left(8.99 \times 10^9 \frac{N\cdot m^2}{C^2}\right)\frac{(4.806 \times 10^{-19} C)(3.204 \times 10^{-19} C)}{(2.00 \times 10^{-7} m)^2} \\ &= 3.45 \times 10^{-14}\,N. \end{aligned}$$

We find that

$$\begin{aligned} F &= \sqrt{F_x^2 + F_y^2} \\ &= 4.07 \times 10^{-14}\,\mathrm{N} \end{aligned}$$

at an angle of

$$\begin{aligned} \phi &= \tan^{-1}\left(\frac{F_y}{F_x}\right) \\ &= \tan^{-1}\left(\frac{3.46 \times 10^{-14} N}{-2.16 \times 10^{-14} N}\right) \\ &= -58^\circ, \end{aligned}$$

that is, 58° above the $-x$-axis, as shown in the diagram.

CHECK Notice that when we substituted the numerical values of the charges, we did not include the negative sign of either q_1 or q_3. Recall that negative signs on vector quantities indicate a reversal of direction of the vector in question. But for electric forces, the direction of the force is determined by the types (signs) of both interacting charges; we determine the force directions by considering whether the signs of the two charges are the same or are opposite. If you also include negative signs from negative charges when you substitute numbers, you run the risk of mathematically reversing the direction of the force you are calculating. Thus, the safest thing to do is to calculate just the magnitude of the force, using the absolute values of the charges, and determine the directions physically.

It is also worth noting that the only new concept in this example is how to calculate the electric forces; everything else (getting the net force from its components, breaking the forces into their components, finding the direction of the net force) is the same as force problems you have done earlier.

Solution #2: Field Superposition

CALCULATE First, calculate the electric field due to q_1 at the test location of q_2. We identify the location of the source charge as $(x_s, y_s) = (0, d)$ and the test location as $(x_t, y_t) = (0, 0)$. As a result,

$$\begin{aligned} \Delta x &= x_t - x_s = 0 - 0 = 0, \\ \Delta y &= y_t - y_s = 0 - d = -d, \end{aligned}$$

and therefore the separation distance is given by

$$r = [(\Delta x)^2 + (\Delta y)^2]^{1/2} = [(0)^2 + (-d)^2]^{1/2} = d.$$

The radial unit vector is then given by

$$\hat{r} = \frac{(\Delta x)\hat{i} + (\Delta y)\hat{j}}{[(\Delta x)^2 + (\Delta y)^2]^{1/2}} = \frac{(0)\hat{i} + (-d)\hat{j}}{d} = -\hat{j}.$$

The electric field at the test location is then given by

$$\vec{E}_1 = k_e \frac{q_1}{r^2}\hat{r} = k_e \frac{2e}{d^2}(-\hat{j})$$

Next, calculate the electric field due to q_3 at the test location of q_2. We identify the location of the source charge as $(x_s, y_s) = (2d, 0)$ and the test location as $(x_t, y_t) = (0, 0)$. As a result,

$$\begin{aligned} \Delta x &= x_t - x_s = 0 - 2d = -2d, \\ \Delta y &= y_t - y_s = 0 - 0 = 0, \end{aligned}$$

and therefore the separation distance is given by

$$r = [(\Delta x)^2 + (\Delta y)^2]^{1/2} = [(2d)^2 + (0)^2]^{1/2} = 2d.$$

The radial unit vector is then given by

$$\hat{r} = \frac{(\Delta x)\hat{i} + (\Delta y)\hat{j}}{[(\Delta x)^2 + (\Delta y)^2]^{1/2}} = \frac{(-2d)\hat{i} + (0)\hat{j}}{2d} = -\hat{i}.$$

The electric field at the test location is then given by

$$\vec{E}_3 = k_e \frac{q_3}{r^2}\hat{r} = k_e \frac{-5e}{(2d)^2}(-\hat{i})$$

The net electric field is

$$\vec{E}_{net} = \vec{E}_1 + \vec{E}_3 = k_e \frac{e}{d^2}(-2\hat{j} + 1.25\hat{i}) \tag{2.5.5}$$

The net electric force is then

$$\begin{aligned} \vec{F}_{net} &= q_2 \vec{E}_{net} \\ &= (-3e)\left[k_e \frac{e}{d^2}(-2\hat{j} + 1.25\hat{i})\right] \\ &= k_e \frac{e^2}{d^2}(6\hat{j} - 3.75\hat{i}) \\ &= \left(8.99 \times 10^9 \frac{\mathrm{N} \cdot \mathrm{m}^2}{\mathrm{C}^2}\right) \frac{(1.6 \times 10^{-19}\ \mathrm{C})^2}{(2.0 \times 10^{-7}\ \mathrm{m})^2}(-3.75\hat{i} + 6\hat{j}) \\ &= (-2.16 \times 10^{-14}\ \mathrm{N})\hat{i} + (3.45 \times 10^{-14}\ \mathrm{N})\hat{j} \end{aligned}$$

This value is the same result as in Solution #1.

CHECK While the second approach may take a bit more math, it has the advantage of following a systematic method of calculation. You do not necessarily have to think separately about which directions the fields are going; these directions automatically come out of the calculation of the products of the source charges with the radial unit vectors. The force calculation is somewhat simpler because you only have to multiply the test charge by the net electric field once.

It is reassuring that both methods give the same result!

Exercise 2.5.2

What would be different in Example $2.5.2$ if q_1 were negative rather than positive?

Answer

The net force would point $58°$ below the $-x$-axis.

As suggested above, a charge that is not held in place will move under the influence of electric forces. By solving the equations of motion with a computer, it is possible to calculate the path of the charge's motion.

PhET Simulation: Electric Field Hockey

The goal of the game of "Electric Field Hockey" is to get the positively-charged puck in the goal by placing other charges in fixed locations on the field.

Instructions:

1. Click on the link to open a new browser window containing the simulation.
2. Click on the arrow in the center of the simulation.
3. In the resulting dialog box, select "Run CheerpJ Browser-Compatible Version" to run the Java simulation in your browser.
4. Place positive and negative charges on to the field in a way that you think will attract or repel the positively-charged puck.
5. Select the "Trace" checkbox to show the path of the puck as it travels.
6. Press "Start" to begin the simulation.
7. If the puck goes into the goal, you win. If it does not go into the goal, press the "Reset" button.
8. If you need some guidance, turn on the electric field display by selecting the "Field" checkbox.
9. At each level of difficulty, you should strive to use as few positive and negative charges as possible to score a goal.

Source: Electric Field Hockey

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

2.6: Electric Field Diagrams

Learning Objectives

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

- Explain the purpose of an electric field diagram
- Describe the relationship between a vector diagram and a field line diagram
- Explain the rules for creating a field diagram and why these rules make physical sense
- Sketch the field of an arbitrary source charge

Now that we have experience calculating electric fields, let's try to gain some insight into their geometry. As mentioned earlier, our model is that the charge on an object (the source charge) alters space in the region around it in such a way that when another charged object (the test charge) is placed in that region of space, that test charge experiences an electric force. The concepts of electric field vectors and field lines and their corresponding diagrams enable us to visualize how the space is altered, allowing us to visualize the field. The purpose of this section is to enable you to create sketches of this geometry, so we will list the specific steps and rules involved in creating an accurate and useful sketch of an electric field.

It is important to remember that electric fields are three-dimensional. Although we include some pseudo-three-dimensional images in this book, several diagrams (both here and in subsequent chapters) will be two-dimensional projections or cross-sections. Always keep in mind that, in fact, you are looking at a three-dimensional phenomenon.

Field-Vector Diagrams

Our starting point is the physical fact that the electric field of the source charge causes a test charge in that field to experience a force. By definition, electric field vectors point in the same direction as the electric force that a (hypothetical) positive test charge would experience if placed in the field (Figure $2.6.1$). This kind of diagram is called a electric **field-vector diagram**.

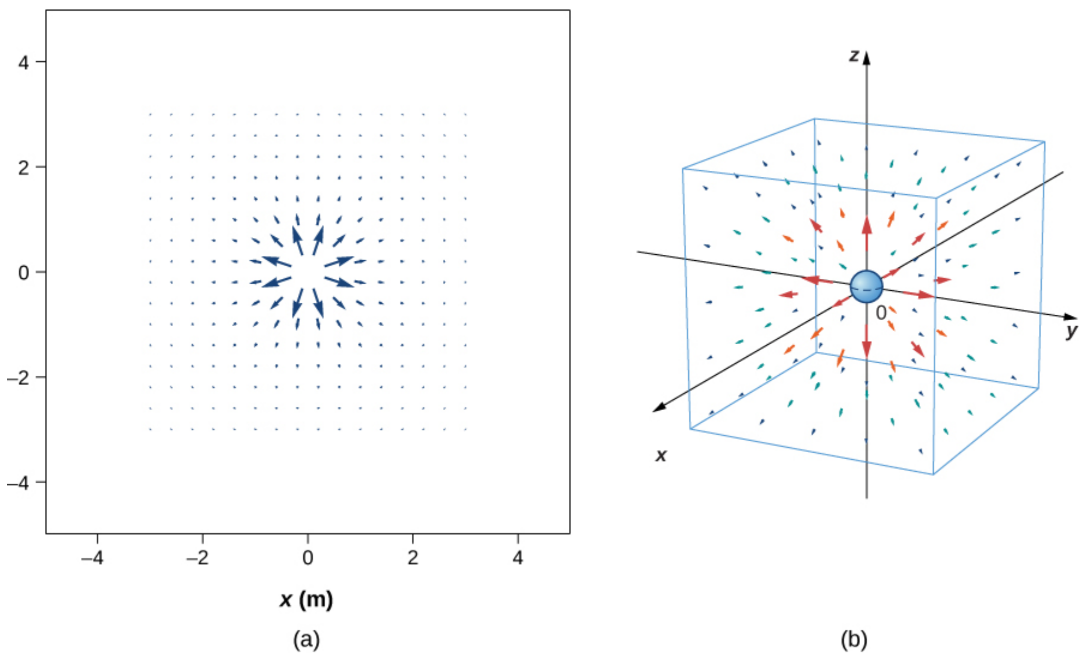


Figure $2.6.1$: Electric field vector diagrams of a positive point charge. A large number of field vectors are shown. Like all vector arrows, the length of each vector is proportional to the magnitude of the field at each point. (a) Field in two dimensions; (b) field in three dimensions.

We plotted many field vectors in the figure using a uniform distribution of points around the source charge. Because the electric field is a vector, the arrows that we draw correspond at every point in space to both the magnitude and the direction of the field at that point. As always, the length of the arrow that we draw corresponds to the magnitude of the field vector at that point. For a point source charge, the length decreases by the square of the distance from the source charge. In addition, the direction of the field vector is radially away from the source charge because a positive test charge would experience a force in that direction in that field. (Again, remember that the actual field is three-dimensional; there are also field lines pointing out of and into the page.)

Finally, it should be noted that the field vectors in these diagrams should not be interpreted to have any spatial extent. In other words, the electric field vector acts only on a test charge placed at its starting point and not other points along the extent of the vector. Neighboring points can have vectors pointing in different directions, and if the density of points is high enough or the field vectors are long enough, the vectors may cross. These cases have no ambiguity or problem because the resulting field vectors are interpreted as acting at *different points* in space.

Field-vector diagrams can sometimes become less useful as the source charge distribution becomes more complicated. For example, consider the vector field diagram of a dipole (Figure $2.6.2$).

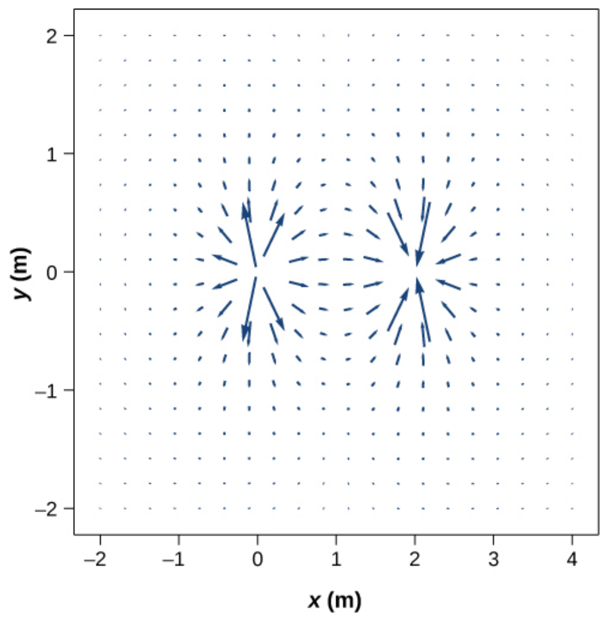


Figure $2.6.2$: The vector field diagram of a dipole. Even with just two identical charges, the vector field diagram becomes difficult to understand.

Field-Line Diagrams

There is a more useful way to present the same information. Rather than drawing a large number of increasingly smaller vector arrows, we connect them all together, forming continuous lines and curves, as shown in Figure $2.6.3$. This style of visualization is called a **field-line diagram**.

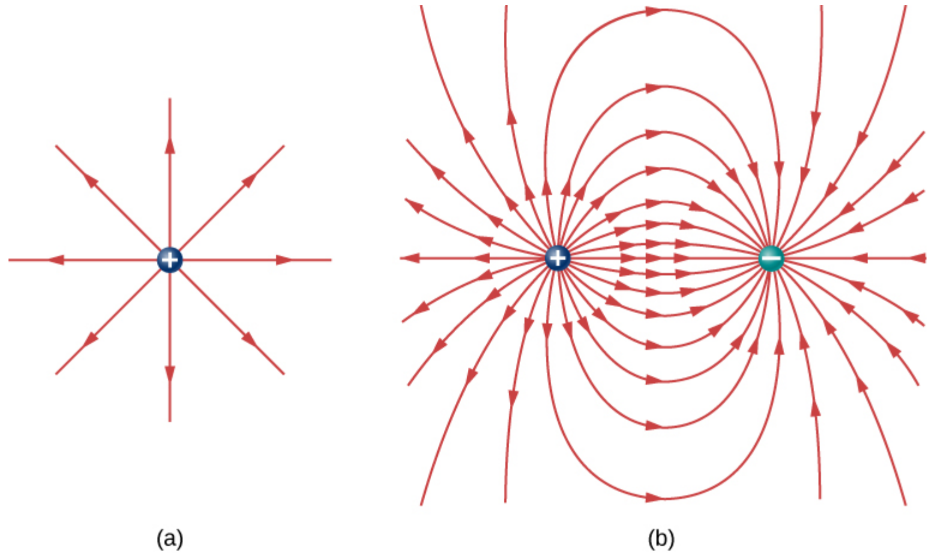


Figure $2.6.3$: (a) The electric field line diagram of a positive point charge. (b) The field line diagram of a dipole. In both diagrams, the magnitude of the field is indicated by the field line density. The field **vectors** (not shown here) are everywhere tangent to the field lines.

Although it may not be obvious initially, these field-line diagrams convey the same information about the electric field as the field-vector diagrams. First, the direction of the field at every point is simply the direction of the field vector at that same point. In other words, at any point in space, the field vector at each point is tangent to the field line at that same point. The arrowhead placed on a field line indicates its direction.

But how is the magnitude of the field indicated since we no longer have the length of a vector to show us? The magnitude is instead indicated by the **field-line density**—that is, the number of field lines per unit area passing through a small cross-sectional area perpendicular to the electric field. This field-line density is drawn to be proportional to the magnitude of the field at that cross-section. As a result, if the field lines are close together (that is, the field line density is greater), the field's magnitude is large in that region. If the field lines are far apart at the cross-section, this indicates the field's magnitude is small. Figure $2.6.4$ shows the idea.

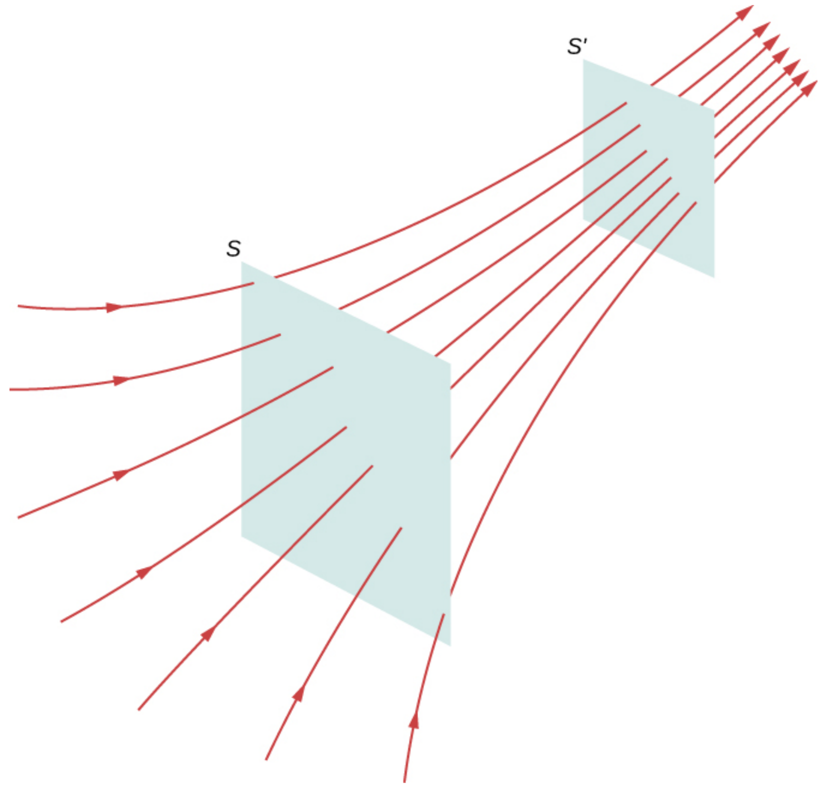


Figure 2.6.4: Electric field lines passing through imaginary areas. Since the number of lines passing through each area is the same, but the areas themselves are different, the field line density is different. This indicates different magnitudes of the electric field at these points.

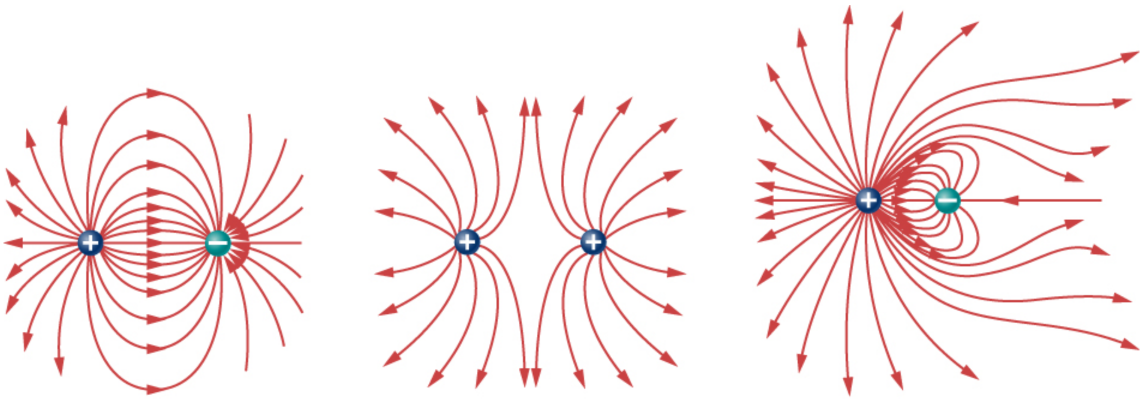
In Figure 2.6.4, the same number of field lines passes through both surfaces S and S', but the surface S is larger than surface S'. Therefore, the density of field lines (number of lines per unit area) is larger at the location of S', indicating that the electric field is stronger at the location of S' than at S. The rules for creating an electric field diagram are as follows.

Problem-Solving Strategy: Drawing Electric Field Lines

1. Electric field lines either originate on positive charges or come in from infinity and either terminate on negative charges or extend out to infinity.
2. The number of field lines originating or terminating at a charge is proportional to the magnitude of that charge. A charge of $2q$ will have twice as many lines as a charge of q.
3. At every point in space, the field vector at that point is tangent to the field line at that same point.
4. The field line density at any point in space is proportional to (and therefore is representative of) the magnitude of the field at that point in space.
5. Field lines can never cross. Since a field line represents the direction of the field at a given point, if two field lines crossed at some point, that would imply that the electric field was pointing in two different directions at a single point. This result would suggest that the (net) force on a test charge placed at that point would point in two different directions. Since this situation is obviously impossible, field lines must never cross.

Always keep in mind that field lines serve only as a convenient way to visualize the electric field; they are not physical entities. Although the direction and relative intensity of the electric field can be deduced from a set of field lines, the lines can also be misleading. For example, the field lines drawn to represent the electric field in a region must, by necessity, be discrete. However, the actual electric field in that region exists at every point in space.

Field lines for three groups of discrete charges are shown in Figure 2.6.5. Since the charges in parts (a) and (b) have the same magnitude, the same number of field lines are shown starting from or terminating on each charge. In (c), however, we draw three times as many field lines leaving the $+3q$ charge as entering the $-q$. The field lines that do not terminate at $-q$ emanate outward from the charge configuration to infinity.



(a) (b) (c)

Figure $2.6.5$: Three typical electric field diagrams. (a) A dipole. (b) Two identical charges. (c) Two charges with opposite signs and different magnitudes. Can you tell from the diagram which charge has the larger magnitude?

The ability to construct an accurate electric field diagram is an important, useful skill; it makes it much easier to estimate, predict, and calculate the electric field of a source charge. The best way to develop this skill is with software that allows you to place source charges and then draw the net field upon request. One such simulation is given below. Practice drawing field diagrams on your own, and then check your predictions with the computer-drawn diagrams.

PhET Simulation: Charges and Fields

Instructions:

1. Make sure that the "Electric Field," "Values," and "Grid" boxes are checked in the box in the upper right corner of the screen.
2. Add one positive 1 nC charge to the center of the main window by dragging and dropping it from the box at the bottom of the window.
3. Observe the resulting electric field. By default, all the arrows will be the same length, and the field's strength will be designated by the arrow's brightness (i.e., dimmer arrows are weaker field strengths).
4. Add a sensor to the window by dragging and dropping it from the box at the bottom. Move the sensor around and observe how the magnitude and direction of the net field change. (Note that the electric field is given in units of V/m, where 1 V/m = 1 N/C).
5. Add a second positive charge to the window. Arrange the two charges in a horizontal line along the midline of the window, like in part (a) of Exercise $2.6.1B$ of the previous section. Move the sensor around the window and observe how the magnitude and direction of the field changes. Does superposition hold? Do you see the same qualitative result as in the exercise?
6. Replace the left positive charge with a negative charge, like in part (b) of Exercise $2.6.1B$ in the previous section. Does superposition hold? Do you see the same qualitative result as in the exercise?
7. Experiment with other combinations of three or more positive and negative charges, and observe how the electric field changes. Does the field pattern make sense based on superposition?

Source: Charges and Fields

Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0 (opens in new window).

2.7: Common Models of Electric Field

LEARNING OBJECTIVES

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

- Explain the concept of continuous charge distribution and explain why it is used.
- List and apply electric field models for some common geometric distributions of charge.
- Evaluate limiting cases of electric field models.

Continuous Charge Distributions

The charge distributions we have seen so far have been composed of **discrete charges**, in other words, made up of individual point particles. However, in many practical situations, there are so many charges in the system (often 10^{22} or more!) that it is not possible to use superposition to calculate the electric field from each charge individual. To deal with this situation, we will introduce the approximation of a **continuous charge distribution** with at least one nonzero dimension. If a charge distribution is continuous rather than discrete, we can generalize the definition of the electric field. We divide the charge into infinitesimal pieces and treat each as a point charge. When we used superposition to calculate the total electric field at a test location using a continuous distribution, the discrete sum of electric-field contributions from discrete charges turns into an integral of electric-field contributions over all the infinitesimal "bits" of charge in the system.

Note that because charge is actually 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 model a bucket of water as a continuous fluid, rather than a collection of individual H_2O molecules.

When we use the assumption of continuous distribution of charge, the mathematics of calculating the electric field can still be challenging. In some special cases, we can construct the integrals, perform the integration analytically, and then write down closed-form formulas for the electric field at test locations in space. However, in general, it is often necessary to use computer simulations to calculate the electric field throughout the space surrounding the charge distribution.

In this chapter, we will summarize several models corresponding to the most common geometric distributions of charge. In each of these models, the electric field can be written in a closed-form expression for a specified region of space around the charge distribution. Students who are interested in the details of the calculations should refer to the section on Electric Fields of Charge Distributions in the chapter on Direct Calculation of Electrical Quantities from Charge Distributions. We will make use of these results in later chapters.

Electric Fields of Common Charge Distributions

In the following results, it is assumed that the **charge density** is uniform on the object or, in other words, the charge per unit length or charge per area is the same everywhere on the object.

Finite Line Segment of Charge

Suppose a finite line segment of charge is located along the x-axis, as illustrated in Fig. $2.7.1$. Assume that the segment has charge density λ (e.g., in coulombs per meters). The electric field along the perpendicular bisector of the segment at a distance z away from the segment is

$$\vec{E}(z)=\frac{1}{4\pi\epsilon_0}\frac{\lambda L}{z\sqrt{z^2+\frac{L^2}{4}}}\hat{k}. \quad \text{(finite line segment)} \tag{2.7.1}$$

Observe that if the charge density λ is positive, then the electric field points away from the segment. Otherwise, if the charge density λ is negative, then the electric field points toward the segment.

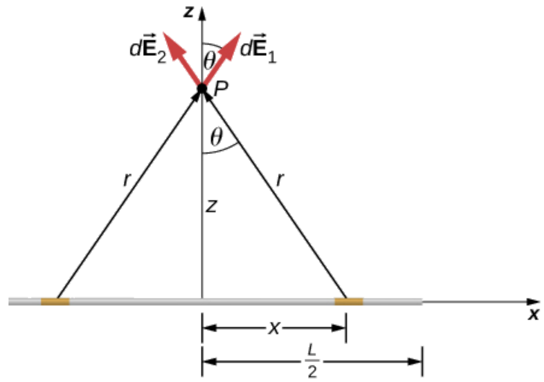


Figure 2.7.1: A uniformly charged segment of wire and its coordinate system. The electric field at point P can be found by applying the superposition principle to symmetrically placed charge elements and then integrating.

Infinite Line of Charge

If the length of the finite segment is allowed to go to infinite in both directions, then Equation 2.7.1 becomes

$$\vec{E}(z)=\frac{1}{4\pi\epsilon_0}\frac{2\lambda}{z}\hat{k}=\frac{\lambda}{2\pi\epsilon_0 z}\hat{k}. \quad \text{(infinite line)} \tag{2.7.2}$$

In contrast to a point charge which falls off as the inverse square of the distance from the charge ($1/r^2$), the electric field of an infinite line charge falls off as the simple reciprocal of the distance from the line ($1/z$). Of course, no actual infinite line charges exist, so what is the purpose of this result? The answer is that if a line charge is sufficiently long as compared to the separation distance of the test location, then Equation 2.7.2 will be a good approximation and certainly simpler than the exact result of Equation 2.7.1.

Ring of Charge

If we take a finite line segment with linear charge density λ and total charge q_0 and wrap it into a closed circular loop, then we get a ring of charge of radius R (Fig. $2.7.2$). The electric field at a distance z from the plane of the ring along its symmetry axis is

$$\vec{E}(z)=\frac{1}{4\pi\epsilon_0}\frac{2\pi\lambda Rz}{(z^2+R^2)^{3/2}}\hat{k}=\frac{1}{4\pi\epsilon_0}\frac{q_{tot}z}{(z^2+R^2)^{3/2}}\hat{k}. \quad \text{(ring)} \tag{2.7.3}$$

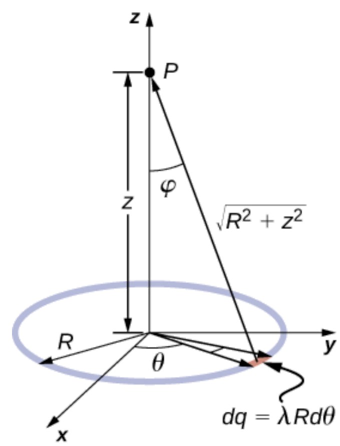


Figure 2.7.2: A ring of charge of radius R and its coordinate system. The electric field at point P can be found by applying the superposition principle to symmetrically placed charge elements and integrating.

Exercise 2.7.1

What is the electric field along the axis of the ring in the limit that $z \gg R$?

Answer

$$\vec{E}\approx\frac{1}{4\pi\epsilon_0}\frac{q_{tot}}{z^2}\hat{z},$$

which is the expression for the field of a point charge q_{tot}. (If you get far enough away from a ring, it looks pointlike!)

Disk of Charge

Once we know the electric field along the axis of a ring of charge, we can use this result to find the electric field along the axis of a disk of charge because a disk of charge can be modeled as a set of rings of charge of increasing radius (Fig. 2.7.3. The electric field for a disk of radius R and uniform surface charge density σ (coulombs per m^2) along the axis of the disk at a distance z from the disk is

$$\vec{E}(z)=\frac{1}{4\pi\epsilon_0}\left(2\pi\sigma-\frac{2\pi\sigma z}{\sqrt{z^2+R^2}}\right)\hat{k}=\frac{\sigma}{2\epsilon_0}\left(1-\frac{z}{\sqrt{z^2+R^2}}\right)\hat{k}. \quad \text{(disk)} \tag{2.7.4}$$

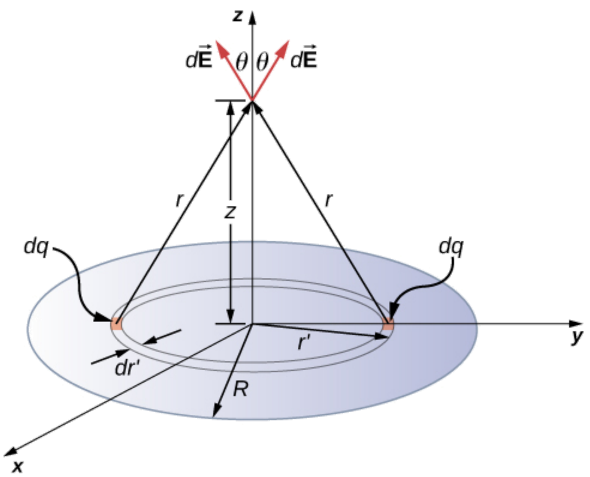


Figure 2.7.4: A ring of charge of radius R and its coordinate system. The electric field at distance z can be found by applying the superposition principle to rings of increasing radius and integrating.

? Exercise 2.7.2

What is the electric field along the axis of the disk in the limit that $z \gg R$?

Answer

$$\vec{E}(z)\approx\frac{1}{4\pi\epsilon_0}\frac{\sigma\pi R^2}{z^2}\hat{k},$$

which is the expression for the field of a point charge $Q=\sigma\pi R^2$. (If you get far enough away from a disk, it looks pointlike!)

? Exercise 2.7.3

How would the above limit change with a uniformly charged rectangle instead of a disk?

Answer

The result would be the same except that the point charge would be $Q=\sigma ab$ where a and b are the sides of the rectangle.

Infinite Plane

As $R\to\infty$, Equation 2.7.4 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:

$$\vec{E}(z)=\begin{cases}\dfrac{\sigma}{2\epsilon_0}\;\hat{k} & \text{if } z>0\\ \dfrac{\sigma}{2\epsilon_0}(-\hat{k}) & \text{if } z<0\end{cases} \tag{2.7.5}$$

Note that this field is constant. This surprising result is, again, an artifact of our limit, although one that we will use 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. The direction of the field below the plane is opposite to the direction above the plane because the field will always either point away from the plane (if positively charged) or toward the plane (if negatively charged).

Example 2.7.4: The Field of Two Infinite Planes

Find the electric field everywhere resulting from two infinite planes with equal but opposite charge densities (Figure 2.7.5).

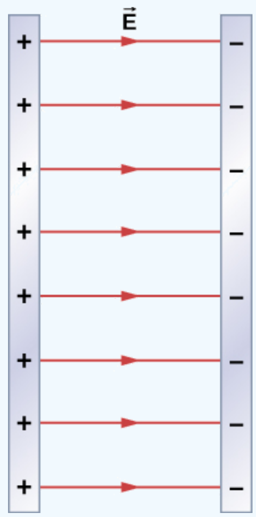


Figure 2.7.5: Two charged infinite planes. Note the direction of the electric field.

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 provide a useful means of creating uniform electric fields. This approach can also be used to model a charged parallel-plate capacitor.

Exercise 2.7.4

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 the planes and have magnitude $\dfrac{\sigma}{\epsilon_0}$ everywhere else.

2.8: Motion of a Charged Particle in an Electric Field

As we have seen, when a particle with charge q is placed in an electric field $\vec{E}$, the field causes an electric force $\vec{F}_E = q\vec{E}$ on the charge, as illustrated in Figure $2.8.1$. A positive charge has a force in the same direction of the electric field, while a negative charge has an electric force in the opposite direction of the field.

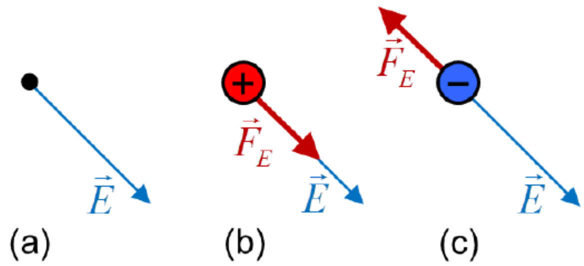


Figure 2.8.1: (a) Electric field at a point in space. (b) A positive charge placed at the same point in space will have a force in the same direction of the electric field. (c) A positive charge placed at the same point in space will have a force in the opposite direction of the electric field. (Ronald Kumon, CC-BY-SA 4.0)

According to Newton's Second Law, the net force $\vec{F}_{\text{net}} = m\vec{a}$, where m is the mass of the object upon which the force is exerted, and $\vec{a}$ is the acceleration of the object. If the electric force is the only force acting on the charged particle, then the acceleration of the particle will be

$$\vec{a} = \frac{\vec{F}}{m} = \frac{q}{m}\vec{E} \tag{2.8.1}$$

We see that the acceleration is proportional to the electric field by a factor of the charge-to-mass ratio of the particle. In general, the electric field will vary from point to point in space and so the acceleration will also vary from point to point, and computational methods may be needed to determine the motion of the particle.

Charged Particle Motion in a Uniform Electric Field

in the special case that the electric field is uniform, then the acceleration is constant, and kinematic equations can be used to calculate the position and velocity of the particle.

Example $2.8.1$

A thin square conducting plate 0.500 m on each side is given a charge of +0.200 nC. An electron at rest is near the center of the place and 1.00 cm above the plate.

(a) How long does it take for the electron to hit the plate?

(b) What is the velocity of the electron when it hits the plate?

Solution

PLAN

We will model the plate as an infinite plane. This assumption is plausible because the electron is away from the edges of the plate and its distance away from the plate is small compared to the width of the plate. According to Common Models of Electric Field, the electric field for an infinite plane is uniform, and therefore we can use kinematic equations (Motion with Constant Acceleration) to solve for the electron's travel time and final velocity. We will also ignore the effects of gravity on the electron given its very small mass and the effects of drag due its very small size.

SKETCH

We sketch out a side view of the plate, showing the electron at a distance $d = 1.00\,\text{cm} = 0.0100\,\text{m}$ above the plate.

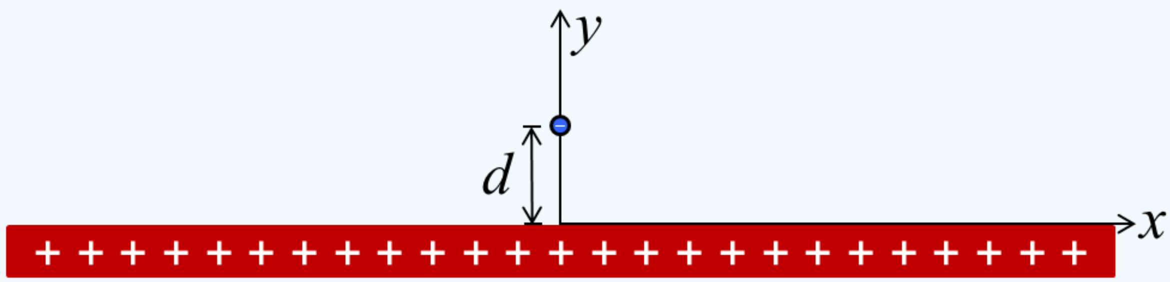


Figure 2.8.2: Electron at a distance d above a positively-charged plate. (Ronald Kumon, CC-BY-SA 4.0)

CALCULATE

We will first calculate the electric field at the position of the electron (see section on the "Infinite Plane" in Common Models of Electric Field). Observe that the surface charge density σ on the square plate is the charge on the plate divided by its surface area

$$\sigma = \frac{Q}{L^2} = \frac{+0.200 \times 10^{-9}\ \mathrm{C}}{(0.500\ \mathrm{m})^2} = 8.00 \times 10^{-10}\mathrm{C/m^2}\ . \tag{2.8.2}$$

The electric field is related to the surface charge density according to

$$\vec{E} = \frac{\sigma}{2\epsilon_0}\ \hat{j} = \frac{2.00 \times 10^{-7}\ \mathrm{C/m^2}}{2 \cdot 8.85 \times 10^{-12}\ \mathrm{C^2/N \cdot m^2}}\ \hat{j} = 45.2\,\mathrm{N/C}\ \hat{j}\ . \tag{2.8.3}$$

The field is directed upwards above the plate because the plate is positively charged. By Eq. 2.8.1, the acceleration of the particle is then

$$\vec{a} = \frac{q}{m}\vec{E} = \frac{-1.6 \times 10^{-19}\ \mathrm{C}}{9.11 \times 10^{-31}\ \mathrm{kg}}(45.2\,\mathrm{N/C}\ \hat{j}) = (-7.94 \times 10^{12}\ \mathrm{m/s^2})\ \hat{j}\ . \tag{2.8.4}$$

This result indicates the acceleration is downward toward the plate, as expected.

(a) According to the kinematic equations, the vertical position of the particle $y = y_0 + v_{y0}\Delta t + \frac{1}{2}a_y(\Delta t)^2$, where y_0 is the initial position and v_{y0} is the initial y-component of the velocity (zero in this case because the particle is starting at rest). Solving for the time of flight yields

$$\Delta t = \left(\frac{2y_0}{a_y}\right)^{1/2} = \left(\frac{2 \cdot 0.0100\ \mathrm{m}}{1.98 \times 10^{15}\ \mathrm{m/s^2}}\right)^{1/2} = 5.02 \times 10^{-8}\ \mathrm{s}\ . \tag{2.8.5}$$

(b) The vertical component of the velocity of the particle can be computed from the kinematic equation

$$v_y = v_{y0} + a_y\Delta t = 0\ \mathrm{m/s} + (-1.98 \times 10^{15}\ \mathrm{m/s^2})(3.17 \times 10^{-9}\ \mathrm{s}) = -3.98 \times 10^5\ \mathrm{m/s}\ . \tag{2.8.6}$$

CHECK

(a) The mass of the electron is very small and the acceleration is very high, so the travel time is very small, as expected.

(b) The final speed of the particle is large but plausible given the large value of the acceleration.

? Exercise 2.8.1

For the scenario as in Example 2.8.1, what is the minimum speed that the particle can be traveling in the x-direction to miss the landing on the plate?

Answer

The motion in the x- and y-directions are independent. In the x-direction, the kinematic equation is

$$x = x_0 + v_{x0}\Delta t + \frac{1}{2}a_x(\Delta t)^2 = v_{x0}\Delta t\ . \tag{2.8.7}$$

because the particle starts with $x = 0$ in the coordinate system in the Example $2.8.1$ and there is no force or acceleration acting in the x-direction. Solving for the x-component of the velocity yields

$$v_{x0} = \frac{x}{\Delta t} = \frac{0.5\ \mathrm{m}}{5.02\times 10^{-8}\ \mathrm{s}} = 9.36\times 10^{6}\ \mathrm{m/s} \tag{2.8.8}$$

This is a high velocity but plausible given the short time of travel before the electron hits the plate.

The principle illustrated by Example $2.8.1$ and Exercise $2.8.1$ is applied in *electrostatic precipitators* [1], which use the principle to pull unwanted particulate matter from the air. By charging the unwanted particles (e.g., fly ash from a coal-fired power plant) and passing the particles over charged plates, the particles can be filtered out of the air, as illustrated in Figure $2.8.3$.

Large electrostatic precipitators are used industrially to remove over 99% of the particles from stack gas emissions associated with the burning of coal and oil. Home precipitators, often in conjunction with the home heating and air conditioning system, are very effective in removing polluting particles, irritants, and allergens.

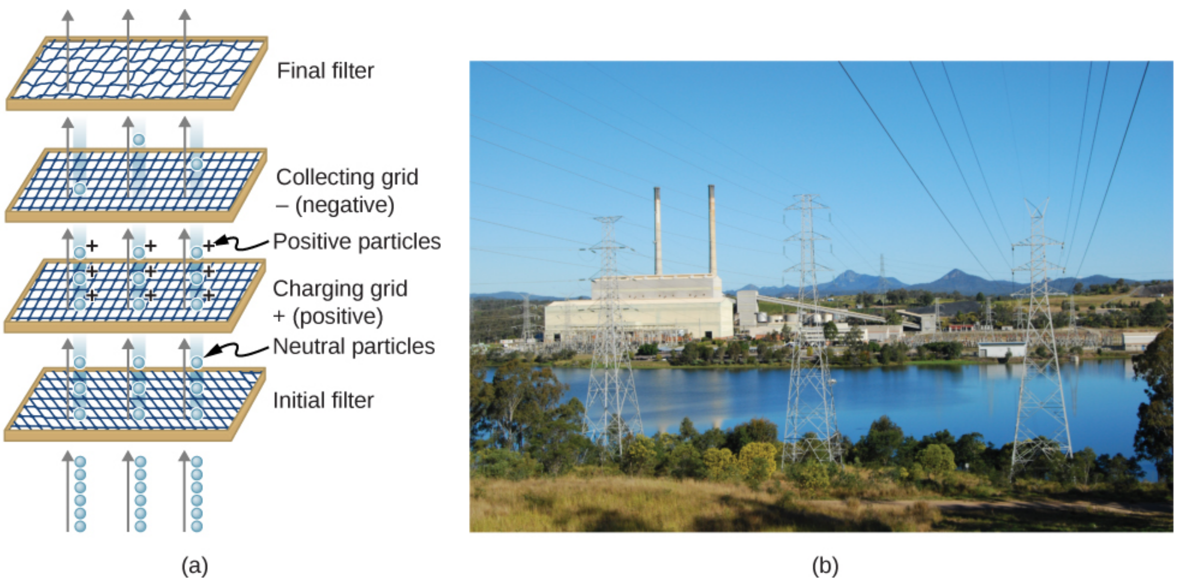


Figure $2.8.3$: (a) Schematic of an electrostatic precipitator. Air is passed through grids of opposite charge. The first grid charges airborne particles, while the second attracts and collects them. (b) The dramatic effect of electrostatic precipitators is seen by the absence of smoke from this power plant. (credit b: modification of work by “Cmdalgleish”/Wikimedia Commons) (Figure from Applications of Electrostatics).

Electric Field in the Ionosphere

Electric fields exist naturally in the atmosphere, and these fields can affect charged particles in the air. Atoms in the upper atmosphere are ionized by energy from the sun in a layer of the atmosphere called the *ionosphere*. (We will see later that the ionosphere can play an important role in the propagation of radio waves.) After ionization, the freed electrons are still attracted to the positive ions and so are not completely lost out into space. The resulting separation charge creates a *polarization electric field* in the ionosphere, sometimes also called an *ambipolar electric field*. We will explore some possible consequences of this field on the atmosphere in Example $2.8.2$.

Example $2.8.2$

The polarization electric field in a part of the ionosphere in altitude has been measured to be $1.09\times 10^{-6}\ \mathrm{N/C}$ in the upward direction (directed radially away from the surface of the earth) [2].

(a) What is the acceleration of a hydrogen ion in this region? What is the direction of the net acceleration?

(b) How long will it take for the ion to reach the speed of sound on the earth's surface (340 m/s) with this acceleration?

Solution

PLAN

To find the acceleration of the ion, we need to first calculate the net force. We will consider the effect of both the electric field and gravitational field on the ion as these fields are acting in opposite directions. If we assume that the acceleration is approximately constant, we can estimate the time to increase the ion's speed to a supersonic value.

SKETCH

We draw a free-body diagram for the ion, as seen in Figure 2.8.4 to see the relative effects of the forces. (The vector lengths come from the calculations below.)

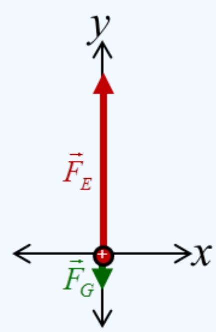


Figure 2.8.1: Free-body diagram for the hydrogen ion. (Ronald Kumon, CC-BY-SA 4.0)

CALCULATE

(a) A hydrogen ion is just a proton, which has mass of $m_p = 1.67 \times 10^{-27}$ kg and a charge of $q_p = e = +1.60 \times 10^{-19}$ C. The electric force on the ion is then

$$\overrightarrow{F_E} = q_p \vec{E} = (+1.6 \times 10^{-19}\ \mathrm{C})(+1.09 \times 10^{-6}\ \mathrm{N/C}\ \hat{j}) = 1.74 \times 10^{-25}\ \mathrm{N}\ \hat{j} \tag{2.8.9}$$

The gravitational force on the ion can be approximated by

$$\overrightarrow{F_G} = m_p \vec{g} = (1.67 \times 10^{-27}\ \mathrm{kg})(-9.8\ \mathrm{m/s^2}\ \hat{j}) = -1.64 \times 10^{-26}\ \mathrm{N}\ \hat{j}\,. \tag{2.8.10}$$

(This is an approximation because the acceleration of gravity decreases with altitude according to Newton's Law of Universal Gravitation; see Figure 1 in [3].)

The net force on the ion is then

$$\vec{F}_{\text{net}} = \vec{F}_E + \vec{F}_G = 1.74 \times 10^{-25}\ \mathrm{N}\ \hat{j} + -0.164 \times 10^{-25}\ \mathrm{N}\ \hat{j} = 1.58 \times 10^{-25}\ \mathrm{N}\ \hat{j}\,. \tag{2.8.11}$$

The acceleration of the ion is

$$\vec{a} = \frac{\vec{F}_{\text{net}}}{m_p} = \frac{1.58 \times 10^{-25}\ \mathrm{N}\ \hat{j}}{1.67 \times 10^{-27}\ \mathrm{kg}} = 94.6\ \mathrm{m/s^2}\ \hat{j}\,, \tag{2.8.12}$$

which is around 9.7 times the acceleration of gravity on the surface of the earth

(b) If we assume that the acceleration is approximately constant and the ion starts at rest, then the time to reach 340 m/s can be estimated from the kinematic equation $v_y = v_{y0} + a_y \Delta t$ such that

$$\Delta t = \frac{340\ \mathrm{m/s}}{94.6\ \mathrm{m/s^2}} = 3.59\ \mathrm{s}\,. \tag{2.8.13}$$

(The value is an estimate because speed of sound increases with altitude [4], so it would actually take longer to reach the local sound speed in the ionosphere.)

CHECK

The calculations indicate that the ambipolar electric field should create supersonic jets in the ionosphere. These *polar winds* have been experimentally observed (see news article [5] and visualizations [6]). The value of the acceleration is close to value of 10.6g given in [2], which is the acceleration that would be estimated by including the effect of the electric field only. Of course, we have simplified the problem considerably here (e.g., by neglecting the effect of the earth's magnetic field), and the actual physics are considerably more complicated; see [2] for further discussion.

References

1. Wikipedia contributors. Electrostatic precipitator [Internet]. Wikipedia, The Free Encyclopedia. Available from: https://en.wikipedia.org/w/index.php?title=Electrostatic_precipitator.
2. Collinson GA, Glocer A, Pfaff R, Barjatya A, Conway R, Breneman A, Clemmons J, Eparvier F, Mitchell R, Imber S, Akbari H, Davis L, Kavanagh A, Robertson E, Swanson D, Xu S, Miller J, Cameron T, Chornay D, Uribe P, Nguyen L, Clayton R, Graves N, Debchoudhury S, Valentine H, Ghalib A, and The Endurance Mission Team. Earth's ambipolar electrostatic field and its role in ion escape to space. Nature 2024; 632: 1021-1024. https://doi.org/10.1038/s41586-024-07480-3
3. Deng Y, Ridley AJ, Wang W. Effect of the altitudinal variation of the gravitational acceleration on the thermosphere simulation. Journal of Geophysical Research 2008; 113: A09302. https://doi.org/10.1029/2008JA013081
4. Longnonné P, Garcia R, Crespon F, Occhipinti G, Kherani A, Artru-Lambin J. Seismic waves in the ionosphere. Europhysics News 2006; 37(4): 11-14. https://doi.org/10.1051/epn:2006401
5. Hatfield M, Lense R. NASA discovers a long-sought global electric field on earth. NASA's Goddard Space Flight Center (28 August 2024). Available from: https://science.nasa.gov/science-research/heliophysics/nasa-discovers-long-sought-global-electric-field-on-earth/.
6. Young L. Discovering earth's third global energy field. NASA's Goddard Space Flight Center (28 August 2024). Available from: https://svs.gsfc.nasa.gov/14628. (Page contains many useful videos.)

2.9: Conclusion

In this chapter, we gained a better understanding of the properties of electric charge. We saw that an electric charge is associated with an electric field and that electric fields can exert electric forces on other electric charges. We also saw how to visualize the electric field through field-vector and field-line diagrams. Finally, we learned about various electric field models for common charge distributions under the assumption of continuously-distributed charge. We will use some of these electric field models in later chapters. These concepts covered in this chapter are fundamental to the study of electricity and will form the foundation of our future studies as we work to understand the components and properties of electric circuits.

In particular, we now know why two charged balloons in Figure $2.9.1$ repel each other. Both balloons must have a net charge (either an excess or deficit of charge) probably because they were rubbed on another object. The net charge on the left balloon creates an electric field around that balloon. The electric field of the left balloon then exerts an electric force on the right balloon. Similarly, by Coulomb's Law, we know that the right balloon also exerts an equal and opposite electric force on the left balloon. Because the balloons have the same sign charge, they are repelled.

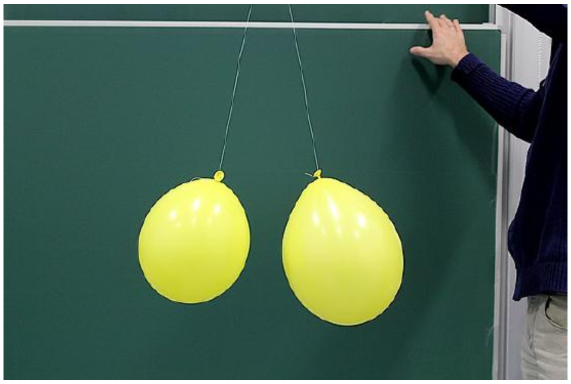


Figure $2.9.1$: Repulsive electric forces between two charged balloons. [1]

In this chapter we have focused on electric fields and forces. However, it is also possible to describe physical systems using an energy-based approach. In next chapter, we will investigate this alternative approach.

References

1. Wikimedia Commons contributors. File:Repulsive-electric-force-between-balloons.jpg [Internet]. Wikimedia Commons. (CC BY-SA 4.0, MikeRun)

2.10: The Electric Field (Summary)

Key Terms

Table 1. Key Terms for Chapter 2

Term	Definition
charging by induction	process by which an electrically charged object brought near a neutral object creates a charge separation in that object
conduction electron	electron that is free to move away from its atomic orbit
conductor	material that allows electrons to move separately from their atomic orbits; object with properties that allow charges to move about freely within it
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
coulomb	SI unit of electric charge
Coulomb force	another term for the electrostatic force
Coulomb's law	mathematical equation calculating the electrostatic force vector between two charged particles
dipole	two equal and opposite charges that are fixed close to each other
electric charge	physical property of an object that causes it to be attracted toward or repelled from another charged object; each charged object generates and is influenced by a force called an electric force
electric field	physical phenomenon created by a charge; it "transmits" a force between a two charges
electric force	noncontact force observed between electrically charged objects
electron	particle surrounding the nucleus of an atom and carrying the smallest unit of negative charge
electrostatic attraction	phenomenon of two objects with opposite charges attracting each other
electrostatic force	amount and direction of attraction or repulsion between two charged bodies; the assumption is that the source charges have no acceleration
electrostatic repulsion	phenomenon of two objects with like charges repelling each other
electrostatics	study of charged objects which are not in motion
field line	smooth, usually curved line that indicates the direction of the electric field
field line density	number of field lines per square meter passing through an imaginary area; its purpose is to indicate the field strength at different points in space
induced dipole	typically an atom, or a spherically symmetric molecule; a dipole created due to opposite forces displacing the positive and negative charges

Term	Definition
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
insulator	material that holds electrons securely within their atomic orbits
ion	atom or molecule with more or fewer electrons than protons
law of conservation of charge	net electric charge of a closed system is constant
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
neutron	neutral particle in the nucleus of an atom, with (nearly) the same mass as a proton
permittivity of vacuum	also called the permittivity of free space, and constant describing the strength of the electric force in a vacuum
polarization	slight shifting of positive and negative charges to opposite sides of an object
principle of superposition	useful fact that we can simply add up all of the forces due to charges acting on an object
proton	particle in the nucleus of an atom and carrying a positive charge equal in magnitude to the amount of negative charge carried by an electron
static electricity	buildup of electric charge on the surface of an object; the arrangement of the charge remains constant ("static")
superposition	concept that states that the net electric field of multiple source charges is the vector sum of the field of each source charge calculated individually
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

Key Equations

Table 2. Key Equations for Chapter 2

Description	Equation
Coulomb's law	$\overrightarrow{F_{12}}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{r}_{12}$
Superposition of electric forces	$\vec{F}(r) = \frac{1}{4\pi\varepsilon_0} Q \sum_{i=1}^{N} \frac{q_i}{r_i^2}\hat{r}_i$
Electric force due to an electric field	$\vec{F} = Q\vec{E}$
Electric field at point P	$\vec{E}(P) \equiv \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i$
Field of an infinite wire	$\vec{E}(z) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\hat{k}$
Field of an infinite plane	$\vec{E} = \frac{\sigma}{2\varepsilon_0}\hat{k}$

Summary

Electric Charge Model

- There are only two types of charge, which we call positive and negative. Like charges repel, unlike charges attract, and the force between charges decreases with the square of the distance.
- Protons carry the vast majority of positive charge in nature, whereas electrons carry the vast majority of negative charge. The electric charge of one electron is equal in magnitude and opposite in sign to the charge of one proton.
- An ion is an atom or molecule that has nonzero total charge due to having unequal numbers of electrons and protons.
- The SI unit for charge is the coulomb (C), with protons and electrons having charges of opposite sign but equal magnitude; the magnitude of this basic charge is $e \equiv 1.602 \times 10^{-19}\ \mathrm{C}$
- Both positive and negative charges exist in neutral objects and can be separated by bringing the two objects into physical contact; rubbing the objects together can remove electrons from the bonds in one object and place them on the other object, increasing the charge separation.
- For macroscopic objects, negatively charged means an excess of electrons, and positively charged means a depletion of electrons.
- The law of conservation of charge states that the net charge of a closed system is constant.

Conduction and Charging

- A conductor is a substance that allows charge to flow freely through its atomic structure.
- An insulator holds charge fixed in place.
- Polarization is the separation of positive and negative charges in a neutral object. Polarized objects have their positive and negative charges concentrated in different areas, giving them a charge distribution.

Electric Fields and Forces

- The electric field is an alteration of space caused by the presence of an electric charge. The electric field mediates the electric force between a source charge and a test charge. The electric field of a point charge is given by
  $$\vec{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r},$$
  where q is the value of the charge, r is the distance away from the charge, and $\hat{r}$ is the radial unit vector.
- The field is a vector; by definition, it points away from positive charges and toward negative charges
- Coulomb's law gives the magnitude of the force between point charges. It is
  $$\overrightarrow{F_{12}}(r) = \frac{1}{4\pi\varepsilon_0}\frac{q_1 q_2}{r_{12}^2}\hat{r}_{12},$$
  where q_1 and q_2 are two point charges separated by a distance **r**. This Coulomb force is a fundamental force because most charges are due to point-like particles. It is responsible for all electrostatic effects and underlies most macroscopic forces.

Electric Fields and Forces of Multiple Charges

- The electric field obeys the superposition principle, which states that the electric fields from multiple source charges can be summed to give the total field at a test location.
- The electric force also obeys the superposition principle.

Electric Field Diagrams

- Electric field diagrams assist in visualizing the field of a source charge.
- The magnitude of the field is proportional to the field line density.
- Field vectors are everywhere tangent to field lines.

Common Electric Field Models

- 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
- Electric field models can be constructed for charge distributions like the finite line segment, infinite line, ring, disk, and infinite plane.

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).

2.11: The Electric Field (Exercises)

Conceptual Questions

Electric Charge Model

1. There are very large numbers of charged particles in most objects. Why, then, don't most objects exhibit static electricity?

2. Why do most objects tend to contain nearly equal numbers of positive and negative charges?

3. A positively charged rod attracts a small piece of cork.

(a) Can we conclude that the cork is negatively charged?

(b) The rod repels another small piece of cork. Can we conclude that this piece is positively charged?

4. Two bodies attract each other electrically. Do they both have to be charged? Answer the same question if the bodies repel one another.

5. How would you determine whether the charge on a particular rod is positive or negative?

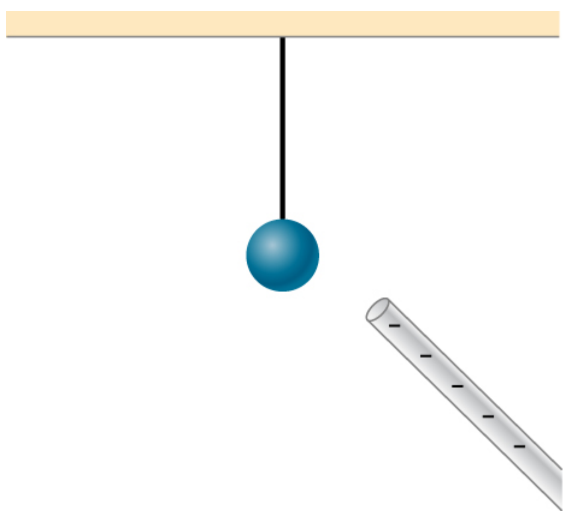
Conduction and Charging

6. An eccentric inventor attempts to levitate a cork ball by wrapping it with foil and placing a large negative charge on the ball and then putting a large positive charge on the ceiling of his workshop. Instead, while attempting to place a large negative charge on the ball, the foil flies off. Explain.

7. When a glass rod is rubbed with silk, it becomes positive and the silk becomes negative—yet both attract dust. Does the dust have a third type of charge that is attracted to both positive and negative? Explain.

8. Why does a car always attract dust right after it is polished? (Note that car wax and car tires are insulators.)

9. Does the uncharged conductor shown below experience a net electric force?



10. While walking on a rug, a person frequently becomes charged because of the rubbing between his shoes and the rug. This charge then causes a spark and a slight shock when the person gets close to a metal object. Why are these shocks so much more common on a dry day?

11. Compare charging by conduction to charging by induction.

12. Small pieces of tissue are attracted to a charged comb. Soon after sticking to the comb, the pieces of tissue are repelled from it. Explain.

13. Trucks that carry gasoline often have chains dangling from their undercarriages and brushing the ground. Why?

14. Why do electrostatic experiments work so poorly in humid weather?

15. Why do some clothes cling together after being removed from the clothes dryer? Does this happen if they're still damp?

16. Can induction be used to produce charge on an insulator?

17. Suppose someone tells you that rubbing quartz with cotton cloth produces a third kind of charge on the quartz. Describe what you might do to test this claim.

18. A handheld copper rod does not acquire a charge when you rub it with a cloth. Explain why.

19. Suppose you place a charge **q** near a large metal plate.

(a) If **q** is attracted to the plate, is the plate necessarily charged?

(b) If **q** is repelled by the plate, is the plate necessarily charged?

Electric Fields and Forces

20. Would defining the charge on an electron to be positive have any effect on Coulomb's law?

21. An atomic nucleus contains positively charged protons and uncharged neutrons. Since nuclei do stay together, what must we conclude about the forces between these nuclear particles?

22. Is the force between two fixed charges influenced by the presence of other charges?

23. When measuring an electric field, could we use a negative rather than a positive test charge?

24. During fair weather, the electric field due to the net charge on Earth points downward. Is Earth charged positively or negatively?

Electric Fields and Forces with Multiple Charges

25. If the electric field at a point on the line between two charges is zero, what do you know about the charges?

26. Two charges lie along the **x**-axis. Is it true that the net electric field always vanishes at some point (other than infinity) along the **x**-axis?

Calculating Electric Fields of Charge Distributions

27. Give a plausible argument as to why the electric field outside an infinite charged sheet is constant.

28. Compare the electric fields of an infinite sheet of charge, an infinite, charged conducting plate, and infinite, oppositely charged parallel plates.

29. Describe the electric fields of an infinite charged plate and of two infinite, charged parallel plates in terms of the electric field of an infinite sheet of charge.

30. A negative charge is placed at the center of a ring of uniform positive charge. What is the motion (if any) of the charge? What if the charge were placed at a point on the axis of the ring other than the center?

Electric Field Diagrams

31. If a point charge is released from rest in a uniform electric field, will it follow a field line? Will it do so if the electric field is not uniform?

32. Under what conditions, if any, will the trajectory of a charged particle not follow a field line?

33. How would you experimentally distinguish an electric field from a gravitational field?

34. A representation of an electric field shows 10 field lines perpendicular to a square plate. How many field lines should pass perpendicularly through the plate to depict a field with twice the magnitude?

35. What is the ratio of the number of electric field lines leaving a charge 10q and a charge q?

Problems

Electric Charge Model

37. Common static electricity involves charges ranging from nanocoulombs to microcoulombs.

(a) How many electrons are needed to form a charge of −2.00 nC?

(b) How many electrons must be removed from a neutral object to leave a net charge of $\mathbf{0.500 \mu C}$?

38. If 1.80×10^{20} electrons move through a pocket calculator during a full day's operation, how many coulombs of charge moved through it?

39. To start a car engine, the car battery moves 3.75×10^{21} electrons through the starter motor. How many coulombs of charge were moved?

40. A certain lightning bolt moves 40.0 C of charge. How many fundamental units of charge is this?

41. A 2.5-g copper penny is given a charge of $-2.0 \times 10^{-9} C$.

(a) How many excess electrons are on the penny?

(b) By what percent do the excess electrons change the mass of the penny?

42. A 2.5-g copper penny is given a charge of $4.0 \times 10^{-9} C$.

(a) How many electrons are removed from the penny?

(b) If no more than one electron is removed from an atom, what percent of the atoms are ionized by this charging process?

Conduction and Charging

43. Suppose a speck of dust in an electrostatic precipitator has 1.0000×10^{12} protons in it and has a net charge of −5.00 nC (a very large charge for a small speck). How many electrons does it have?

44. An amoeba has 1.00×10^{16} protons and a net charge of 0.300 pC.

(a) How many fewer electrons are there than protons?

(b) If you paired them up, what fraction of the protons would have no electrons?

45. A 50.0-g ball of copper has a net charge of **2.00μC**. What fraction of the copper's electrons has been removed? (Each copper atom has 29 protons, and copper has an atomic mass of 63.5.)

46. What net charge would you place on a 100-g piece of sulfur if you put an extra electron on 1 in 10^{12} of its atoms? (Sulfur has an atomic mass of 32.1 u.)

47. How many coulombs of positive charge are there in 4.00 kg of plutonium, given its atomic mass is 244 and that each plutonium atom has 94 protons?

Electric Fields and Forces

48. Two point particles with charges **+3μC** and **+5μC** are held in place by 3-N forces on each charge in appropriate directions. (a) Draw a free-body diagram for each particle. (b) Find the distance between the charges.

49. Two charges **+3μC** and **+12μC** are fixed 1 m apart, with the second one to the right. Find the magnitude and direction of the net force on a −2-nC charge when placed at the following locations:

(a) halfway between the two

(b) half a meter to the left of the **+3μC** charge

(c) half a meter above the **+12μC** charge in a direction perpendicular to the line joining the two fixed charges

50. In a salt crystal, the distance between adjacent sodium and chloride ions is $2.82 \times 10^{-10} m$. What is the force of attraction between the two singly charged ions?

51. Protons in an atomic nucleus are typically $10^{-15} m$ apart. What is the electric force of repulsion between nuclear protons?

52. Suppose Earth and the Moon each carried a net negative charge **−Q**. Approximate both bodies as point masses and point charges.

(a) What value of **Q** is required to balance the gravitational attraction between Earth and the Moon?

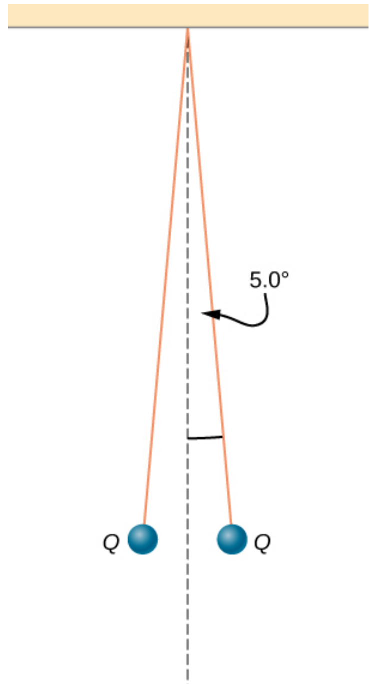
(b) Does the distance between Earth and the Moon affect your answer? Explain.

(c) How many electrons would be needed to produce this charge?

53. Point charges $q_1 = 50\mu C$ and $q_2 = -25\mu C$ are placed 1.0 m apart. What is the force on a third charge $q_3 = 20\mu C$ placed midway between q_1 and q_2?

54. Where must q_3 of the preceding problem be placed so that the net force on it is zero?

55. Two small balls, each of mass 5.0 g, are attached to silk threads 50 cm long, which are in turn tied to the same point on the ceiling, as shown below. When the balls are given the same charge **Q**, the threads hang at **5.0°** to the vertical, as shown below. What is the magnitude of **Q**? What are the signs of the two charges?

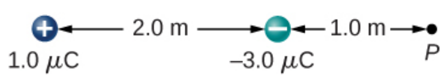


56. Point charges $Q_1 = 2.0\mu C$ and $Q_2 = 4.0\mu C$ are located at $\vec{r_1} = (4.0\hat{i} - 2.0\hat{j} + 5.0\hat{k})m$ and $\vec{r_2} = (8.0\hat{i} + 5.0\hat{j} - 9.0\hat{k})m$. What is the force of Q_2 on Q_1?

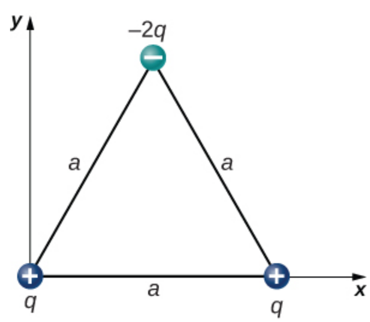
57. The net excess charge on two small spheres (small enough to be treated as point charges) is **Q**. Show that the force of repulsion between the spheres is greatest when each sphere has an excess charge **Q/2**. Assume that the distance between the spheres is so large compared with their radii that the spheres can be treated as point charges.

58. Two small, identical conducting spheres repel each other with a force of 0.050 N when they are 0.25 m apart. After a conducting wire is connected between the spheres and then removed, they repel each other with a force of 0.060 N. What is the original charge on each sphere?

59. A charge **q=2.0μC** is placed at the point **P** shown below. What is the force on **q**?

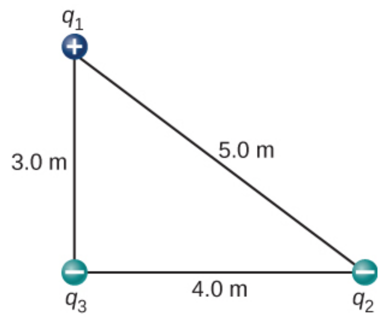


60. What is the net electric force on the charge located at the lower right-hand corner of the triangle shown here?

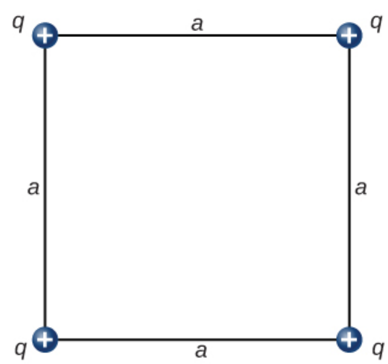


61. Two fixed particles, each of charge $5.0 \times 10^{-6} C$, are 24 cm apart. What force do they exert on a third particle of charge $-2.5 \times 10^{-6} C$ that is 13 cm from each of them?

62. The charges $q_1 = 2.0 \times 10^{-7} C, q_2 = -4.0 \times 10^{-7} C$, and $q_3 = -1.0 \times 10^{-7} C$ are placed at the corners of the triangle shown below. What is the force on q_1?



63. What is the force on the charge q at the lower-right-hand corner of the square shown here?



64. Point charges $q_1 = 10\mu C$ and $q_2 = -30\mu C$ are fixed at $r_1 = (3.0\hat{i} - 4.0\hat{j})m$ and $r_2 = (9.0\hat{i} + 6.0\hat{j})m$. What is the force of q_2 on q_1?

65. A particle of charge $2.0 \times 10^{-8} C$ experiences an upward force of magnitude $4.0 \times 10^{-6} N$ when it is placed in a particular point in an electric field.

(a) What is the electric field at that point?

(b) If a charge $q = -1.0 \times 10^{-8} C$ is placed there, what is the force on it?

66. On a typical clear day, the atmospheric electric field points downward and has a magnitude of approximately 100 N/C. Compare the gravitational and electric forces on a small dust particle of mass $2.0 \times 10^{-15} g$ that carries a single electron charge. What is the acceleration (both magnitude and direction) of the dust particle?

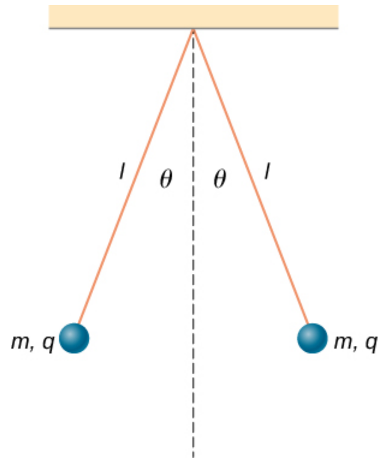
67. Consider an electron that is 10^{-10} m from an alpha particle (q=3.2×10−19C).(q=3.2×10−19C).

(a) What is the electric field due to the alpha particle at the location of the electron?

(b) What is the electric field due to the electron at the location of the alpha particle?

(c) What is the electric force on the alpha particle? On the electron?

68. Each the balls shown below carries a charge q and has a mass m. The length of each thread is l, and at equilibrium, the balls are separated by an angle 2θ. How does θ vary with **q** and **l**? Show that θ satisfies \(\displaystyle sin(θ)^2tan(θ)=\frac{q^2}{16πε0gl^2m\).



69. What is the electric field at a point where the force on a $-2.0\times10^{-6}-C$ charge is $(4.0\hat{i}-6.0\hat{j})\times10^{-6}N$?

70. A proton is suspended in the air by an electric field at the surface of Earth. What is the strength of this electric field?

71. The electric field in a particular thundercloud is $2.0\times10^{5}N/C$. What is the acceleration of an electron in this field?

72. A small piece of cork whose mass is 2.0 g is given a charge of $5.0\times10^{-7}C$. What electric field is needed to place the cork in equilibrium under the combined electric and gravitational forces?

73. If the electric field is 100 N/C at a distance of 50 cm from a point charge q, what is the value of q?

74. What is the electric field of a proton at the first Bohr orbit for hydrogen $(r=5.29\times10^{-11}m)$? What is the force on the electron in that orbit?

75. (a) What is the electric field of an oxygen nucleus at a point that is $10^{-10}m$ from the nucleus?

(b) What is the force this electric field exerts on a second oxygen nucleus placed at that point?

Electric Field and Forces with Multiple Charges

76. Two point charges, $q_1=2.0\times10^{-7}C$ and $q_2=-6.0\times10^{-8}C$, are held 25.0 cm apart.

(a) What is the electric field at a point 5.0 cm from the negative charge and along the line between the two charges?

(b)What is the force on an electron placed at that point?

77. Point charges $q_1=50\mu C$ and $q_2=-25\mu C$ are placed 1.0 m apart.

(a) What is the electric field at a point midway between them?

(b) What is the force on a charge $q_3=20\mu C$ situated there?

78. Can you arrange the two point charges $q_1=-2.0\times10^{-6}C$ and $q_2=4.0\times10^{-6}C$ along the x-axis so that $E=0$ at the origin?

79. Point charges $q_1=q_2=4.0\times10^{-6}C$ are fixed on the x-axis at $x=-3.0m$ and $x=3.0m$. What charge q must be placed at the origin so that the electric field vanishes at x=0,y=3.0m?

Common Electric Fields Models

80. A thin conducting plate 1.0 m on the side is given a charge of $-2.0\times10^{-6}C$. An electron is placed 1.0 cm above the center of the plate. What is the acceleration of the electron?

81. Calculate the magnitude and direction of the electric field 2.0 m from a long wire that is charged uniformly at $\lambda=4.0\times10^{-6}C/m$.

82. Two thin conducting plates, each 25.0 cm on a side, are situated parallel to one another and 5.0 mm apart. If 10^{11} electrons are moved from one plate to the other, what is the electric field between the plates?

85. Two thin parallel conducting plates are placed 2.0 cm apart. Each plate is 2.0 cm on a side; one plate carries a net charge of **8.0μC,** and the other plate carries a net charge of **−8.0μC**. What is the charge density on the inside surface of each plate?

What is the electric field between the plates?

86. A thin conducting plate 2.0 m on a side is given a total charge of **−10.0μC**.

(a) What is the electric field **1.0cm** above the plate?

(b) What is the force on an electron at this point?

(c) Repeat these calculations for a point 2.0 cm above the plate.

(d) When the electron moves from 1.0 to 2,0 cm above the plate, how much work is done on it by the electric field?

88. Charge is distributed along the entire **x**-axis with uniform density **λ**. How much work does the electric field of this charge distribution do on an electron that moves along the **y**-axis from $y=a$ to $y=b$?

89. Charge is distributed along the entire **x**-axis with uniform density λ_x and along the entire **y**-axis with uniform density λ_y. Calculate the resulting electric field at

(a) $\vec{r}=a\hat{i}+b\hat{j}$ and

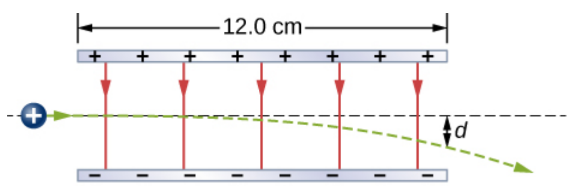
(b) $\vec{r}=c\hat{k}$.

91. A proton moves in the electric field $\vec{E}=200\hat{i}N/C$. (a) What are the force on and the acceleration of the proton? (b) Do the same calculation for an electron moving in this field.

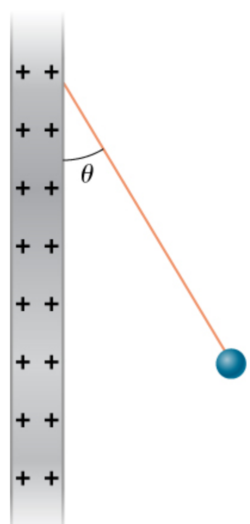
92. An electron and a proton, each starting from rest, are accelerated by the same uniform electric field of 200 N/C. Determine the distance and time for each particle to acquire a kinetic energy of $3.2\times10^{-16}J$.

93. A spherical water droplet of radius 25 μm carries an excess 250 electrons. What vertical electric field is needed to balance the gravitational force on the droplet at the surface of the earth?

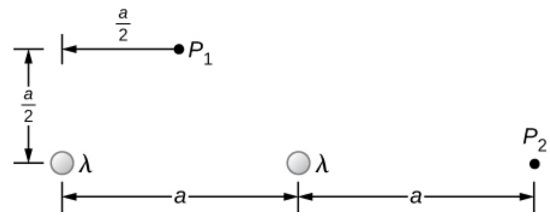
94. A proton enters the uniform electric field produced by the two charged plates shown below. The magnitude of the electric field is $4.0\times10^5N/C$, and the speed of the proton when it enters is $1.5\times10^7m/s$. What distance **d** has the proton been deflected downward when it leaves the plates?



95. Shown below is a small sphere of mass 0.25 g that carries a charge of $9.0\times10^{-10}C$. The sphere is attached to one end of a very thin silk string 5.0 cm long. The other end of the string is attached to a large vertical conducting plate that has a charge density of $30\times10^{-6}C/m^2$. What is the angle that the string makes with the vertical?



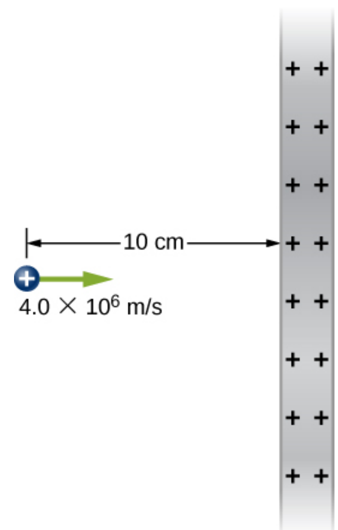
96. Two infinite rods, each carrying a uniform charge density λ, are parallel to one another and perpendicular to the plane of the page. (See below.) What is the electrical field at P_1? At P_2?



98. From a distance of 10 cm, a proton is projected with a speed of $v = 4.0 \times 10^6 m/s$ directly at a large, positively charged plate whose charge density is $\sigma = 2.0 \times 10^{-5} C/m^2$.. (See below.)

(a) Does the proton reach the plate?

(b) If not, how far from the plate does it turn around?



99. A particle of mass m and charge $-q$ moves along a straight line away from a fixed particle of charge **Q**. When the distance between the two particles is r_0, $-q$ is moving with a speed v_0.

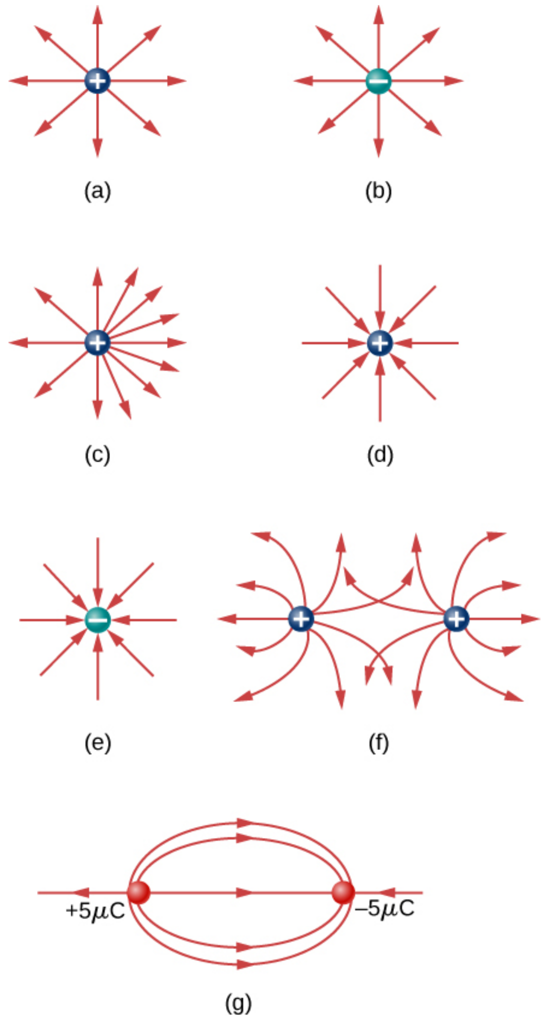
(a) Use the work-energy theorem to calculate the maximum separation of the charges.

(b) What do you have to assume about v_0 to make this calculation?

(c) What is the minimum value of v_0 such that $-q$ escapes from **Q**?

Electric Field Diagrams

100. Which of the following electric field lines are incorrect for point charges? Explain why.



101. In this exercise, you will practice drawing electric field lines. Make sure you represent both the magnitude and direction of the electric field adequately. Note that the number of lines into or out of charges is proportional to the charges.

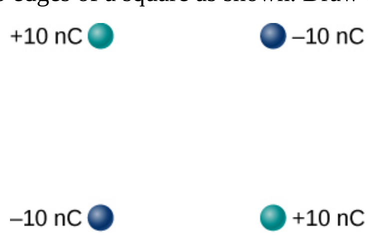
(a) Draw the electric field lines map for two charges **+20μC** and **−20μC** situated 5 cm from each other.

(b) Draw the electric field lines map for two charges **+20μC** and **+20μC** situated 5 cm from each other.

(c) Draw the electric field lines map for two charges **+20μC** and **−30μC** situated 5 cm from each other.

102. Draw the electric field for a system of three particles of charges **+1μC, +2μC** and **−3μC** fixed at the corners of an equilateral triangle of side 2 cm.

103. Two charges of equal magnitude but opposite sign make up an electric dipole. A quadrupole consists of two electric dipoles that are placed anti-parallel at two edges of a square as shown. Draw the electric field of the charge distribution.

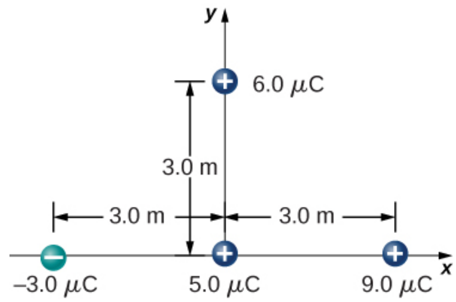


104. Suppose the electric field of an isolated point charge decreased with distance as $1/r^{2+\delta}$ rather than as $1/r^2$. Show that it is then impossible to draw continuous field lines so that their number per unit area is proportional to **E**.

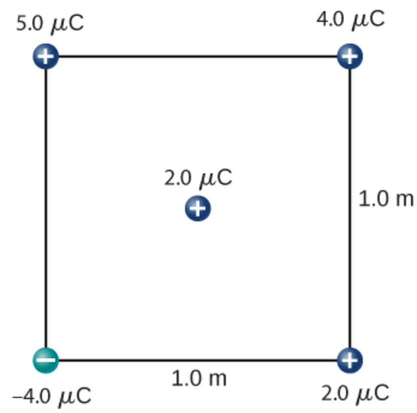
Additional Problems

108. Point charges $q_1 = 2.0\mu C$ and $q_1 = 4.0\mu C$ are located at $r_1 = (4.0\hat{i} - 2.0\hat{j} + 2.0\hat{k})m$ and $r_2 = (8.0\hat{i} + 5.0\hat{j} - 9.0\hat{k})m$. What is the force of q_2 on q_1?

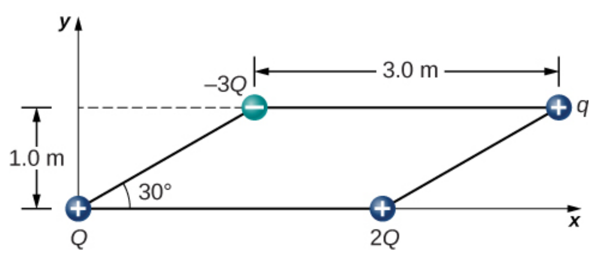
109. What is the force on the **5.0-μC** charges shown below?



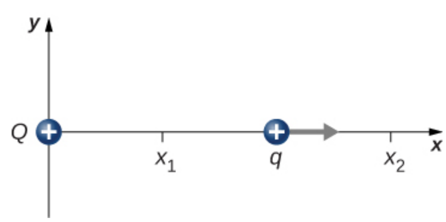
110. What is the force on the **2.0-μC** charge placed at the center of the square shown below?



111. Four charged particles are positioned at the corners of a parallelogram as shown below. If $q = 5.0\mu C$ and $Q = 8.0\mu C$, what is the net force on **q**?

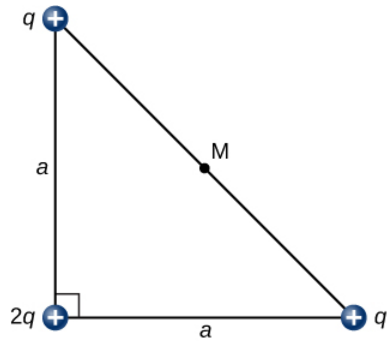


112. A charge **Q** is fixed at the origin and a second charge **q** moves along the **x**-axis, as shown below. How much work is done on **q** by the electric force when **q** moves from x_1 to x_2?

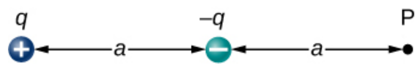


113. A charge **q=−2.0μC** is released from rest when it is 2.0 m from a fixed charge $Q = 6.0\mu C$. What is the kinetic energy of **q** when it is 1.0 m from **Q**?

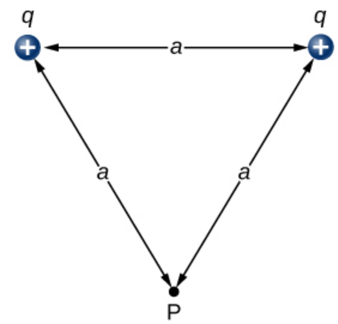
114. What is the electric field at the midpoint M of the hypotenuse of the triangle shown below?



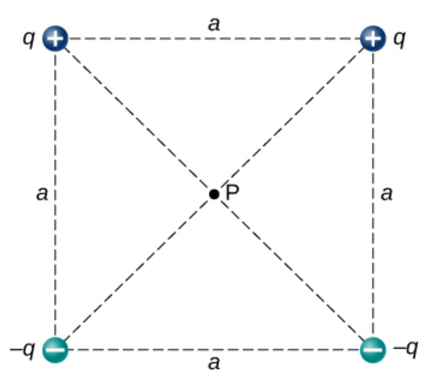
115. Find the electric field at P for the charge configurations shown below.



(a)

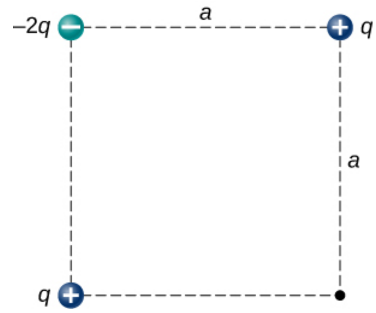


(b)

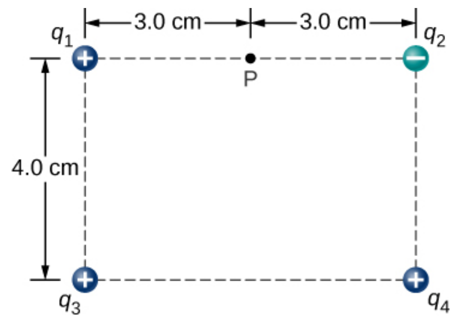


(c)

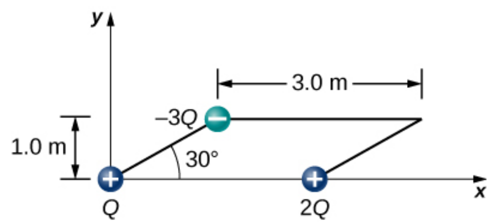
116. (a) What is the electric field at the lower-right-hand corner of the square shown below? (b) What is the force on a charge **q** placed at that point?



117. Point charges are placed at the four corners of a rectangle as shown below: $q_1 = 2.0 \times 10^{-6} C, q_2 = -2.0 \times 10^{-6} C, q_3 = 4.0 \times 10^{-6} C,$ and $q_4 = 1.0 \times 10^{-6} C$. What is the electric field at **P**?



118. Three charges are positioned at the corners of a parallelogram as shown below. (a) If $Q = 8.0 \mu C$, what is the electric field at the unoccupied corner? (b) What is the force on a **5.0-μC** charge placed at this corner?



119. A positive charge **q** is released from rest at the origin of a rectangular coordinate system and moves under the influence of the electric field $\vec{E} = E_0(1 + x/a)\hat{i}$. What is the kinetic energy of **q** when it passes through $x = 3a$?

120. A particle of charge $-q$ and mass **m** is placed at the center of a uniformly charged ring of total charge **Q** and radius **R**. The particle is displaced a small distance along the axis perpendicular to the plane of the ring and released. Assuming that the particle is constrained to move along the axis, show that the particle oscillates in simple harmonic motion with a frequency $f = \frac{1}{2\pi}\sqrt{\frac{qQ}{4\pi\varepsilon_0 m R^3}}$.

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).

2.12: The Electric Field (Answers)

Note: Answers are provided for only the odd-numbered questions.

Conceptual Questions

1. There are mostly equal numbers of positive and negative charges present, making the object electrically neutral.

3. a. yes; b. yes

5. Take an object with a known charge, either positive or negative, and bring it close to the rod. If the known charged object is positive and it is repelled from the rod, the rod is charged positive. If the positively charged object is attracted to the rod, the rod is negatively charged.

7. No, the dust is attracted to both because the dust particle molecules become polarized in the direction of the silk.

9. Yes, polarization charge is induced on the conductor so that the positive charge is nearest the charged rod, causing an attractive force.

11. Charging by conduction is charging by contact where charge is transferred to the object. Charging by induction first involves producing a polarization charge in the object and then connecting a wire to ground to allow some of the charge to leave the object, leaving the object charged.

13. This is so that any excess charge is transferred to the ground, keeping the gasoline receptacles neutral. If there is excess charge on the gasoline receptacle, a spark could ignite it.

15. The dryer charges the clothes. If they are damp, the presence of water molecules suppresses the charge.

17. There are only two types of charge, attractive and repulsive. If you bring a charged object near the quartz, only one of these two effects will happen, proving there is not a third kind of charge.

19. a. No, since a polarization charge is induced. b. Yes, since the polarization charge would produce only an attractive force.

21. The force holding the nucleus together must be greater than the electrostatic repulsive force on the protons.

23. Either sign of the test charge could be used, but the convention is to use a positive test charge.

25. The charges are of the same sign.

27. At infinity, we would expect the field to go to zero, but because the sheet is infinite in extent, this is not the case. Everywhere you are, you see an infinite plane in all directions.

29. The infinite charged plate would have $E = \frac{\sigma}{2\varepsilon_0}$ everywhere. The field would point toward the plate if it were negatively charged and point away from the plate if it were positively charged. The electric field of the parallel plates would be zero between them if they had the same charge, and **E** would be $E = \frac{\sigma}{\varepsilon_0}$ everywhere else. If the charges were opposite, the situation is reversed, zero outside the plates and $E = \frac{\sigma}{\varepsilon_0}$ between them.

31. yes; no

33. At the surface of Earth, the gravitational field is always directed in toward Earth's center. An electric field could move a charged particle in a different direction than toward the center of Earth. This would indicate an electric field is present.

35. 10

Problems

37. a. $2.00 \times 10^{-9} C(\frac{1}{1.602 \times 10^{-19}} e/C) = 1.248 \times 10^{10} electrons2$;

b. $0.500 \times 10^{-6} C(\frac{1}{1.602 \times 10^{-19}} e/C) = 3.121 \times 10^{12} electrons$

39. $\frac{3.750 \times 10^{21} e}{6.242 \times 10^{18} e/C} = -600.8C$

41. a. $2.0\times 10^{-9}C(6.242\times 10^{18}e/C)=1.248\times 10^{10}e$;

b. $9.109\times 10^{-31}kg(1.248\times 10^{10}e)=1.137\times 10^{-20}kg$, $\frac{1.137\times 10^{-20}kg}{2.5\times 10^{-3}kg}=4.548\times 10^{-18}$ or 4.545×10^{-16}

43. $5.00\times 10^{-9}C(6.242\times 10^{18}e/C)=3.121\times 10^{10}e; 3.121\times 10^{10}e+1.0000\times 10^{12}e=1.0312\times 10^{12}e$.

45. atomic mass of copper atom times $1u=1.055\times 10^{-25}kg$; number of copper atoms = $4.739\times 10^{23}atoms$; number of electrons equals 29 times number of atoms or $1.374\times 10^{25}electrons$; $\frac{2.00\times 10^{-6}C(6.242\times 10^{18}e/C)}{1.374\times 10^{25}e}=9.083\times 10^{-13}$ or 9.083×10^{-11}.

47. $244.00u(1.66\times 10^{-27}kg/u)=4.050\times 10^{-25}kg$; $\frac{4.00kg}{4.050\times 10^{-25}kg}=9.877\times 10^{24}atoms$ $9.877\times 10^{24}(94)=9.284\times 10^{26}protons$ $9.284\times 10^{26}protons; 9.284\times 10^{26}(1.602\times 10^{-19}C/p)=1.487\times 10^{8}C$

49. a. charge 1 is $3\mu C$; charge 2 is $12\mu C$, $F_{31}=2.16\times 10^{-4}N$ to the left,

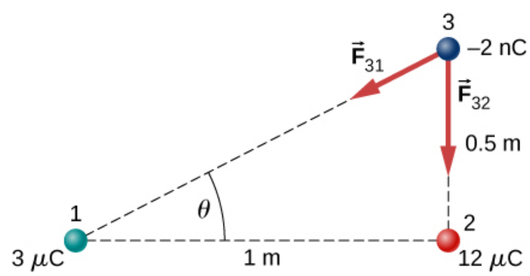
$F_{32}=8.63\times 10^{-4}N$ to the right,

$F_{net}=6.47\times 10^{-4}N$ to the right;

b. $F_{31}=2.16\times 10^{-4}N$ to the right,

$F_{32}=9.59\times 10^{-5}N$ to the right,

$F_{net}=3.12\times 10^{-4}N$ to the right,



c. $\vec{F}_{31x}=-2.76\times 10^{-5}N\hat{i}$,

$\vec{F}_{31y}=-1.38\times 10^{-5}N\hat{j}$,

$\vec{F}_{32y}=-8.63\times 10^{-4}N\hat{j}$,

$\vec{F}_{net}=-2.76\times 10^{-5}N\hat{i}-8.77\times 10^{-4}N\hat{j}$

51. $F=230.7N$

53. $F=53.94N$

55. The tension is $T=0.049N$. The horizontal component of the tension is $0.0043N$

$d=0.088m, q=6.1\times 10^{-8}C$.

The charges can be positive or negative, but both have to be the same sign.

57. Let the charge on one of the spheres be **rQ**, where r is a fraction between 0 and 1. In the numerator of Coulomb's law, the term involving the charges is $rQ(1-r)Q$. This is equal to $(r-r^2)Q^2$. Finding the maximum of this term gives $1-2r=0\Rightarrow r=\frac{1}{2}$

59. Define right to be the positive direction and hence left is the negative direction, then $F=-0.05N$

61. The particles form triangle of sides 13, 13, and 24 cm. The **x**-components cancel, whereas there is a contribution to the y-component from both charges 24 cm apart. The y-axis passing through the third charge bisects the 24-cm line, creating two right triangles of sides 5, 12, and 13 cm. $F_y=2.56N$ in the negative **y**-direction since the force is attractive. The net force from both charges is $\vec{F}_{net}=-5.12N\hat{j}$

63. The diagonal is $\sqrt{2}a$ and the components of the force due to the diagonal charge has a factor $cos\theta = \frac{1}{\sqrt{2}}$;
$\vec{F}_{net} = [k\frac{q^2}{a^2} + k\frac{q^2}{2a^2}\frac{1}{\sqrt{2}}]\hat{i} - [k\frac{q^2}{a^2} + k\frac{q^2}{2a^2}\frac{1}{\sqrt{2}}]\hat{j}$

65. $a.\,E = 2.0\times10^{-2}\frac{N}{C}$;

$b.\,F = 2.0\times10^{-19}N$

67. a. $E = 2.88\times10^{11}N/C$;

b. $E = 1.44\times10^{11}N/C$;

c. $F = 4.61\times10^{-8}N$ on alpha particl

$F = 4.61\times10^{-8}N$ on electron

69. $E = (-2.0\hat{i} + 3.0\hat{j})N$

71. $F = 3.204\times10^{-14}N$,

$a = 3.517\times10^{16}m/s^2$

73. $q = 2.78\times10^{-9}C$

75. a. $E = 1.15\times10^{12}N/C$;

b. $F = 1.47\times10^{-6}N$

77. If the q_2 is to the right of q_1, the electric field vector from both charges point to the right.

a. $E = 2.70\times10^{6}N/C$;

b. $F = 54.0N$

79. There is **45°** right triangle geometry. The **x**-components of the electric field at $y = 3m$ cancel. The **y**-components give $E(y=3m) = 2.83\times10^{3}N/C$.

At the origin we have a a negative charge of magnitide $q = -2.83\times10^{-6}C$

81. $\vec{E}(z) = 3.6\times10^{4}N\hat{k}$

85. $\sigma = 0.02C/m^2\; E = 2.26\times10^{9}N/C$

89. a. $\vec{E}(\vec{r}) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda_x}{a}\hat{i} + \frac{1}{4\pi\varepsilon_0}\frac{2\lambda_y}{b}\hat{j}$;

b. $\frac{1}{4\pi\varepsilon_0}\frac{2(\lambda_x+\lambda_y)}{c}\hat{k}$

91. a. $\vec{F} = 3.2\times10^{-17}N\hat{i}$,

$\vec{a} = 1.92\times10^{10}m/s^2\hat{i}$;

b. $\vec{F} = -3.2\times10^{-17}N\hat{i}$,

$\vec{a} = -3.51\times10^{13}m/s^2\hat{i}$

93. $m = 6.5\times10^{-11}kg$,

$E = 1.6\times10^{7}N/C$

95. $E = 1.70\times10^{6}N/C$,

$F = 1.53\times10^{-3}NTcos\theta = mgTsin\theta = qE$,

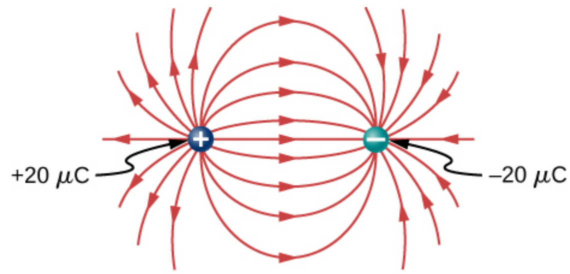
$tan\theta = 0.62 \Rightarrow \theta = 32.0^\circ$,

This is independent of the length of the string.

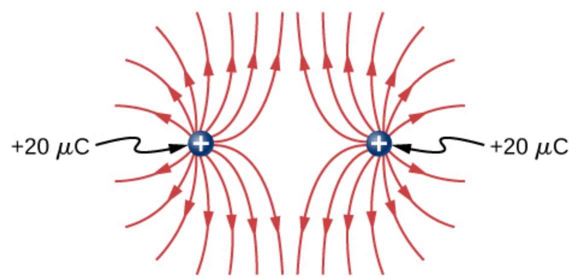
99. a. $W=\frac{1}{2}m(v^2-v_0^2), \frac{Qq}{4\pi\varepsilon_0}(\frac{1}{r}-\frac{1}{r_0})=\frac{1}{2}m(v^2-v_0^2)\Rightarrow r_0-r=\frac{4\pi\varepsilon_0}{Qq}\frac{1}{2}rr_0m(v^2-v_0^2)$;

b. r_0-r is negative; therefore, $v_0>v, r\to\infty$, and $v\to 0$: $\frac{Qq}{4\pi\varepsilon_0}(-\frac{1}{r_0})=-\frac{1}{2}mv_0^2\Rightarrow v_0=\sqrt{\frac{Qq}{2\pi\varepsilon_0 mr_0}}$

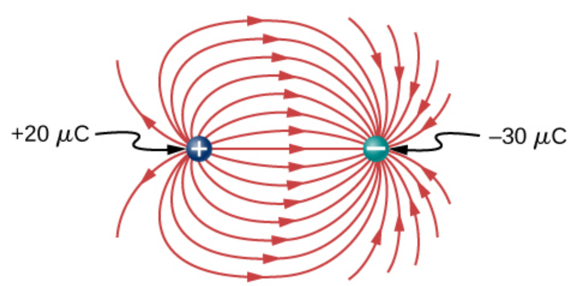
101.



(a)

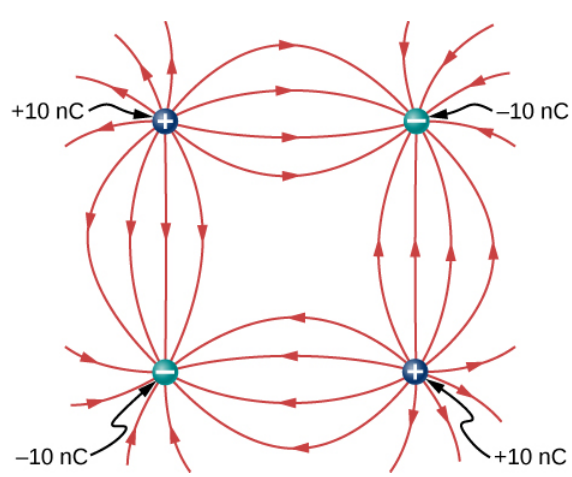


(b)



(c)

103.



Additional Problems

109. $\vec{F}_{net} = [-8.99\times10^9\frac{3.0\times10^{-6}(5.0\times10^{-6})}{(3.0m)^2} - 8.99\times10^9\frac{9.0\times10^{-6}(5.0\times10^{-6})}{(3.0m)^2}]\hat{i}, -8.99$ $\times10^9\frac{6.0\times10^{-6}(5.0\times10^{-6})}{(3.0m)^2}\hat{j} = -0.06N\hat{i} - 0.03N\hat{j}$

111. Charges **Q** and **q** form a right triangle of sides 1 m and $3+\sqrt{3}m.$ Charges **2Q** and **q** form a right triangle of sides 1 m and $\sqrt{3}m.$

$F_x = 0.049N,$

$F_y = 0.09N,$

$\vec{F}_{net} = 0.049N\hat{i} + 0.09N\hat{j}$

113. W=0.054J

115. a. $\vec{E} = \frac{1}{4\pi\varepsilon_0}(\frac{q}{(2a)^2} - \frac{q}{a^2})\hat{i}$;

b. $\vec{E} = \frac{\sqrt{3}}{4\pi\varepsilon_0}\frac{q}{a^2}(-\hat{j})$;

c. $\vec{E} = \frac{2}{\pi\varepsilon_0}\frac{q}{a^2}\frac{1}{\sqrt{2}}(-\hat{j})$

117. $\vec{E} = 6.4\times10^6(\hat{i}) + 1.5\times10^7(\hat{j})N/C$

119. $F = qE_0(1 + x/a)\ W = \frac{1}{2}m(v^2 - v_0^2)$,

$\frac{1}{2}mv^2 = qE_0(\frac{15a}{2})J$

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).

CHAPTER OVERVIEW

3: The Electric Potential

3.1: Introduction

In The Electric Field, we just scratched the surface (or at least rubbed it) of electrical phenomena by examining some of the basic properties of electric charge, fields, and forces. This chapter offers a complementary view of electrical phenomena focused on an energy-based approach.

We know that it takes energy to illuminate light bulbs (Fig. $3.1.1$). If those light bulbs are connected to a battery and energy is conserved, potential energy must be stored in the battery. We would like to be able to explain this process more clearly and quantitatively.

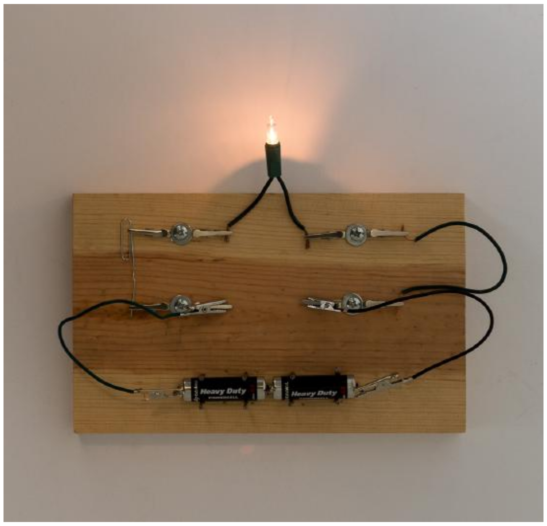


Figure $3.1.1$: Batteries in a circuit used stored electric potential energy to illuminate a light bulb. [1]

We also know that electrical energy is transmitted cross-country through power lines and may even spontaneously jump from clouds to trees in the form of lightning (Fig. $3.1.2$).

Figure $3.1.2$: The energy released in a lightning strike is an excellent illustration of the vast quantities of energy that may be stored and released by an electric potential difference. In this chapter, we calculate just how much energy can be released in a lightning strike and how this varies with the height of the clouds from the ground. (credit: Anthony Quintano)

Two terms commonly used to describe electricity are its energy and voltage, which we show in this chapter are directly related to the potential energy in a system. Batteries are typically a few volts, the outlets in your home frequently produce 120 volts, and power lines can be as high as hundreds of thousands of volts. However, these terms are not interchangeable. A motorcycle battery, for example, is small and would not be very successful in replacing a much larger car battery, yet each has the same voltage. In this chapter, we examine the relationship between electrical energy and voltage (or, more formally, electric potential) and begin to explore some of the many applications of electricity.

References

1. Exploratorium. [Internet] CircuitWorkbench_DSC_4543_H_0.jpg. Exploratorium. Available from: Exploratorium Circuit Workbench. (CC-BY-NC-SA 4.0)

3.2: Work and Energy

Learning Objectives

- Calculate the work done by any force.
- Calculate the kinetic energy of a particle given its mass and velocity or momentum.
- Relate the difference of potential energy to work done on a particle for a system without friction or air drag.
- Explain the meaning of the zero of the potential energy function for a system.
- Define conservative force and non-conservative force.
- Formulate the principle of conservation of mechanical energy, with or without the presence of non-conservative forces

Work

In physics, **work** represents a type of energy. Work is done when a force acts on something that undergoes a displacement from one position to another. Forces can vary as a function of position, and displacements can be along various paths between two points. We first define the increment of work dW done by a force $\vec{F}$ acting through an infinitesimal displacement d$\vec{r}$ as the dot product of these two vectors:

$$dW = \vec{F} \cdot d\vec{r} = |\vec{F}||d\vec{r}| \cos \theta. \tag{3.2.1}$$

Then, we can add the contributions for infinitesimal displacements along a path between two positions to get the total work.

Definition: Work Done by a Force

The work done by a force is the integral of the force with respect to displacement along the path of the displacement:

$$W_{AB} = \int_{path\ AB} \vec{F} \cdot d\vec{r}. \tag{3.2.2}$$

The vectors involved in the definition of the work done by a force acting on a particle are illustrated in Figure $3.2.1$.

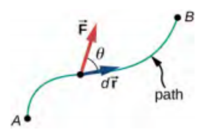


Figure $3.2.1$: Vectors used to define work. The force acting on a particle and its infinitesimal displacement are shown at one point along the path between A and B. The infinitesimal work is the dot product of these two vectors; the total work is the integral of the dot product along the path.

We choose to express the dot product in terms of the magnitudes of the vectors and the cosine of the angle between them because the meaning of the dot product for work can be put into words more directly in terms of magnitudes and angles. We could equally well have expressed the dot product in terms of the various components introduced in Vectors. In two dimensions, these were the x- and y-components in Cartesian coordinates, or the r- and φ-components in polar coordinates; in three dimensions, it was just x-, y-, and z-components. Which choice is more convenient depends on the situation. In words, you can express Equation 3.2.1 for the work done by a force acting over a displacement as a product of one component acting parallel to the other component. From the properties of vectors, it doesn't matter if you take the component of the force parallel to the displacement or the component of the displacement parallel to the force—you get the same result either way.

Recall that the magnitude of a force times the cosine of the angle the force makes with a given direction is the component of the force in the given direction. The components of a vector can be positive, negative, or zero, depending on whether the angle between the vector and the component-direction is between 0° and 90° or 90° and 180°, or is equal to 90°. As a result, the work done by a force can be positive, negative, or zero, depending on whether the force is generally in the direction of the displacement, generally opposite to the displacement, or perpendicular to the displacement. The maximum work is done by a given force when it is along the direction of the displacement (cos θ = ± 1), and zero work is done when the force is perpendicular to the displacement (cos θ = 0).

The unit of work is units of force multiplied by units of length, which in the SI system is newtons times meters, $\mathrm{N \cdot m}$. This combination is called a joule and is abbreviated as J.

Kinetic Energy

Kinetic energy is the energy of motion. For particles with speeds slow compared to the speed of light, we can define the kinetic energy of a particle with mass m and speed v or momentum magnitude $p = mv$.

Definition: Kinetic Energy

The kinetic energy of a particle is:

$$K = \frac{1}{2}mv^2 = \frac{p^2}{2m}. \tag{3.2.3}$$

We then extend this definition to any system of particles by adding up the kinetic energies of all N of its constituent particles:

$$K = \sum_{i=1}^{N} \frac{1}{2} m_i v_i^2. \tag{3.2.4}$$

The units of kinetic energy are joules (J).

If force is applied to a particle, it will accelerate, thereby changing its speed and kinetic energy. However, from Equation 3.2.2, the applied force must also do work on the particle. As a result, work and kinetic energy must be related, and this relationship is defined in the **work-kinetic energy theorem**.

Definition: Work-Kinetic Energy Theorem

The net work done on a particle equals the change in the particle's kinetic energy:

$$W_{net} = K_B - K_A. \tag{3.2.5}$$

Potential Energy

Potential energy is "stored" energy. As an example, consider kicking a football as illustrated in Figure 3.2.1. For simplicity, let's ignore friction and air resistance and define our coordinate system with the vertical axis in the positive direction. As the football rises, the work done by the gravitational force on the football is negative, because the ball’s displacement is positive vertically and the force due to gravity is negative vertically. The ball slows down until it reaches its highest point in the motion, implying decreasing kinetic energy. Under the principle of conservation of energy, this loss in kinetic energy must translate into a gain in gravitational potential energy of the football-Earth system so that the total energy of the system is conserved (constant). In some sense, energy is being temporarily "stored" by gravity in the increased height of the football.

As the football falls toward Earth, the work done on the football is now positive, because the displacement and the gravitational force both point vertically downward. The ball also speeds up, which indicates an increase in kinetic energy. Therefore, energy is converted from gravitational potential energy back into kinetic energy.

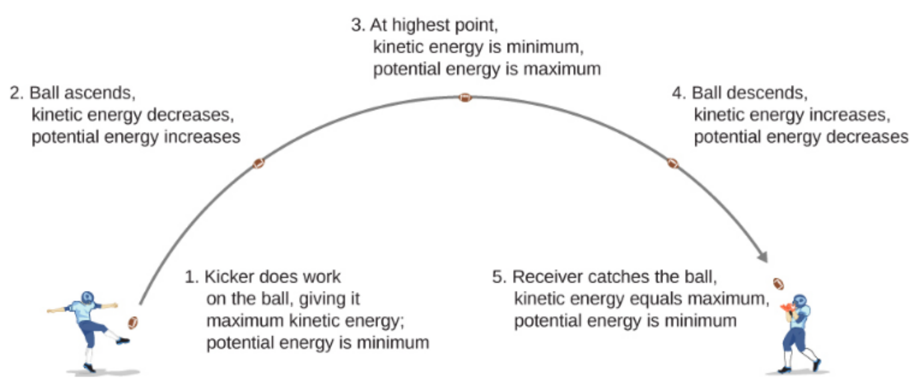


Figure 3.2.1: As a football starts its descent toward the wide receiver, gravitational potential energy is converted back into kinetic energy.

Based on this scenario, we can define the difference of potential energy from point A to point B as the negative of the work done:

$$\Delta U_{AB} = U_B - U_A = -W_{AB} \tag{3.2.6}$$

This formula explicitly states a **potential energy difference**, not just an absolute potential energy. Therefore, we need to define potential energy at a given position in such a way as to state standard values of potential energy on their own, rather than potential energy differences. We do this by rewriting the potential energy function in terms of an arbitrary constant,

$$\Delta U = U(\overrightarrow{\mathbf{r}}) - U\left(\overrightarrow{\mathbf{r}}_0\right) \tag{3.2.7}$$

The choice of the potential energy at a starting location of $\vec{r}_0$ is made out of convenience in the given problem. Most importantly, the choice should be stated and kept consistent throughout the given problem. There are some well-accepted choices of initial potential energy. For example, the lowest height in a problem is usually defined as zero potential energy, or if an object is in space, the farthest point away from the system is often defined as zero potential energy. Then, the potential energy, with respect to zero at $\vec{r}_0$, is just $U(\vec{r})$.

Common types of potential energy include gravitational potential energy and elastic potential energy, i.e., energy stored in compressed or extended springs. (We will see that electric charges can also have electric potential energy.)

Conservation of Energy

We can now mathematically define the concept of **energy conservation**.

Definition: Conservation of Energy

The mechanical energy E of a particle stays constant unless forces outside the system or non-conservative forces do work on it, in which case, the change in the mechanical energy is equal to the work done by the non-conservative forces:

$$W_{nc,\ AB} = \Delta(K+U)_{AB} = \Delta E_{AB}. \tag{3.2.8}$$

Non-conservative forces are dissipative forces such as friction or air resistance. These forces take energy away from the system as the system progresses, energy that you can't get back. These forces are path-dependent; it matters where the object starts and stops. **Conservative forces** are forces for which the change in energy is path-independent. Gravity is an example of a conservative force. Conservative forces can have a potential energy associated with them. (We will see that the electric force is a conservative force, and therefore we will be able to define an electric potential energy.) Mathematically, these concepts can be defined in terms of integrals of force over a path.

Definition: Conservative Force

The work done by a conservative force is independent of the path; in other words, the work done by a conservative force is the same for any path connecting two points:

$$W_{AB,\ path-1} = \int_{AB,\ path-1} \vec{F}_{cons} \cdot d\vec{r} = W_{AB,\ path-2} = \int_{AB,\ path-2} \vec{F}_{cons} \cdot d\vec{r}. \tag{3.2.9}$$

The work done by a non-conservative force depends on the path taken. Equivalently, a force is conservative if the work it does around any closed path is zero:

$$W_{closed\ path} = \oint \vec{E}_{cons} \cdot d\vec{r} = 0. \tag{3.2.10}$$

In Equation 3.2.10, we use the notation of a circle in the middle of the integral sign for a line integral over a closed path, a notation found in most physics and engineering texts.] Equations 3.2.9 and 3.2.10 are equivalent because any closed path is the sum of two paths: the first going from A to B, and the second going from B to A. The work done going along a path from B to A is the negative of the work done going along the same path from A to B, where A and B are any two points on the closed path:

$$\begin{aligned} 0 = \int \vec{F}_{cons} \cdot d\vec{r} &= \int_{AB,\ path-1} \vec{F}_{cons} \cdot d\vec{r} + \int_{BA,\ path-2} \vec{F}_{cons} \cdot d\vec{r} \\ &= \int_{AB,\ path-1} \vec{F}_{cons} \cdot d\vec{r} - \int_{AB,\ path-2} \vec{F}_{cons} \cdot d\vec{r} = 0. \end{aligned}$$

It is sometimes convenient to separate the case where the work done by non-conservative forces is zero, either because no such forces are assumed present or, like the normal force, they do zero work when the motion is parallel to the surface. Then

Conservation of Energy without Non-Conservative Forces

The mechanical energy of a particle does not change if all the non-conservative forces that may act on it do no work.

$$0 = W_{nc,\ AB} = \Delta(K+U)_{AB} = \Delta E_{AB}. \tag{3.2.11}$$

In many cases, there exist systems for which Equation 3.2.11 will hold exactly or approximately. In such situations, conservation of energy can be a powerful tool for understanding the behavior of the system.

3.3: Electric Potential Energy

Learning Objectives

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

- Define the work done by an electric force.
- Define electric potential energy.
- Apply work and potential energy in a system of a point charge in uniform electric field.

In this section, we begin to explore the use of energy to describe physical systems with electric charges through the definition of electric potential energy. Suppose a positively-charged particle is placed in between a positive plate and another equally-charged, parallel negative plate (Figure $3.3.1$). From Common Electric Field Models, we know that a nearly uniform electric field will exist between the plates directed from the positive plate to the negative plate. From Electric Fields and Forces, we also know that the electric field will exert a force on the positive charge in the same direction as the field. When the free positive charge q is accelerated by an electric field, it gains kinetic energy. The process is analogous to an object being accelerated by a gravitational field, as if the charge were going down an electrical hill where its electric potential energy is converted into kinetic energy, although, of course, the sources of the forces are very different. Let us explore the work done on a charge q by the electric field in this process, so that we may develop a definition of **electric potential energy**.

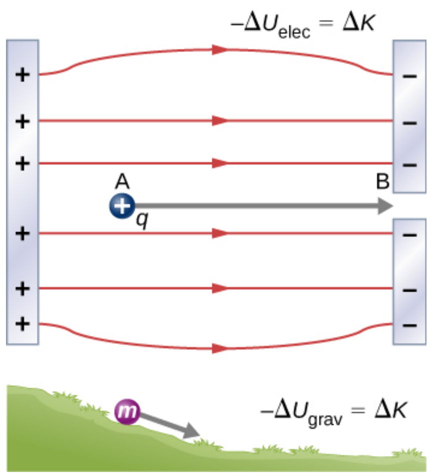


Figure $3.3.1$: A charge accelerated by an electric field is analogous to a mass going down a hill. In both cases, potential energy decreases as kinetic energy increases, $-\Delta U = \Delta K$. Work is done by a force, but since this force is conservative, we can write $W = -\Delta U$.

As we will demonstrate later, the electric force is a conservative force, which means that the work done on q is independent of the path taken. This is exactly analogous to the gravitational force. When a force is conservative, it is possible to define a potential energy associated with the force. It is usually easier to work with the potential energy (because it depends only on position) than to calculate the work directly.

Example $3.3.1$ shows how the gravitational potential energy can be calculated for the analogous system of a mass in a gravitational field.

Example $3.3.1$

What is the gravitational potential energy of a mass m at a height y above the ground near the surface of the Earth?

Solution

From the example in Electric Fields and Forces, we know how to calculate the gravitational force on a mass from the nearly-constant gravitational field near the surface of the Earth. From the expression for work in Work and Energy, we can then calculate the work done on the mass by the gravitational force as the mass falls from an initial height to its final height. Because we know the gravitational force is a conservative force, we can then relate the work done on the mass to the change in the gravitational potential energy in the system.

Figure $3.3.2$ shows a diagram of the system, including the gravitational field and gravitational force, in its initial state and then, after it has fallen, in its final state. Observe that the displacement is parallel to the gravitational force.

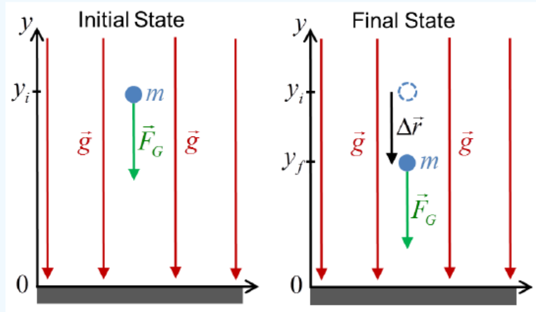


Figure $3.3.2$: Initial and final states of a mass falling in a gravitational field. [1]

CALCULATE

The gravitational force on the mass is $\vec{F} = m\vec{g} = mg(-\hat{j})$. The work done by the gravitational force on the mass during its displacement $\Delta\vec{r} = (y_i - y_f)(-\hat{j})$ is then

$$W = \vec{F} \cdot \Delta\vec{r} = |\vec{F}||\Delta\vec{r}|\cos 0^\circ = mg(y_i - y_f).$$

For a conservative force, the change in the gravitational potential energy is

$$\Delta U_G = -W = -mg(y_i - y_f) = mgy_f - mgy_i = U_f - U_i.$$

Hence it follows that the gravitational potential energy

$$U_G = mgy + U_0, \tag{3.3.1}$$

where U_0 is the value of the potential energy at the ground ($y = 0$). This value is arbitrary, and we choose to set it to zero for convenience. Only differences in energy are physically significant so adding a constant value will not affect the physical response of the system.

CHECK

We know from experience that the gravitational potential energy is greater at larger heights, which is consistent with the given result.

Observe that the parallel-charged plates create a uniform electric field much like the uniform gravitational field near the surface of the Earth. As a result, Example $3.3.2$ shows that we can follow a similar analysis to find the electric potential energy of the system in Fig. $3.3.1$.

✓ Example $3.3.2$

What is the electrical potential energy of a charge q at a distance y away from the negative plate of the oppositely-charged parallel plates in Fig. $3.3.1$?

Solution

From Electric Fields and Forces, we know how to calculate the electric force on a charge. From the expression for work in Work and Energy, we then can calculate the work done on the charge by the electric force as the charge moves from its initial position to its final position. If we assume the electric force is a conservative force, we can then relate the work done on the charge to the change in the electrical potential energy in the system.

Figure $3.3.3$ shows a diagram of the system, including the electric field and electric force, in its initial state and then, after the particle has moved, in its final state. Observe that the displacement is parallel to the electric force.

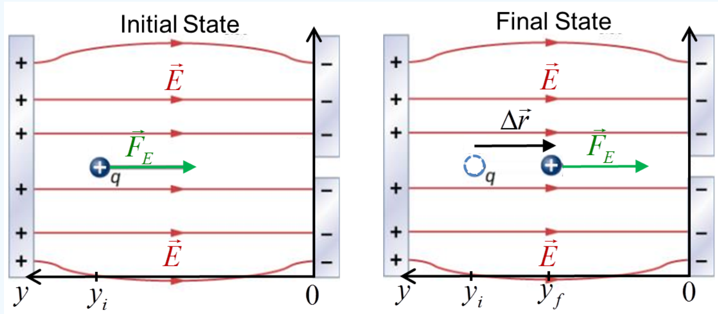


Figure 3.3.3: Initial and final states of a positive charge attracted to the negative plate in a set of oppositely-charged parallel plates. Each plate has an equal amount of charge. [2]

CALCULATE

The electric force on the charge is $\vec{F} = q\vec{E} = qE(-\hat{j})$. The work done by the electric force on the charge during its displacement
$\Delta\vec{r} = (y_i - y_f)(-\hat{j})$ is then

$$W = \vec{F} \cdot \Delta\vec{r} = |\vec{F}||\Delta\vec{r}|\cos 0^\circ = qE(y_i - y_f).$$

For a conservative force, the change in the gravitational potential energy is

$$\Delta U_E = -W = -qE(y_i - y_f) = qEy_f - qEy_i = U_f - U_i.$$

Hence it follows that the gravitational potential energy

$$U_E = qEy + U_0, \quad \text{(electric potential energy of point charge in a uniform field)} \tag{3.3.2}$$

where U_0 is the value of the potential energy at the ground ($y = 0$), which you could set to be zero, if you so choose. Note that while the figure shows a positive charge, all the steps in the calculation are still correct if the charge is negative.

CHECK

Equation 3.3.2 shows that the electric potential energy is greater when the charge is further away from the negative plate. This result makes sense because the positive charge is attracted to the negative plate, and, thus, it will take work to push it away from the negative plate. Moreover, we know that the positive charge will accelerate toward the negative plate, thereby gaining kinetic energy. By conservation of energy, the increase in kinetic energy of the charge must come from a decrease in the electric potential energy of the charge as it moves toward the plate.

For the system in Fig. $3.3.1$, we can construct an **energy diagram** (Fig. $3.3.1$). This energy diagram plots energy on the vertical axis and the distance from the negative plate on the horizontal axis. Given Equation 3.3.2 with $U_0 = 0$, the electric potential should be zero at the negative plate and then increase linearly away from the plate (recall that the positive y-axis is to the left). If a positively-charged particle is close to the negative plate and then is given a "kick" toward the positively charged plate, it will have a certain total amount of mechanical energy E_{mech}. The difference between the total mechanical energy and electric potential energy U_E is then kinetic energy K. However, the positive charge is repelled from the positive plate, and therefore should slow down. At some distance y_{max}, the particle comes to a stop as the electric potential energy $U_E = E_{mech}$ and therefore $K = 0$. This position is called the **turning point**, as the particle will stop there, and then turn around and accelerate back toward the negative plate.

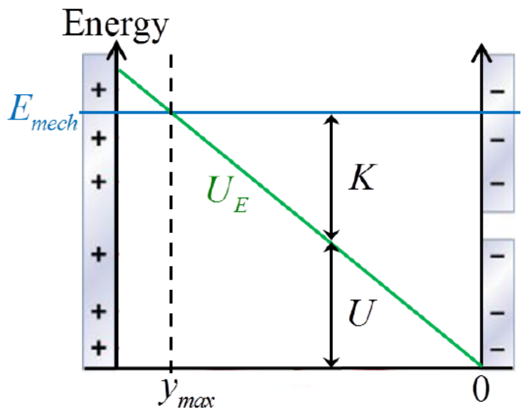


Figure 3.3.4: Energy diagram for a positively charged particle in between parallel plates with equal but opposite charge. [3]

Exercise 3.3.1

How will the energy diagram change if the particle in between the charged plates of Fig. 3.3.1 is negative?

Answer

According to Equation 3.3.2, the potential energy will be still linear, but negatively-valued with its maximum at $y = 0$. As particles tend toward a lower energy state, this means that the particle will move to the left losing potential energy (becoming more negative) and gaining kinetic energy. This outcome is consistent with the force perspective because we expect that the negatively-charged particle will be attracted to the positive plate and will accelerate as it moves in that direction.

The electric potential energy of a charged particle in a uniform field is a good starting point for our discussion, but it is a special case. In the next section, we will consider the more general case of the electric potential energy of a point charge in the electric field created by other point charges.

References

1. Ronald E. Kumon. [Internet] GravitationalPotential.png. (CC-BY-SA 4.0)
2. Background Image: University Physics, Volume 2. OpenStax; 2021. Modifications by Ronald E. Kumon. [Internet] ElectricPotential_UniformField.png. (CC-BY-SA 4.0)
3. Background Image: University Physics, Volume 2. OpenStax; 2021. Modifications by Ronald E. Kumon. [Internet] ElectricPotential_UniformField.png. (CC-BY-SA 4.0)

3.4: Electric Potential Energy of Point Charges

Learning Objectives

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

- Calculate the electric potential energy due to multiple point charges
- Prove that the electric force is a conservative force.
- Apply work and potential energy in systems with electric charges

In the previous section, we showed that we could calculate the electric potential energy from the work done by the electric force on a charge in a uniform field. We will now generalize that result to a system of point charges.

To begin, consider an electric charge $+q$ fixed at the origin and move another charge $+Q$ toward q in such a manner that, at each instant, the applied force $\vec{F}$ exactly balances the electric force $\vec{F}_e$ on Q (Figure $3.4.1$). The work done by the applied force $\vec{F}$ on the charge Q changes the potential energy of Q. We call this potential energy the **electrical potential energy** of Q.

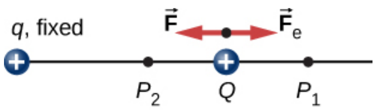


Figure $3.4.1$: Displacement of "test" charge Q in the presence of fixed "source" charge q.

The work W_{12} done by the applied force $\vec{F}$ when the particle moves from P_1 to P_2 may be calculated by

$$W_{12} = \int_{P_1}^{P_2} \vec{F} \cdot d\vec{l}. \tag{3.4.1}$$

Since the applied force $\vec{F}$ balances the electric force $\vec{F}_e$ on Q, the two forces have equal magnitude and opposite directions. Therefore, the applied force is

$$\vec{F} = -\vec{F}_e = -\frac{kqQ}{r^2}\hat{r}, \tag{3.4.2}$$

where we have defined positive to be pointing away from the origin and r is the distance from the origin. The directions of both the displacement and the applied force in the system in Figure $3.4.1$ are parallel, and thus the work done on the system is positive.

We use the letter U to denote electric potential energy, which has units of joules (J). When a conservative force does negative work, the system gains potential energy. When a conservative force does positive work, the system loses potential energy, $\Delta U = -W$. In the system in Figure $3.4.1$, the Coulomb force acts in the opposite direction to the displacement; therefore, the work is negative. However, we have increased the potential energy in the two-charge system.

Example $3.4.1$: Kinetic Energy of a Charged Particle

A $+3.0$ nC charge Q is initially at rest a distance of 10 cm (r_1) from a $+5.0$ nC charge q fixed at the origin (Figure $3.4.2$). Naturally, the Coulomb force accelerates Q away from q, eventually reaching 15 cm (r_2).

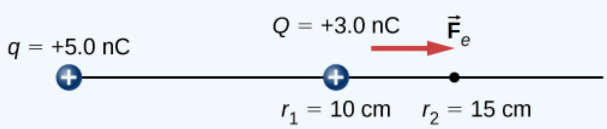


Figure $3.4.2$: The charge Q is repelled by q, thus having work done on it and gaining kinetic energy.

a. What is the work done by the electric field between r_1 and r_2?
b. How much kinetic energy does Q have at r_2?

Strategy

Calculate the work with the usual definition. Since Q started from rest, this is the same as the kinetic energy.

Solution

Integrating force over distance, we obtain

$$\begin{aligned} W_{12} &= \int_{r_1}^{r_2} \vec{F} \cdot d\vec{r} \\ &= \int_{r_1}^{r_2} \frac{kqQ}{r^2} dr \\ &= -\frac{kqQ}{r}\bigg|_{r_1}^{r_2} \\ &= kqQ\left[\frac{-1}{r_2} + \frac{1}{r_1}\right] \\ &= (8.99 \times 10^9 \,\mathrm{N\,m^2/C^2})(5.0 \times 10^{-9}\,\mathrm{C})(3.0 \times 10^{-9}\,\mathrm{C})\left[\frac{-1}{0.15\,\mathrm{m}} + \frac{1}{0.10\,\mathrm{m}}\right] \\ &= 4.5 \times 10^{-7}\,\mathrm{J}. \end{aligned}$$

This is also the value of the kinetic energy at r_2.

Significance

Charge Q was initially at rest; the electric field of q did work on Q, so now Q has kinetic energy equal to the work done by the electric field.

Exercise 3.4.1

If Q has a mass of $4.00\,\mu\mathrm{g}$, what is the speed of Q at r_2?

Answer

$K = \frac{1}{2}mv^2$, $v = \sqrt{2\frac{\mathrm{K}}{\mathrm{m}}} = \sqrt{2\frac{4.5\times10^{-7}\,\mathrm{J}}{4.00\times10^{-9}\,\mathrm{kg}}} = 15\,\mathrm{m/s}.$

In this example, the work W done to accelerate a positive charge from rest is positive and results from a loss in U, or a negative ΔU. A value for U can be found at any point by taking one point as a reference and calculating the work needed to move a charge to the other point.

Electric Potential Energy

Work W done to accelerate a positive charge from rest is positive and results from a loss in U, or a negative ΔU. Mathematically,

$$W = -\Delta U. \tag{3.4.3}$$

Potential energy accounts for work done by a conservative force and gives added insight regarding energy and energy transformation without the necessity of dealing with the force directly. It is much more common, for example, to use the concept of electric potential energy than to deal with the Coulomb force directly in real-world applications.

The Conservative Electric Force

We have previously asserted that the electric force is a conservative force. In this section, we will provide justification for this claim. First let's return to the scenario in in Fig. $3.4.2$. In polar coordinates with q at the origin and Q located at r, the displacement element vector is $d\vec{l} = \hat{r}\,dr$ and thus the work becomes

$$\begin{aligned} W_{12} &= kqQ\int_{r_1}^{r_2} \frac{1}{r^2}\hat{r}\cdot\hat{r}\,dr \\ &= \underbrace{kqQ\frac{1}{r_2}}_{\text{final point}} - \underbrace{kqQ\frac{1}{r_1}}_{\text{initial point}}. \end{aligned} \tag{3.4.4}$$

Notice that this result only depends on the endpoints and is otherwise independent of the path taken. To explore this further, compare path P_1 to P_2 with path $P_1P_3P_4P_2$ in Figure 3.4.3.

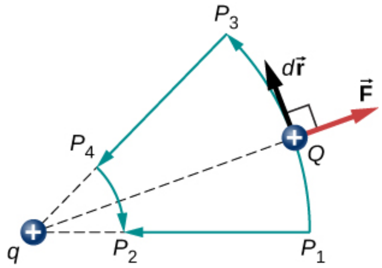


Figure 3.4.3: Two paths for displacement P_1 to P_2. The work on segments P_1P_3 and P_4P_2 are zero due to the electrical force being perpendicular to the displacement along these paths. Therefore, work on paths P_1P_2 and $P_1P_3P_4P_2$ are equal.

The segments P_1P_3 and P_4P_2 are arcs of circles centered at q. Since the force on Q points either toward or away from q, no work is done by a force balancing the electric force, because it is perpendicular to the displacement along these arcs. Therefore, the only work done is along segment P_3P_4 which is identical to P_1P_2.

One implication of this work calculation is that if we were to go around the path $P_1P_3P_4P_2P_1$, the net work would be zero (Figure 3.4.4). Recall that this is how we determine whether a force is conservative or not. Hence, because the electric force is related to the electric field by $\vec{F} = q\vec{E}$, the electric field is itself conservative. That is,

$$\oint \vec{E} \cdot d\vec{l} = 0. \tag{3.4.5}$$

Note that Q is a constant.

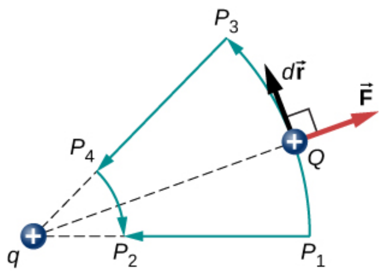


Figure 3.4.4: A closed path in an electric field. The net work around this path is zero.

Electric Potential Energy of Two Point Charges

Another implication is that we may define an electric potential energy for point charges. Recall that the work done by a conservative force is also expressed as the difference in the potential energy corresponding to that force. Therefore, the work W_{ref} to bring a charge from a reference point to a point of interest may be written as

$$W_{ref} = \int_{r_{\text{ref}}}^{r} \vec{F} \cdot d\vec{l} \tag{3.4.6}$$

and, by Equation 3.4.3, the difference in potential energy ($U_2 - U_1$) of the test charge Q between the two points is

$$\Delta U = -\int_{r_{\text{ref}}}^{r} \vec{F} \cdot d\vec{l}. \tag{3.4.7}$$

Therefore, we can write a general expression for the potential energy of two point charges (in spherical coordinates):

$$\Delta U = -\int_{r_{\text{ref}}}^{r} \frac{kqQ}{r^2} dr = -\left[-\frac{kqQ}{r}\right]_{r_{\text{ref}}}^{r} = kqQ\left[\frac{1}{r} - \frac{1}{r_{\text{ref}}}\right]. \tag{3.4.8}$$

We may take the second term to be an arbitrary constant reference level, which serves as the zero reference:

$$U(r) = k\frac{qQ}{r} - U_{\text{ref}}. \tag{3.4.9}$$

A convenient choice of reference that relies on our common sense is that when the two charges are infinitely far apart, there is no interaction between them. (Recall the discussion of reference potential energy in Potential Energy and Conservation of Energy.) Taking the potential energy of this state to be zero removes the term U_{ref} from the equation (just like when we say the ground is zero potential energy in a gravitational potential energy problem), and the potential energy of Q when it is separated from q by a distance r assumes the form

$$\underbrace{U(r) = k\frac{qQ}{r}}_{\text{zero reference at } r=\infty} . \tag{3.4.10}$$

This formula is symmetrical with respect to q and Q, so it is best described as the potential energy of the two-charge system.

Example 3.4.2: Potential Energy of a Charged Particle

A $+3.0$ nC charge Q is initially at rest a distance of 10 cm (r_1) from a $+5.0$ nC charge q fixed at the origin (Figure 3.4.5). Naturally, the Coulomb force accelerates Q away from q, eventually reaching 15 cm (r_2).

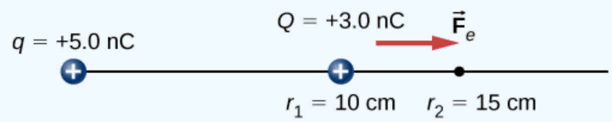


Figure 3.4.5: The charge Q is repelled by q, thus having work done on it and losing potential energy.

What is the change in the potential energy of the two-charge system from r_1 to r_2?

Strategy

Calculate the potential energy with the definition given above:

$\Delta U_{12} = -\int_{r_1}^{r_2} \vec{F} \cdot d\vec{r}$. Since Q started from rest, this is the same as the kinetic energy.

Solution

We have

$$\begin{aligned}
\Delta U_{12} &= -\int_{r_1}^{r_2} \vec{F} \cdot d\vec{r} \\
&= -\int_{r_1}^{r_2} \frac{kqQ}{r^2} dr \\
&= -\left[-\frac{kqQ}{r}\right]_{r_1}^{r_2} \\
&= kqQ\left[\frac{1}{r_2} - \frac{1}{r_1}\right] \\
&= (8.99 \times 10^9 \,\mathrm{N\,m^2/C^2})(5.0 \times 10^{-9}\,\mathrm{C})(3.0 \times 10^{-9}\,\mathrm{C})\left[\frac{1}{0.15\,\mathrm{m}} - \frac{1}{0.10\,\mathrm{m}}\right] \\
&= -4.5 \times 10^{-7}\,\mathrm{J}.
\end{aligned} \tag{3.4.11}$$

Significance

The change in the potential energy is negative, as expected, and equal in magnitude to the change in kinetic energy in this system. Recall from Example 3.4.1 that the change in kinetic energy was positive.

Exercise 3.4.2

What is the potential energy of Q relative to the zero reference at infinity at r_2 in the above example?

Answer

It has kinetic energy of 4.5×10^{-7} J at point r_2 and potential energy of 9.0×10^{-7} J, which means that as Q approaches infinity, its kinetic energy totals three times the kinetic energy at r_2, since all of the potential energy gets converted to kinetic.

Due to Coulomb's law, the forces due to multiple charges on a test charge Q superimpose; they may be calculated individually and then added. This implies that the work integrals and hence the resulting potential energies exhibit the same behavior. To demonstrate this, we consider an example of assembling a system of four charges.

Example 3.4.3: Assembling Four Positive Charges

Find the amount of work an external agent must do in assembling four charges $+2.0\,\mu$C, $+3.0\,\mu$C, $+4.0\,\mu$C, and $+5.0\,\mu$C at the vertices of a square of side 1.0 cm, starting each charge from infinity (Figure 3.4.6).

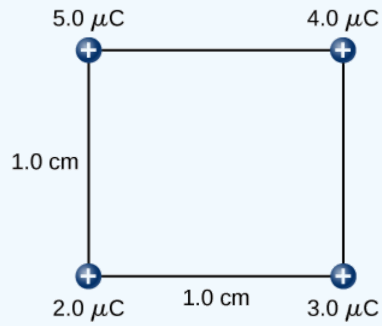


Figure 3.4.6: How much work is needed to assemble this charge configuration?

Strategy

We bring in the charges one at a time, giving them starting locations at infinity and calculating the work to bring them in from infinity to their final location. We do this in order of increasing charge.

Solution

Step 1. First bring the $+2.0\,\mu$C charge to the origin. Since there are no other charges at a finite distance from this charge yet, no work is done in bringing it from infinity,

$$W_1 = 0. \tag{3.4.12}$$

Step 2. While keeping the $+2.0\,\mu$C charge fixed at the origin, bring the $+3.0\,\mu$C charge to $(x, y, z) = (1.0\,\text{cm},\ 0,\ 0)$ (Figure 3.4.7). Now, the applied force must do work against the force exerted by the $+2.0\,\mu$C charge fixed at the origin. The work done equals the change in the potential energy of the $+3.0\,\mu$C charge:

$$\begin{aligned} W_2 &= k\frac{q_1 q_2}{r12} \\ &= \left(9.0 \times 10^9\,\frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\frac{(2.0 \times 10^{-6}\,\text{C})(3.0 \times 10^{-6}\,\text{C})}{1.0 \times 10^{-2}\,\text{m}} \\ &= 5.4\,\text{J}. \end{aligned}$$

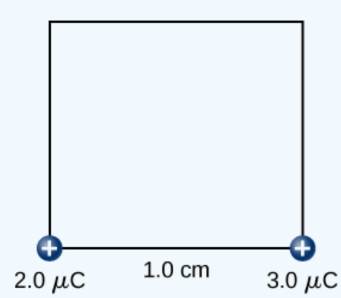


Figure 3.4.7: Step 2. Work W_2 to bring the $+3.0\,\mu$C charge from infinity.

Step 3. While keeping the charges of $+2.0\,\mu\text{C}$ and $+3.0\,\mu\text{C}$ fixed in their places, bring in the $+4.0\,\mu\text{C}$ charge to $(x, y, z) = (1.0\,\text{cm},\ 1.0\,\text{cm},\ 0)$(Figure $3.4.8$). The work done in this step is

$$\begin{aligned} W_3 &= k\frac{q_1 q_3}{r_{13}} + k\frac{q_2 q_3}{r_{23}} \\ &= \left(9.0 \times 10^9 \frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)\left[\frac{(2.0\times 10^{-6}\,\text{C})(4.0\times 10^{-6}\,\text{C})}{\sqrt{2}\times 10^{-2}\,\text{m}} + \frac{(3.0\times 10^{-6}\,\text{C})(4.0\times 10^{-6}\,\text{C})}{1.0\times 10^{-2}\,\text{m}}\right] \\ &= 15.9\,\text{J}. \end{aligned}$$

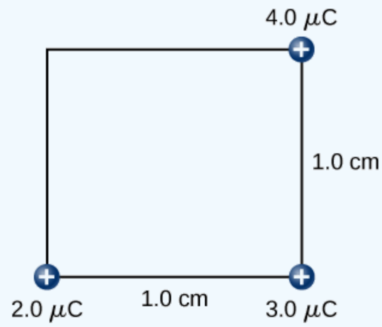


Figure $3.4.8$: Step 3. The work W_3 to bring the $+4.0\ \mu\text{C}$ charge from infinity.

Step 4. Finally, while keeping the first three charges in their places, bring the $+5.0\,\mu\text{C}$ charge to $(x, y, z) = (0,\ 1.0\,\text{cm},\ 0)$ (Figure $3.4.9$). The work done here is

$$\begin{aligned} W_4 &= kq_4\left[\frac{q_1}{r_{14}} + \frac{q_2}{r_{24}} + \frac{q_3}{r_{34}}\right], \\ &= \left(9.0 \times 10^9 \frac{\text{N}\cdot\text{m}^2}{\text{C}^2}\right)(5.0\times 10^{-6}\,\text{C})\left[\frac{(2.0\times 10^{-6}\,\text{C})}{1.0\times 10^{-2}\,\text{m}} + \frac{(3.0\times 10^{-6}\,\text{C})}{\sqrt{2}\times 10^{-2}\,\text{m}} + \frac{(4.0\times 10^{-6}\,\text{C})}{1.0\times 10^{-2}\,\text{m}}\right] \\ &= 36.5\,\text{J}. \end{aligned}$$

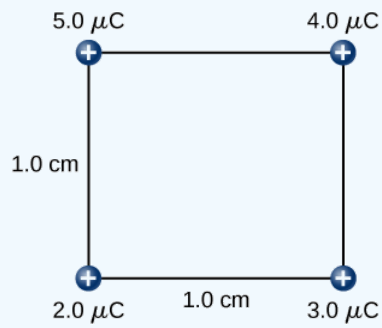


Figure $3.4.9$: Step 4. The work W_4 to bring the $+5.0\ \mu\text{C}$ charge from infinity.

Hence, the total work done by the applied force in assembling the four charges is equal to the sum of the work in bringing each charge from infinity to its final position:

$$\begin{aligned} W_T &= W_1 + W_2 + W_3 + W_4 \\ &= 0 + 5.4\,\text{J} + 15.9\,\text{J} + 36.5\,\text{J} \\ &= 57.8\,J. \end{aligned}$$

Significance

The work on each charge depends only on its pairwise interactions with the other charges. No more complicated interactions need to be considered; the work on the third charge only depends on its interaction with the first and second charges, the interaction between the first and second charge does not affect the third.

? Exercise $3.4.3$

(1) Is the electrical potential energy of two point charges positive or negative if the charges are of the same sign? (2) Opposite signs? (3) How does this relate to the work necessary to bring the charges into proximity from infinity?

Answer

(1) positive, (2) negative, and (3) these quantities are the same as the work you would need to do to bring the charges in from infinity

Note that the electrical potential energy is positive if the two charges are of the same type, either positive or negative, and negative if the two charges are of opposite types. This makes sense if you think of the change in the potential energy ΔU as you bring the two charges closer or move them farther apart. Depending on the relative types of charges, you may have to work on the system or the system would do work on you, that is, your work is either positive or negative. If you have to do positive work on the system (actually push the charges closer), then the energy of the system should increase. If you bring two positive charges or two negative charges closer, you have to do positive work on the system, which raises their potential energy. Since potential energy is proportional to $1/r$, the potential energy goes up when r goes down between two positive or two negative charges.

On the other hand, if you bring a positive and a negative charge nearer, you have to do negative work on the system (the charges are pulling you), which means that you take energy away from the system. This reduces the potential energy. Since potential energy is negative in the case of a positive and a negative charge pair, the increase in $1/r$ makes the potential energy more negative, which is the same as a reduction in potential energy.

The result from Example $3.4.2$ may be extended to systems with any arbitrary number of charges. In this case, it is most convenient to write the formula as

$$W_{12 \ldots N} = \frac{k}{2} \sum_{i}^{N} \sum_{j}^{N} \frac{q_i q_j}{r_{ij}} \, for \, i \neq j. \tag{3.4.13}$$

The factor of 1/2 accounts for adding each pair of charges twice.

3.5: Electric Potential

Learning Objectives

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

- Define electric potential, voltage, and potential difference.
- Calculate electric potential and potential difference from potential energy.
- Define the electron-volt.
- Describe systems in which the electron-volt is a useful unit.
- Apply conservation of energy to systems with electric charges.

Recall that earlier we defined electric field to be a quantity independent of the test charge in a given system, which would nonetheless allow us to calculate the force that would result on an arbitrary test charge. (The default assumption in the absence of other information is that the test charge is positive.) We briefly defined a field for gravity, but gravity is always attractive, whereas the electric force can be either attractive or repulsive. Therefore, although potential energy is perfectly adequate in a gravitational system, it is convenient to define a quantity that allows us to calculate the work on a charge independent of the magnitude of the charge. Calculating the work directly may be difficult, since $W = \vec{F} \cdot \vec{d}$ and the direction and magnitude of $\vec{F}$ can be complex for multiple charges, for odd-shaped objects, and along arbitrary paths. But we do know that because $\vec{F}$, the work, and hence ΔU is proportional to the test charge q. To have a physical quantity that is independent of test charge, we define **electric potential** V (or simply potential, since electric is understood) to be the potential energy per unit charge:

Electric Potential

The electric potential energy per unit charge is

$$V = \frac{U}{q}. \tag{3.5.1}$$

Since the potential energy U is proportional to the test charge q, the dependence on q cancels. Thus, V does not depend on q. The change in potential energy ΔU is crucial, so we are concerned with the difference in potential or potential difference ΔV between two points, where

$$\Delta V = V_B - V_A = \frac{\Delta U}{q} \tag{3.5.2}$$

Electric Potential Difference

The **electric potential difference** between points A and B, $V_B - V_A$ is defined to be the change in potential energy of a charge q moved from A to B, divided by the charge. Units of potential difference are joules per coulomb, given the name volt (V) after Alessandro Volta.

$$1 \text{ V} = 1 \text{ J/C} \tag{3.5.3}$$

The familiar term **voltage** is the common name for electric potential difference. Keep in mind that whenever a voltage is quoted, it is understood to be the potential difference between two points. For example, every battery has two terminals, and its voltage is the potential difference between them. More fundamentally, the point you choose to be zero volts is arbitrary. This is analogous to the fact that gravitational potential energy has an arbitrary zero, such as sea level or perhaps a lecture hall floor. It is worthwhile to emphasize the distinction between potential difference and electrical potential energy.

Potential Difference and Electrical Potential Energy

The relationship between potential difference (or voltage) and electrical potential energy is given by

$$\Delta V = \frac{\Delta U}{q} \tag{3.5.4}$$

or

$$\Delta U = q\Delta V. \tag{3.5.5}$$

Voltage is not the same as energy. Voltage is the energy per unit charge. Thus, a motorcycle battery and a car battery can both have the same voltage (more precisely, the same potential difference between battery terminals), yet one stores much more energy than the other because $\Delta U = q\Delta V$. The car battery can move more charge than the motorcycle battery, although both are 12-V batteries.

Example 3.5.1: Calculating Energy

You have a 12.0-V motorcycle battery that can move 5000 C of charge, and a 12.0-V car battery that can move 60,000 C of charge. How much energy does each deliver? (Assume that the numerical value of each charge is accurate to three significant figures.)

Strategy

To say we have a 12.0-V battery means that its terminals have a 12.0-V potential difference. When such a battery moves charge, it puts the charge through a potential difference of 12.0 V, and the charge is given a change in potential energy equal to $\Delta U = q\Delta V$. To find the energy output, we multiply the charge moved by the potential difference.

Solution

For the motorcycle battery, $q = 5000\,\text{C}$ and $\Delta V = 12.0\,\text{V}$. The total energy delivered by the motorcycle battery is

$$\Delta U_{cycle} = (5000\,\text{C})(12.0\,\text{V}) = (5000\,\text{C})(12.0\,\text{J/C}) = 6.00 \times 10^4\,\text{J}.$$

Similarly, for the car battery, $q = 60,000\,\text{C}$ and

$$\Delta U_{car} = (60,000\,\text{C})(12.0\,\text{V}) = 7.20 \times 10^5\,\text{J}.$$

Significance

Voltage and energy are related, but they are not the same thing. The voltages of the batteries are identical, but the energy supplied by each is quite different. A car battery has a much larger engine to start than a motorcycle. Note also that as a battery is discharged, some of its energy is used internally and its terminal voltage drops, such as when headlights dim because of a depleted car battery. The energy supplied by the battery is still calculated as in this example, but not all of the energy is available for external use.

Exercise 3.5.1

How much energy does a 1.5-V AAA battery have that can move 100 C?

Answer

$$\Delta U = q\Delta V = (100\,\text{C})(1.5\,\text{V}) = 150\,\text{J}$$

Note that the energies calculated in the previous example are absolute values. The change in potential energy for the battery is negative, since it loses energy. These batteries, like many electrical systems, actually move negative charge—electrons in particular. The batteries repel electrons from their negative terminals (A) through whatever circuitry is involved and attract them to their positive terminals (B), as shown in Figure 3.5.1. The change in potential is $\Delta V = V_B - V_A = +12\,V$ and the charge q is negative, so that $\Delta U = q\Delta V$ is negative, meaning the potential energy of the battery has decreased when q has moved from A to B.

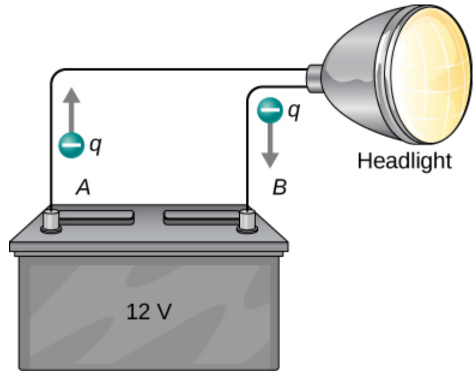


Figure 3.5.1: A battery moves negative charge from its negative terminal through a headlight to its positive terminal. Appropriate combinations of chemicals in the battery separate charges so that the negative terminal has an excess of negative charge, which is repelled by it and attracted to the excess positive charge on the other terminal. In terms of potential, the positive terminal is at a higher voltage than the negative terminal. Inside the battery, both positive and negative charges move.

Example 3.5.2: How Many Electrons Move through a Headlight Each Second?

When a 12.0-V car battery powers a single 30.0-W headlight, how many electrons pass through it each second?

Strategy

To find the number of electrons, we must first find the charge that moves in 1.00 s. The charge moved is related to voltage and energy through the equations $\Delta U = q\Delta V$. A 30.0-W lamp uses 30.0 joules per second. Since the battery loses energy, we have $\Delta U = -30\text{ J}$ and, since the electrons are going from the negative terminal to the positive, we see that $\Delta V = +12.0\,\text{V}$.

Solution

To find the charge q moved, we solve the equation $\Delta U = q\Delta V$:

$$q = \frac{\Delta U}{\Delta V}. \tag{3.5.6}$$

Entering the values for ΔU and ΔV, we get

$$q = \frac{-30.0\,\text{J}}{+12.0\,\text{V}} = \frac{-30.0\,\text{J}}{+12.0\,\text{J/C}} = -2.50\,\text{C}. \tag{3.5.7}$$

The number of electrons n_e is the total charge divided by the charge per electron. That is,

$$n_e = \frac{-2.50\,\text{C}}{-1.60\times10^{-19}\text{C}/e^-} = 1.56\times10^{19}\,\text{electrons}. \tag{3.5.8}$$

Significance

This is a very large number. It is no wonder that we do not ordinarily observe individual electrons with so many being present in ordinary systems. In fact, electricity had been in use for many decades before it was determined that the moving charges in many circumstances were negative. Positive charge moving in the opposite direction of negative charge often produces identical effects; this makes it difficult to determine which is moving or whether both are moving.

Exercise 3.5.2

How many electrons would go through a 24.0-W lamp?

Answer

$-2.00\,\text{C},\; n_e = 1.25\times10^{19}\,\text{electrons}$

The Electron-Volt

The energy per electron is very small in macroscopic situations like that in the previous example—a tiny fraction of a joule. But on a submicroscopic scale, such energy per particle (electron, proton, or ion) can be of great importance. For example, even a tiny fraction of a joule can be great enough for these particles to destroy organic molecules and harm living tissue. The particle may do its damage by direct collision, or it may create harmful X-rays, which can also inflict damage. It is useful to have an energy unit related to submicroscopic effects.

Figure $3.5.2$ shows a situation related to the definition of such an energy unit. An electron is accelerated between two charged metal plates, as it might be in an old-model television tube or oscilloscope. The electron gains kinetic energy that is later converted into another form—light in the television tube, for example. (Note that in terms of energy, "downhill" for the electron is "uphill" for a positive charge.) Since energy is related to voltage by $\Delta U = q\Delta V$, we can think of the joule as a coulomb-volt.

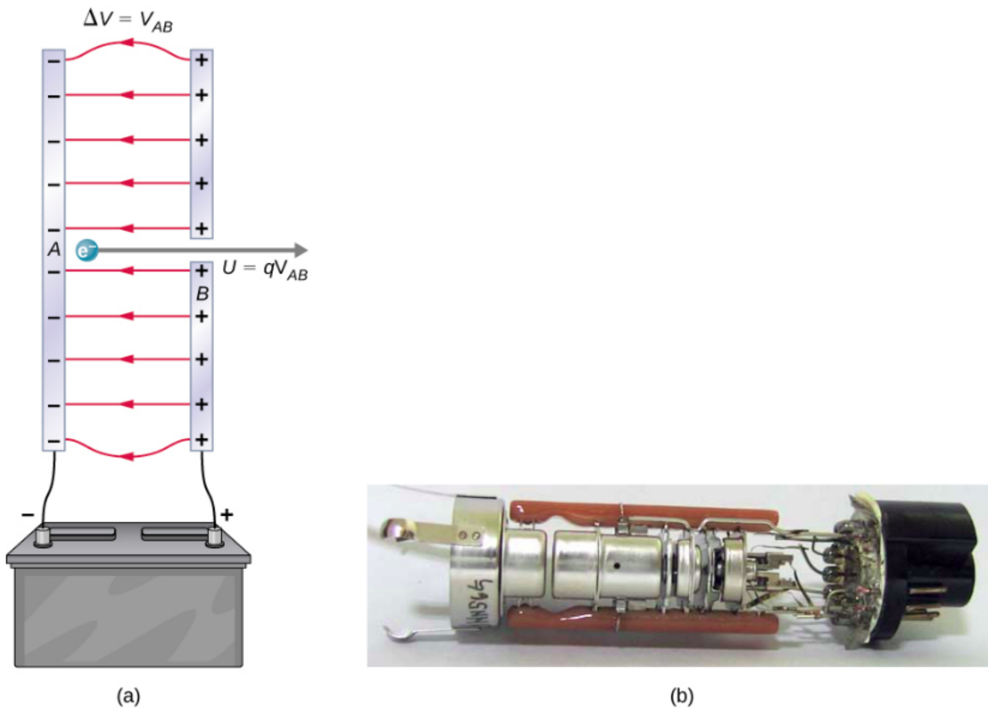


Figure $3.5.2$: A typical electron gun accelerates electrons using a potential difference between two separated metal plates. By conservation of energy, the kinetic energy has to equal the change in potential energy, so $KE = qV$. The energy of the electron in electron-volts is numerically the same as the voltage between the plates. For example, a 5000-V potential difference produces 5000-eV electrons. The conceptual construct, namely two parallel plates with a hole in one, is shown in (a), while a real electron gun is shown in (b).

The Electron-Volt Unit

On the submicroscopic scale, it is more convenient to define an energy unit called the **electron-volt** (eV), which is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$1\text{ eV} = (1.60 \times 10^{-19}\text{C})(1\text{ V}) = (1.60 \times 10^{-19}\text{C})(1\text{ J/C}) = 1.60 \times 10^{-19}\text{ J}. \tag{3.5.9}$$

An electron accelerated through a potential difference of 1 V is given an energy of 1 eV. It follows that an electron accelerated through 50 V gains 50 eV. A potential difference of 100,000 V (100 kV) gives an electron an energy of 100,000 eV (100 keV), and so on. Similarly, an ion with a double positive charge accelerated through 100 V gains 200 eV of energy. These simple relationships between accelerating voltage and particle charges make the electron-volt a simple and convenient energy unit in such circumstances.

The electron-volt is commonly employed in submicroscopic processes—chemical valence energies and molecular and nuclear binding energies are among the quantities often expressed in electron-volts. For example, about 5 eV of energy is required to break up certain organic molecules. If a proton is accelerated from rest through a potential difference of 30 kV, it acquires an energy of 30 keV (30,000 eV) and can break up as many as 6000 of these molecules $(30,000\text{ eV} \div 5\text{ eV per molecule} = 6000\text{ molecules})$. Nuclear decay energies are on the order of 1 MeV (1,000,000 eV) per event and can thus produce significant biological damage.

Conservation of Energy

The total energy of a system is conserved if there is no net addition (or subtraction) due to work or heat transfer. For conservative forces, such as the electrostatic force, conservation of energy states that mechanical energy is a constant.

Mechanical energy is the sum of the kinetic energy and potential energy of a system; that is, $K+U=\text{constant}$. A loss of U for a charged particle becomes an increase in its K. Conservation of energy is stated in equation form as

$$K+U=\text{constant} \tag{3.5.10}$$

or

$$K_i+U_i=K_f+U_f \tag{3.5.11}$$

where i and f stand for initial and final conditions. As we have found many times before, considering energy can give us insights and facilitate problem solving.

Example 3.5.3: Electrical Potential Energy Converted into Kinetic Energy

Calculate the final speed of a free electron accelerated from rest through a potential difference of 100 V. (Assume that this numerical value is accurate to three significant figures.)

Strategy

We have a system with only conservative forces. Assuming the electron is accelerated in a vacuum, and neglecting the gravitational force (we will check on this assumption later), all of the electrical potential energy is converted into kinetic energy. We can identify the initial and final forms of energy to be

$K_i=0, K_f=\frac{1}{2}mv^2, U_i=qV, U_f=0.$

Solution

Conservation of energy states that

$$K_i+U_i=K_f+U_f. \tag{3.5.12}$$

Entering the forms identified above, we obtain

$$qV=\frac{mv^2}{2}. \tag{3.5.13}$$

We solve this for v:

$$v=\sqrt{\frac{2qV}{m}}. \tag{3.5.14}$$

Entering values for q, V, and m gives

$$v=\sqrt{\frac{2(-1.60\times10^{-19}\,\text{C})(-100\,\text{J/C})}{9.11\times10^{-31}\text{kg}}}=5.93\times10^{6}\ \text{m/s}. \tag{3.5.15}$$

Significance

Note that both the charge and the initial voltage are negative, as in Figure $3.5.2$. From the discussion of electric charge and electric field, we know that electrostatic forces on small particles are generally very large compared with the gravitational force. The large final speed confirms that the gravitational force is indeed negligible here. The large speed also indicates how easy it is to accelerate electrons with small voltages because of their very small mass. Voltages much higher than the 100 V in this problem are typically used in electron guns. These higher voltages produce electron speeds so great that effects from special relativity must be taken into account and will be discussed elsewhere. That is why we consider a low voltage (accurately) in this example.

Exercise $3.5.3$

How would this example change with a positron? A positron is identical to an electron except the charge is positive.

Answer

It would be going in the opposite direction, with no effect on the calculations as presented.

Before presenting problems involving electrostatics, we suggest a problem-solving strategy to follow for this topic.

Problem-Solving Strategy: Electrostatics

1. Examine the situation to determine if static electricity is involved; this may concern separated stationary charges, the forces among them, and the electric fields they create.
2. Identify the system of interest. This includes noting the number, locations, and types of charges involved.
3. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful. Determine whether the electric force is to be considered directly—if so, it may be useful to draw a free-body diagram, using electric field lines.
4. Make a list of what is given or can be inferred from the problem as stated (identify the knowns). For example, it is important to distinguish the electric force $\vec{F}$ from the electric field $\vec{E}$ or the electric potential energy U from the electric potential V.
5. Solve the appropriate equation for the quantity to be determined (the unknown) or draw the field lines as requested.
6. Examine the answer to see if it is reasonable: Does it make sense? Are units correct and the numbers involved reasonable?

3.6: Electric Potential of a Point Charge

Learning Objectives

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

- Calculate the potential due to a point charge
- Calculate the potential of a system of multiple point charges
- Calculate the potential of a continuous charge distribution

Point charges, such as electrons, are among the fundamental building blocks of matter. Furthermore, spherical charge distributions (such as charge on a metal sphere) create external electric fields exactly like a point charge. The electric potential due to a point charge is, thus, a case we need to consider.

We can use calculus to find the work needed to move a test charge q from a large distance away to a distance of r from a point charge q. Noting the connection between work and potential $W = -q\Delta V$, as in the last section, we can obtain the following result.

Electric Potential V of a Point Charge

The electric potential V of a point charge is given by

$$\underbrace{V = \frac{kq}{r}}_{\text{point charge}} \tag{3.6.1}$$

where k is a constant equal to $8.99 \times 10^9\ \mathrm{N \cdot m^2/C^2}$.

The potential in Equation 3.6.1 at infinity is chosen to be zero. Thus, V for a point charge decreases with distance, whereas $\vec{E}$ for a point charge decreases with distance squared:

$$E = \frac{F}{q_t} = \frac{kq}{r^2} \tag{3.6.2}$$

Recall that the electric potential V is a scalar and has no direction, whereas the electric field $\vec{E}$ is a vector. To find the voltage due to a combination of point charges, you add the individual voltages as numbers. To find the total electric field, you must add the individual fields as vectors, taking magnitude and direction into account. This is consistent with the fact that V is closely associated with energy, a scalar, whereas $\vec{E}$ is closely associated with force, a vector.

Example 3.6.1: What Voltage Is Produced by a Small Charge on a Metal Sphere?

Charges in static electricity are typically in the nanocoulomb (nC) to microcoulomb (μC) range. What is the voltage 5.00 cm away from the center of a 1-cm-diameter solid metal sphere that has a –3.00-nC static charge?

Strategy

As we discussed in Electric Charges and Fields, charge on a metal sphere spreads out uniformly and produces a field like that of a point charge located at its center. Thus, we can find the voltage using the equation $V = \frac{kq}{r}$.

Solution

Entering known values into the expression for the potential of a point charge (Equation 3.6.1), we obtain

$$\begin{aligned} V &= k\frac{q}{r} \\ &= (9.00 \times 10^9\ \mathrm{N \cdot m^2/C^2})\left(\frac{-3.00 \times 10^{-9}\,\mathrm{C}}{5.00 \times 10^{-2}\,\mathrm{m}}\right) \\ &= -539\ \mathrm{V}. \end{aligned}$$

Significance

The negative value for voltage means a positive charge would be attracted from a larger distance, since the potential is lower (more negative) than at larger distances. Conversely, a negative charge would be repelled, as expected.

✔ Example $3.6.2$: What Is the Excess Charge on a Van de Graaff Generator?

A demonstration **Van de Graaff generator** has a 25.0-cm-diameter metal sphere that produces a voltage of 100 kV near its surface (Figure). What excess charge resides on the sphere? (Assume that each numerical value here is shown with three significant figures.)

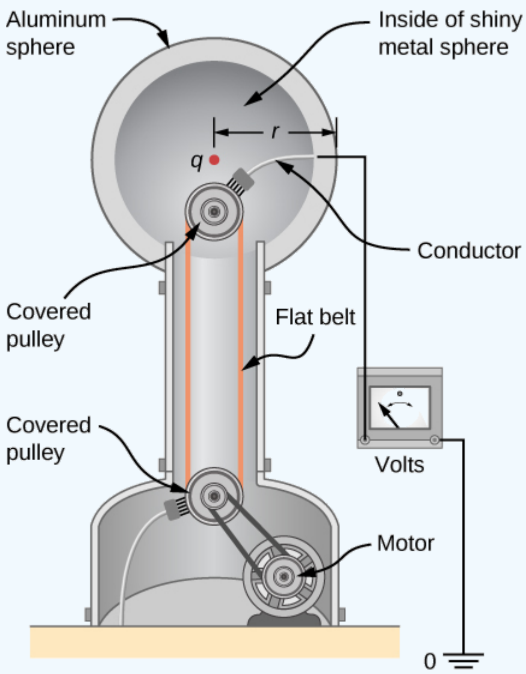


Figure $3.6.1$: The voltage of this demonstration Van de Graaff generator is measured between the charged sphere and ground. Earth's potential is taken to be zero as a reference. The potential of the charged conducting sphere is the same as that of an equal point charge at its center.

Strategy

The potential on the surface is the same as that of a point charge at the center of the sphere, 12.5 cm away. (The radius of the sphere is 12.5 cm.) We can thus determine the excess charge using Equation $3.6.1$

$$V=\frac{kq}{r}. \tag{3.6.3}$$

Solution

Solving for q and entering known values gives

\[

$$\begin{aligned} q &= \frac{rV}{k} \\ &= \frac{(0.125\,\mathrm{m})(100\times 10^3\,\mathrm{V})}{8.99\times 10^9\,\mathrm{N\cdot m^2/C^2}} \\ &= 1.39\times 10^{-6}\,\mathrm{C} \\ &= 1.39\,\mu\mathrm{C}. \end{aligned} \tag{3.6.4}$$

\nonumber

Significance

This is a relatively small charge, but it produces a rather large voltage. We have another indication here that it is difficult to store isolated charges.

Exercise 3.6.1

What is the potential inside the metal sphere in Example 3.6.1?

Answer

$$\begin{aligned} V &= k\frac{q}{r} \\ &= (8.99\times10^{9}\,\mathrm{N\cdot m^2/C^2})\left(\frac{-3.00\times10^{-9}\,\mathrm{C}}{5.00\times10^{-3}\,\mathrm{m}}\right) \\ &= -5390\,\mathrm{V} \end{aligned}$$

Recall that the electric field inside a conductor is zero. Hence, any path from a point on the surface to any point in the interior will have an integrand of zero when calculating the change in potential, and thus the potential in the interior of the sphere is identical to that on the surface.

The voltages in both of these examples could be measured with a meter that compares the measured potential with ground potential. Ground potential is often taken to be zero (instead of taking the potential at infinity to be zero). It is the potential difference between two points that is of importance, and very often there is a tacit assumption that some reference point, such as Earth or a very distant point, is at zero potential. As noted earlier, this is analogous to taking sea level as $h=0$ when considering gravitational potential energy $U_g=mgh$.

Systems of Multiple Point Charges

Just as the electric field obeys a superposition principle, so does the electric potential. Consider a system consisting of N charges $q_1, q_2, \ldots, q_N$. What is the net electric potential V at a space point P from these charges? Each of these charges is a source charge that produces its own electric potential at point P, independent of whatever other changes may be doing. Let $V_1, V_2, \ldots, V_N$ be the electric potentials at P produced by the charges $q_1, q_2, \ldots, q_N$, respectively. Then, the net electric potential V_p at that point is equal to the sum of these individual electric potentials. You can easily show this by calculating the potential energy of a test charge when you bring the test charge from the reference point at infinity to point P:

$$V_p = V_1 + V_2 + \ldots + V_N = \sum_1^N V_i. \tag{3.6.5}$$

Note that electric potential follows the same principle of superposition as electric field and electric potential energy. To show this more explicitly, note that a test charge q_i at the point P in space has distances of $r_1, r_2, \ldots, r_N$ from the N, charges fixed in space above, as shown in Figure 3.6.2. Using our formula for the potential of a point charge for each of these (assumed to be point) charges, we find that

$$V_p = \sum_1^N k\frac{q_i}{r_i} = k\sum_1^N \frac{q_i}{r_i}. \tag{3.6.6}$$

Therefore, the electric potential energy of the test charge is

$$U_p = q_t V_p = q_t k \sum_1^N \frac{q_i}{r_i}, \tag{3.6.7}$$

which is the same as the work to bring the test charge into the system, as found in the first section of the chapter.

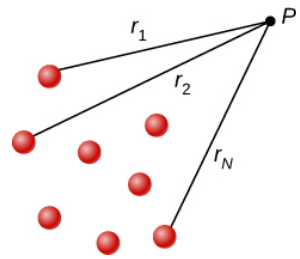


Figure 3.6.2: Notation for direct distances from charges to a space point P.

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 3.6.3: Electric Potential of a Dipole

Consider the dipole in Figure 3.6.3 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 cm); (b) (0, 0, –5.0 cm); (c) (3.0 cm, 0, 2.0 cm).

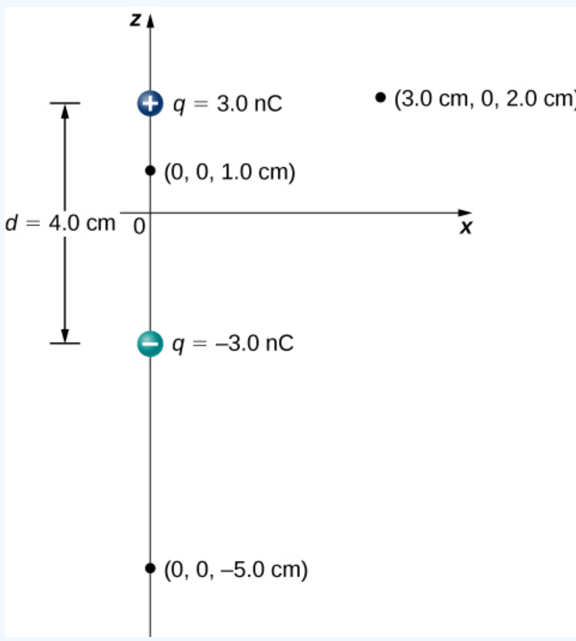


Figure 3.6.3: A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point P in space.

Strategy

Apply $V_p = k\sum_1^N \frac{q_i}{r_i}$ to each of these three points.

Solution

a. $V_p = k\sum_1^N \frac{q_i}{r_i} = (9.0\times 10^9\ \text{N}\cdot\text{m}^2/\text{C}^2)\left(\frac{3.0\ \text{nC}}{0.010\ \text{m}} - \frac{3.0\ \text{nC}}{0.030\ \text{m}}\right) = 1.8\times 10^3\ \text{V}$

b. $V_p = k\sum_1^N \frac{q_i}{r_i} = (9.0\times 10^9\ \text{N}\cdot\text{m}^2/\text{C}^2)\left(\frac{3.0\ \text{nC}}{0.070\ \text{m}} - \frac{3.0\ \text{nC}}{0.030\ \text{m}}\right) = -5.1\times 10^2\ \text{V}$

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 3.6.1

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 $z >> d$, again the potential goes to zero due to cancellation.

3.7: Common Models of Electric Potential

Learning Objectives

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

- List and apply common models for the electric potential of continuous charge distributions.

Continuous Charge Distributions

As we saw previously in Common Models of Electric Field, there are situations where there are so many charges that the system can be modeled as a continuous distribution of charges. In other words, the superposition of electric potential as a sum of discrete contributions

$$V_p = k\sum \frac{q_i}{r_i}. \tag{3.7.1}$$

can be modeled instead as a continuous charge distribution consisting of a collection of infinitesimally separated individual points. This approach yields the integral

$$V_p = \int \frac{dq}{r} \tag{3.7.2}$$

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 before, the infinitesimal charges are given by

$$\underbrace{dq = \lambda\, dl}_{\text{one dimension}} \tag{3.7.3}$$

$$\underbrace{dq = \sigma\, dA}_{\text{two dimensions}} \tag{3.7.4}$$

where λ is linear charge density and σ is the charge per unit area.

In this chapter, we will summarize several models corresponding to the most common geometric distributions of charge. In each of these models, the electric potential can be written in a closed-form expression for a specified region of space around the charge distribution. Students who are interested in the details of the calculations should refer to the section Calculating Electric Potentials of Charge Distributions in the chapter Direct Calculation of Electrical Quantities from Charge Distributions.

Electric Potentials of Common Charge Distributions

Finite Line Segment of Charge

Suppose a finite line segment of charge is located along the y-axis as illustrated in Fig. $3.7.1$. Assume that the segment has charge density λ (e.g., in coulombs per meters). The electric potential along the perpendicular bisector of the segment at a distance x away from the segment is

$$V(x) = k\lambda \ln\left[\frac{L+\sqrt{L^2+4x^2}}{-L+\sqrt{L^2+4x^2}}\right] \tag{3.7.5}$$

.

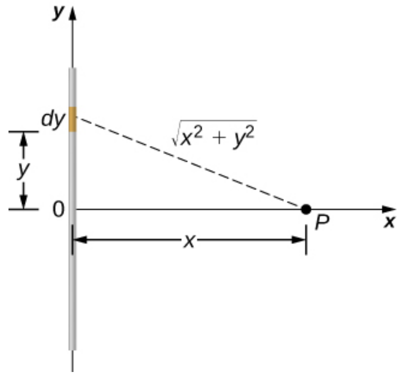


Figure 3.7.1: Geometry for calculating electric potential due to a finite line of uniform charge density.

Ring of Charge

If we take a finite line segment with linear charge density λ and total charge q_0 and wrap it into a closed circular loop, then we get a ring of charge of radius R (Fig. 3.7.2). The electric field at a distance z from the plane of the ring along its symmetry axis is

$$V(z) = \frac{2\pi k\lambda R}{\sqrt{z^2+R^2}} = k\frac{q_{tot}}{\sqrt{z^2+R^2}} \tag{3.7.6}$$

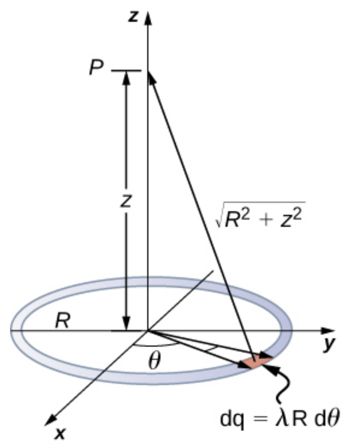


Figure 3.7.2: Geometry for calculating electric potential due to a ring of uniform charge density.

Disk of Charge

Once we know the electric field along the axis of a ring of charge, we can use this result to find the electric field along the axis of a disk of charge because a disk of charge can be modeled as a set of rings of charge of increasing radius (Fig. 3.7.3. The electric field for a disk of radius R and uniform surface charge density σ (coulombs per m^2) along the axis of the disk at a distance z from the disk is

$$V(z) = k2\pi\sigma(\sqrt{z^2+R^2} - \sqrt{z^2}) \tag{3.7.7}$$

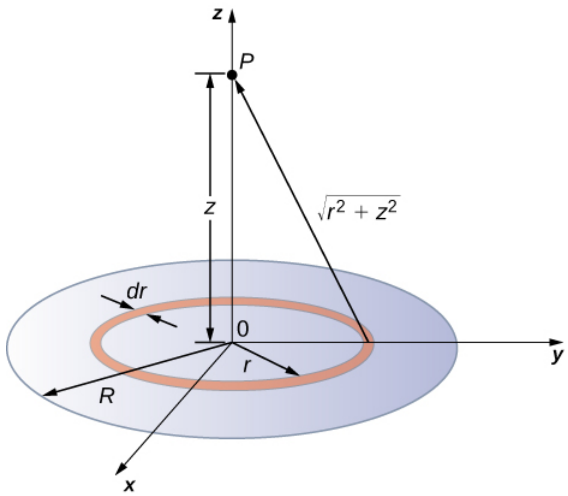


Figure 3.7.3: Geometry for calculating the electric potential due to a disk of uniform surface charge density.

Infinite Line of Charge

Consider an infinite line of charge with linear charge density λ aligned along the y-axis. Assume the reference potential $V_R = 0$ V at $x = 1$ m from the line of charge. The electric potential at a distance x in meters perpendicular to the line is

$$V(x) = -2k\lambda \ln x. \tag{3.7.8}$$

Note that unlike the result for the finite line segment Equation 3.7.5 where the reference potential for zero voltage was taken at infinity, here the reference voltage was taken at a finite distance of 1 meter from the wire. This specification is necessary because the charge itself reaches to infinity, and therefore the potential cannot be set to zero there.

Even though no actual infinite line charge exists, the result of Equation 3.7.8 can still provide a simple and good approximation for the potential provided that the separation distance to the test location is small compared to the length of the line charge.

3.8: Electric Potential (Summary)

Key Terms

Chapter 3

Term	Definition
electric dipole	system of two equal but opposite charges a fixed distance apart
electric potential	potential energy per unit charge
electric potential difference	the change in potential energy of a charge q moved between two points, divided by the charge.
electric potential energy	potential energy stored in a system of charged objects due to the charges
electron-volt	energy given to a fundamental charge accelerated through a potential difference of one volt
voltage	change in potential energy of a charge moved from one point to another, divided by the charge; units of potential difference are joules per coulomb, known as volt

Key Equations

Chapter 3

Description	Equation
Potential energy of a two-charge system	$U(r) = k\frac{qQ}{r}$
Work done to assemble a system of charges	$W_{12\cdots N} = \frac{k}{2}\sum_{i}^{N}\sum_{j}^{N}\frac{q_iq_j}{r_{ij}} \text{ for } i \neq j$
Potential difference	$\Delta V = \frac{\Delta U}{q} \text{ or } \Delta U = q\Delta V$
Electric potential	$V = \frac{U}{q} = -\int_R^P \vec{E}\cdot\vec{dl}$
Electric potential of a point charge	$V = \frac{kq}{r}$
Electric potential of a system of point charges	$V_P = k\sum_{1}^{N}\frac{q_i}{r_i}$
Electric potential of a continuous charge distribution	$V_P = k\int\frac{dq}{r}$

Summary

Work and Energy

- The work done by a force, acting over a finite path, is the integral of the infinitesimal increments of work done along the path, which are given by the dot product of the force and the infinitesimal displacements.
- The kinetic energy of a particle is the product of one-half its mass and the square of its speed (for non-relativistic speeds), and the kinetic energy of a system is the sum of the kinetic energies of all the particles in the system.
- The integral for the net work done on the particle is equal to the change in the particle's kinetic energy. This is the work-kinetic energy theorem.
- For a single-particle system, the difference of potential energy is the opposite of the work done by the forces acting on the particle as it moves from one position to another.

- A conservative force is one for which the work done is independent of path. Equivalently, a force is conservative if the work done over any closed path is zero.
- If non-conservative forces do no work and there are no external forces, the mechanical energy of a particle stays constant. This is a statement of the conservation of mechanical energy and there is no change in the total mechanical energy.

Electric Potential Energy

- The work done to move a charge from point A to B in an electric field is path independent, and the work around a closed path is zero. Therefore, the electric field and electric force are conservative.
- The superposition principle holds for electric potential energy; the potential energy of a system of multiple charges is the sum of the potential energies of the individual pairs.

Electric Potential Energy of Point Charges

- We can define an electric potential energy, which between point charges is $U(r)=k\frac{qQ}{r}$, with the zero reference taken to be at infinity.

Electric Potential

- Electric potential is potential energy per unit charge.
- The potential difference between points A and B, V_B-V_A , that is, the change in potential of a charge q moved from A to B, is equal to the change in potential energy divided by the charge.
- Potential difference is commonly called voltage, represented by the symbol ΔV:

$$\Delta V=\frac{\Delta U}{q} \text{ or } \Delta U=q\Delta V.$$

- An electron-volt is the energy given to a fundamental charge accelerated through a potential difference of 1 V. In equation form,

$$1\text{ eV}=(1.60\times10^{-19}\text{ C})(1\text{ V})=(1.60\times10^{-19}\text{ C})(1\text{ J/C})=1.60\times10^{-19}\text{ J}.$$

Electric Potential of a Point Charge

- Electric potential is a scalar, whereas electric field is a vector.
- The addition of voltages as numbers gives the voltage due to a combination of point charges, allowing us to use the principle of superposition: $V_P=k\sum_1^N\frac{q_i}{r_i}$.
- An electric dipole consists of two equal and opposite charges a fixed distance apart, with a dipole moment $\vec{p}=q\vec{d}$.

Common Models of Electric Potential

- Continuous charge distributions may be calculated with $V_P=k\int\frac{dq}{r}$.
- Results are for the electric potential provided for common continuous charge distributions including a line segment, ring, disk, and infinite line.

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).

3.9: The Electric Potential (Exercises)

Conceptual Questions

Electric Potential Energy

1. Would electric potential energy be meaningful if the electric field were not conservative?

2. Why do we need to be careful about work done **on** the system versus work done **by** the system in calculations?

3. Does the order in which we assemble a system of point charges affect the total work done?

Electric Potential Energy
Electric Potential Energy of Point Charges
Electric Potential
Electric Potential of a Point Charge

4. Discuss how potential difference and electric field strength are related. Give an example.

5. What is the strength of the electric field in a region where the electric potential is constant?

6. If a proton is released from rest in an electric field, will it move in the direction of increasing or decreasing potential? Also answer this question for an electron and a neutron. Explain why.

7. Voltage is the common word for potential difference. Which term is more descriptive, voltage or potential difference?

8. If the voltage between two points is zero, can a test charge be moved between them with zero net work being done? Can this necessarily be done without exerting a force? Explain.

9. What is the relationship between voltage and energy? More precisely, what is the relationship between potential difference and electric potential energy?

10. Voltages are always measured between two points. Why?

11. How are units of volts and electron-volts related? How do they differ?

12. Can a particle move in a direction of increasing electric potential, yet have its electric potential energy decrease? Explain

Problems

Electric Potential Energy
Electric Potential Energy of Point Charges

29. Consider a charge $Q_1 (+5.0\mu C)$ fixed at a site with another charge Q_2 (charge $+3.0\mu C$, mass $6.0\mu g$) moving in the neighboring space. (a) Evaluate the potential energy of Q_2 when it is 4.0 cm from Q_1. (b) If Q_2 starts from rest from a point 4.0 cm from Q_1, what will be its speed when it is 8.0 cm from Q_1? (**Note:** Q_1 is held fixed in its place.)

30. Two charges $Q_1 (+2.00\mu C)$ and $Q_2 (+2.00\mu C)$ are placed symmetrically along the **x**-axis at $x = \pm 3.00 cm$. Consider a charge Q_3 of charge $+4.00\mu C$ and mass 10.0 mg moving along the **y**-axis. If Q_3 starts from rest at $y = 2.00 cm$, what is its speed when it reaches $y = 4.00 cm$?

31. To form a hydrogen atom, a proton is fixed at a point and an electron is brought from far away to a distance of $0.529 \times 10^{-10} m$, the average distance between proton and electron in a hydrogen atom. How much work is done?

32. (a) What is the average power output of a heart defibrillator that dissipates 400 J of energy in 10.0 ms? (b) Considering the high-power output, why doesn't the defibrillator produce serious burns?

Electric Potential
Electric Potential of a Point Charge

33. Find the ratio of speeds of an electron and a negative hydrogen ion (one having an extra electron) accelerated through the same voltage, assuming non-relativistic final speeds. Take the mass of the hydrogen ion to be $1.67 \times 10^{-27} kg$.

34. An evacuated tube uses an accelerating voltage of 40 kV to accelerate electrons to hit a copper plate and produce X-rays. Non-relativistically, what would be the maximum speed of these electrons?

35. Show that units of V/m and N/C for electric field strength are indeed equivalent.

36. What is the strength of the electric field between two parallel conducting plates separated by 1.00 cm and having a potential difference (voltage) between them of $1.50 \times 10^4 V$?

37. The electric field strength between two parallel conducting plates separated by 4.00 cm is $7.50 \times 10^4 V$.

(a) What is the potential difference between the plates?

(b) The plate with the lowest potential is taken to be zero volts. What is the potential 1.00 cm from that plate and 3.00 cm from the other?

38. The voltage across a membrane forming a cell wall is 80.0 mV and the membrane is 9.00 nm thick. What is the electric field strength? (The value is surprisingly large, but correct.) You may assume a uniform electric field.

39. Two parallel conducting plates are separated by 10.0 cm, and one of them is taken to be at zero volts.

(a) What is the electric field strength between them, if the potential 8.00 cm from the zero volt plate (and 2.00 cm from the other) is 450 V?

(b) What is the voltage between the plates?

40. Find the maximum potential difference between two parallel conducting plates separated by 0.500 cm of air, given the maximum sustainable electric field strength in air to be $3.0 \times 10^6 V/m$.

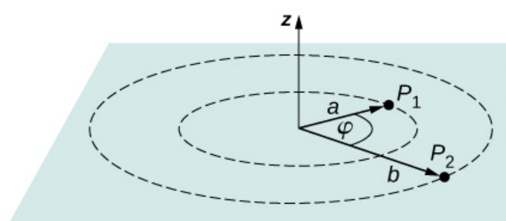
41. An electron is to be accelerated in a uniform electric field having a strength of $2.00 \times 10^6 V/m$.

(a) What energy in keV is given to the electron if it is accelerated through 0.400 m?

(b) Over what distance would it have to be accelerated to increase its energy by 50.0 GeV?

42. Use the definition of potential difference in terms of electric field to deduce the formula for potential difference between $r = r_a$ and $r = r_b$ for a point charge located at the origin. Here r is the spherical radial coordinate.

43. The electric field in a region is pointed away from the **z**-axis and the magnitude depends upon the distance s from the axis. The magnitude of the electric field is given as $E = \frac{\alpha}{s}$ where α is a constant. Find the potential difference between points P_1 and P_2, explicitly stating the path over which you conduct the integration for the line integral.



44. Singly charged gas ions are accelerated from rest through a voltage of 13.0 V. At what temperature will the average kinetic energy of gas molecules be the same as that given these ions?

Common Models of Electric Potential

45. A 0.500-cm-diameter plastic sphere, used in a static electricity demonstration, has a uniformly distributed 40.0-pC charge on its surface. What is the potential near its surface?

46. How far from a $1.00 - \mu C$ point charge is the potential 100 V? At what distance is it $2.00 \times 10^2 V$?

47. If the potential due to a point charge is $5.00 \times 10^2 V$ at a distance of 15.0 m, what are the sign and magnitude of the charge?

48. In nuclear fission, a nucleus splits roughly in half. (a) What is the potential $2.00 \times 10^{-14} m$ from a fragment that has 46 protons in it? (b) What is the potential energy in MeV of a similarly charged fragment at this distance?

49. A research Van de Graaff generator has a 2.00-m-diameter metal sphere with a charge of 5.00 mC on it. Assume the potential energy is zero at a reference point infinitely far away from the Van de Graaff.

(a) What is the potential near its surface?

(b) At what distance from its center is the potential 1.00 MV?

(c) An oxygen atom with three missing electrons is released near the Van de Graaff generator. What is its kinetic energy in MeV when the atom is at the distance found in part b?

50. An electrostatic paint sprayer has a 0.200-m-diameter metal sphere at a potential of 25.0 kV that repels paint droplets onto a grounded object.

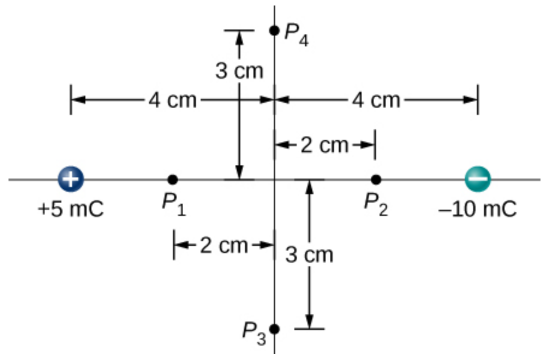
(a) What charge is on the sphere?

(b) What charge must a 0.100-mg drop of paint have to arrive at the object with a speed of 10.0 m/s?

51. (a) What is the potential between two points situated 10 cm and 20 cm from a $3.0-\mu C$ point charge?

(b) To what location should the point at 20 cm be moved to increase this potential difference by a factor of two?

52. Find the potential at points P_1, P_2, P_3, and P_4 in the diagram due to the two given charges.



53. Two charges $-2.0\mu C$ and $+2.0\mu C$ are separated by 4.0 cm on the **z**-axis symmetrically about origin, with the positive one uppermost. Two space points of interest P_1 and P_2 are located 3.0 cm and 30 cm from origin at an angle 30° with respect to the **z**-axis. Evaluate electric potentials at P_1 and P_2 in two ways:

(a) Using the exact formula for point charges, and

(b) using the approximate dipole potential formula.

54. (a) Plot the potential of a uniformly charged 1-m rod with 1 C/m charge as a function of the perpendicular distance from the center. Draw your graph from $s=0.1m$ to $s=1.0m$.

(b) On the same graph, plot the potential of a point charge with a 1-C charge at the origin.

(c) Which potential is stronger near the rod? (d) What happens to the difference as the distance increases? Interpret your result.

Additional Problems

88. A 12.0-V battery-operated bottle warmer heats 50.0 g of glass, $2.50\times10^2 g$ of baby formula, and $2.00\times10^2 g$ of aluminum from **20.0°C** to **90.0°C**.

(a) How much charge is moved by the battery?

(b) How many electrons per second flow if it takes 5.00 min to warm the formula? (Hint: Assume that the specific heat of baby formula is about the same as the specific heat of water.)

89. A battery-operated car uses a 12.0-V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a $2.00\times10^2-m$ high hill, and finally cause it to travel at a constant 25.0 m/s while climbing with $5.00\times10^2-N$ force for an hour.

90. (a) Find the voltage near a 10.0 cm diameter metal sphere that has 8.00 C of excess positive charge on it.

(b) What is unreasonable about this result?

(c) Which assumptions are responsible?

91. A uniformly charged half-ring of radius 10 cm is placed on a nonconducting table. It is found that 3.0 cm above the center of the half-ring the potential is –3.0 V with respect to zero potential at infinity. How much charge is in the half-ring?

92. A glass ring of radius 5.0 cm is painted with a charged paint such that the charge density around the ring varies continuously given by the following function of the polar angle $\theta, \lambda = (3.0 \times 10^{-6} C/m) cos^2\theta.$. Find the potential at a point 15 cm above the center.

93. A CD disk of radius ($R = 3.0cm$) is sprayed with a charged paint so that the charge varies continually with radial distance rfrom the center in the following manner: $\sigma = -(6.0C/m)r/R$ Find the potential at a point 4 cm above the center.

94. (a) What is the final speed of an electron accelerated from rest through a voltage of 25.0 MV by a negatively charged Van de Graff terminal? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

95. A large metal plate is charged uniformly to a density of $\sigma = 2.0 \times 10^{-9} C/m^2$. How far apart are the equipotential surfaces that represent a potential difference of 25 V?

96. Your friend gets really excited by the idea of making a lightning rod or maybe just a sparking toy by connecting two spheres as shown in Figure 7.39, and making R_2 so small that the electric field is greater than the dielectric strength of air, just from the usual 150 V/m electric field near the surface of the Earth. If R_1 is 10 cm, how small does R_2 need to be, and does this seem practical? (**Hint**: recall the calculation for electric field at the surface of a conductor from Gauss's Law.)

97. (a) Find $x >> L$ limit of the potential of a finite uniformly charged rod and show that it coincides with that of a point charge formula. (b) Why would you expect this result?

98. A small spherical pith ball of radius 0.50 cm is painted with a silver paint and then $-10\mu C$ of charge is placed on it. The charged pith ball is put at the center of a gold spherical shell of inner radius 2.0 cm and outer radius 2.2 cm.

(a) Find the electric potential of the gold shell with respect to zero potential at infinity.

(b) How much charge should you put on the gold shell if you want to make its potential 100 V?

99. 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,

(a) what is the potential difference between the plates?

(b) What is the potential difference between the positive plate and a point 1.25 cm from it that is between the plates?

100. A point charge of $q = 5.0 \times 10^{-8} C$ 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 potential at

(a) $r = 4.0cm$,

(b) $r = 8.0cm$,

(c) $r = 12.0cm$.

101. Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface.

(a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center?

(b) What acceleration will the field produce on a free electron near Earth's surface?

(c) What mass object with a single extra electron will have its weight supported by this field?

102. Point charges of $25.0\mu C$ and $45.0\mu C$ are placed 0.500 m apart.

(a) At what point along the line between them is the electric field zero?

(b) What is the electric field halfway between them?

103. What can you say about two charges q_1 and q_2, if the electric field one-fourth of the way from q_1 to q_2 is zero?

104. Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is $0.530 \times 10^{-10} m$. You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.

105. An electron has an initial velocity of $5.00 \times 10^6 m/s$ in a uniform $2.00 \times 10^5 - N/C$ electric field. The field accelerates the electron in the direction opposite to its initial velocity.

(a) What is the direction of the electric field?

(b) How far does the electron travel before coming to rest?

(c) How long does it take the electron to come to rest?

(d) What is the electron's velocity when it returns to its starting point?

Challenge Problems

106. Three Na^+ and three Cl^- ions are placed alternately and equally spaced around a circle of radius 50 nm. Find the electrostatic energy stored.

107. Look up (presumably online, or by dismantling an old device and making measurements) the magnitude of the potential deflection plates (and the space between them) in an ink jet printer. Then look up the speed with which the ink comes out the nozzle. Can you calculate the typical mass of an ink drop?

108. Use the electric field of a finite sphere with constant volume charge density to calculate the electric potential, throughout space. Then check your results by calculating the electric field from the potential.

109. Calculate the electric field of a dipole throughout space from the potential.

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).

3.10: The Electric Potential (Answers)

Conceptual Questions

1. No. We can only define potential energies for conservative fields.

3. No, though certain orderings may be simpler to compute.

5. The electric field strength is zero because electric potential differences are directly related to the field strength. If the potential difference is zero, then the field strength must also be zero.

7. Potential difference is more descriptive because it indicates that it is the difference between the electric potential of two points.

9. They are very similar, but potential difference is a feature of the system; when a charge is introduced to the system, it will have a potential energy which may be calculated by multiplying the magnitude of the charge by the potential difference.

11. An electron-volt is a volt multiplied by the charge of an electron. Volts measure potential difference, electron-volts are a unit of energy.

Problems

29. a. $U=3.4J$;

b. $\frac{1}{2}mv^2=kQ_1Q_2(\frac{1}{r_i}-\frac{1}{r_f})\rightarrow v=750m/s$

31. $U=4.36\times10^{-18}J$

33. $\frac{1}{2}m_ev_e^2=qV$, $\frac{1}{2}m_Hv_H^2=qV$, so that $\frac{m_ev_e^2}{m_Hv_H^2}=1$ or $\frac{v_e}{v_H}=42.8$.

35. $1V=1J/C; 1J=1N\cdot m\rightarrow 1V/m=1N/C$

37. a. $V_{AB}=3.00kV$;

b. $V_{AB}=7.50kV$

39. a. $V_{AB}=Ed\rightarrow E=5.63kV/m$;

b. $V_{AB}=563V$

41. a. $\Delta K=q\Delta V$ and $V_{AB}=Ed$,so that $\Delta K=800keV$;

b. $d=25.0km$

43. One possibility is to stay at constant radius and go along the arc from P_1 to P_2, which will have zero potential due to the path being perpendicular to the electric field. Then integrate from a to b: $V_{ab}=\alpha ln(\frac{b}{a})$

45. $V=144V$

47. $V=\frac{kQ}{r}\rightarrow Q=8.33\times10^{-7}C$; The charge is positive because the potential is positive.

49. a. $V=45.0MV$;

b. $V=\frac{kQ}{r}\rightarrow r=45.0m$;

c. $\Delta U=132MeV$

51. $V=kQ/r$; a. Relative to origin, find the potential at each point and then calculate the difference. $\Delta V=135\times10^3V$;

b. To double the potential difference, move the point from 20 cm to infinity; the potential at 20 cm is halfway between zero and that at 10 cm.

Additional Problems

89. $E_{tot} = 4.67 \times 10^7 J \; E_{tot} = qV \to q = \dfrac{E_{tot}}{V} = 3.89 \times 10^6 C$

91. $V_P = k\dfrac{q_{tot}}{\sqrt{z^2+R^2}} \to q_{tot} = -3.5 \times 10^{-11} C$

93. $V_P = -2.2 GV$

95. Recall from the previous chapter that the electric field $E_P = \dfrac{\sigma_0}{2\varepsilon_0}$ is uniform throughout space, and that for uniform fields we have $E = -\dfrac{\Delta V}{\Delta z}$ for the relation. Thus, we get $\dfrac{\sigma}{2\varepsilon_0} = \dfrac{\Delta V}{\Delta z} \to \Delta z = 0.22 m$ for the distance between 25-V equipotentials.

97. a. Take the result from Example 7.13, divide both the numerator and the denominator by x, take the limit of that, and then apply a Taylor expansion to the resulting log to get: $V_P \approx k\lambda \dfrac{L}{x}$;

b. which is the result we expect, because at great distances, this should look like a point charge of $q = \lambda L$

99. a. $V = 9.0 \times 10^3 V$;

b. $-9.0 \times 10^3 V(\dfrac{1.25 cm}{2.0 cm}) = -5.7 \times 10^3 V$

101. a. $E = \dfrac{KQ}{r^2} \to Q = -6.76 \times 10^5 C$;

b. $F = ma = qE \to a = \dfrac{qE}{m} = 2.63 \times 10^{13} m/s^2 (upwards)$;

c. $F = -mg = qE \to m = \dfrac{-qE}{g} = 2.45 \times 10^{-18} kg$

103. If the electric field is zero ¼ from the way of q_1 and q_2, then we know from $E = k\dfrac{Q}{r^2}$ that $|E_1| = |E_2| \to \dfrac{Kq_1}{x^2} = \dfrac{Kq^2}{(3x)^2}$ so that $\dfrac{q_2}{q_1} = \dfrac{(3x)^2}{x^2} = 9$; the charge q_2 is 9 times larger than q_1.

105. a. The field is in the direction of the electron's initial velocity.

b. $v^2 = v_0^2 + 2ax \to x = -\dfrac{v_0^2}{2a} (v = 0)$. Also, $F = ma = qE \to a = \dfrac{qE}{m}$, $x = 3.56 \times 10^{-4} m$;

c. $v_2 = v_0 + at \to t = -\dfrac{v_0 m}{qE} (v = 0), \therefore t = 1.42 \times 10^{-10} s$;

d. $v = -(\dfrac{2qEx}{m})^{1/2} - 5.00 \times 10^6 m/s$ (opposite its initial velocity)

Challenge Problems

107. Answers will vary. This appears to be proprietary information, and ridiculously difficult to find. Speeds will be 20 m/s or less, and there are claims of 10^{-7} grams for the mass of a drop.

109. Apply $\vec{E} = -\vec{\nabla} V$ with $\vec{\nabla} = \hat{r}\dfrac{\partial}{\partial r} + \hat{\theta}\dfrac{1}{r}\dfrac{\partial}{\partial \theta} + \hat{\varphi}\dfrac{1}{r sin\theta}\dfrac{\partial}{\partial \varphi}$ to the potential calculated earlier, $V_P = k\dfrac{\vec{p}\cdot\hat{r}}{r^2}$ with $\vec{p} = q\vec{d}$, and assume that the axis of the dipole is aligned with the **z**-axis of the coordinate system. Thus, the potential is $V_P = k\dfrac{q\vec{d}\cdot\hat{r}}{r^2} = k\dfrac{qd cos\theta}{r^2}$.

$$\vec{E} = 2kqd(\frac{cos\theta}{r^3})\hat{r} + kqd(\frac{sin\theta}{r^3})\hat{\theta}$$

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).

CHAPTER OVERVIEW

4: Potential and Field Relationships

4.1: Electric Potential from Electric Field

> **Learning Objectives**
>
> By the end of this section, you will be able to:
>
> - Calculate electric potential change from the electric field in a region of space.
> - Relate the voltage drop across charged parallel plates to their intervening electric field.

So far, we have explored the relationship between voltage and energy. Now we want to explore the relationship between voltage and electric field. We will start with the general case for a non-uniform $\vec{E}$ field. Recall that our general formula for the potential energy of a test charge q at point P relative to reference point R is

$$U_p = -\int_R^P \vec{F}\cdot d\vec{l}. \tag{4.1.1}$$

When we substitute in the definition of electric field $(\vec{E}=\vec{F}/q)$, this becomes

$$U_p = -q\int_R^P \vec{E}\cdot d\vec{l}. \tag{4.1.2}$$

Applying our definition of potential $(V=U/q)$ to this potential energy, we find that, in general,

$$V_p = -\int_R^P \vec{E}\cdot d\vec{l}. \tag{4.1.3}$$

From our previous discussion of the potential energy of a charge in an electric field, the result is independent of the path chosen, and hence we can pick the integral path that is most convenient.

Consider the special case of a positive point charge q at the origin. To calculate the potential caused by q at a distance r from the origin relative to a reference of 0 at infinity (recall that we did the same for potential energy), let $P=r$ and $R=\infty$, with $d\vec{l}=d\vec{r}=\hat{r}dr$ and use $\vec{E}=\frac{kq}{r^2}\hat{r}$. When we evaluate the integral

$$V_p = -\int_R^P \vec{E}\cdot d\vec{l} \tag{4.1.4}$$

for this system, we have

$$V_r = -\int_\infty^r \frac{kq}{r^2}dr = \frac{kq}{r} - \frac{kq}{\infty} = \frac{kq}{r}. \tag{4.1.5}$$

This result,

$$V_r = \frac{kq}{r} \tag{4.1.6}$$

is the standard form of the potential of a point charge. This will be explored further in the next section.

To examine another interesting special case, suppose a uniform electric field $\vec{E}$ is produced by placing a potential difference (or voltage) ΔV across two parallel metal plates, labeled A and B (Figure 4.1.1). Examining this situation will tell us what voltage is needed to produce a certain electric field strength. It will also reveal a more fundamental relationship between electric potential and electric field.

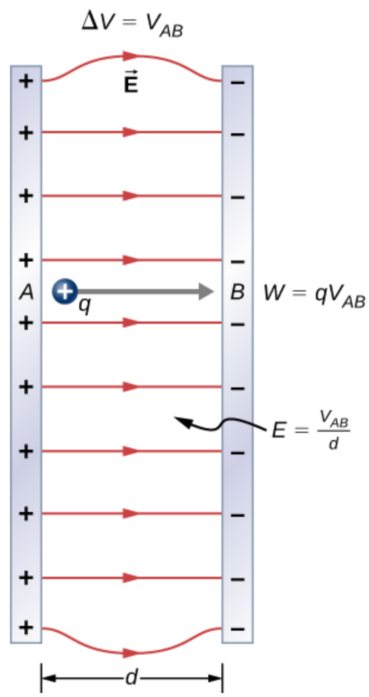


Figure 4.1.1: The relationship between V and E for parallel conducting plates is $E = V/d$. (Note that $\Delta V = V_{AB}$ in magnitude. For a charge that is moved from plate A at higher potential to plate B at lower potential, a minus sign needs to be included as follows: $-\Delta V = V_A - V_B = V_{AB}$.)

From a physicist's point of view, either ΔV or $\vec{E}$ can be used to describe any interaction between charges. However, ΔV is a scalar quantity and has no direction, whereas $\vec{E}$ is a vector quantity, having both magnitude and direction. (Note that the magnitude of the electric field, a scalar quantity, is represented by E.) The relationship between ΔV and $\vec{E}$ is revealed by calculating the work done by the electric force in moving a charge from point A to point B. But, as noted earlier, arbitrary charge distributions require calculus. We therefore look at a uniform electric field as an interesting special case.

The work done by the electric field in Figure 4.1.1 to move a positive charge q from A, the positive plate, higher potential, to B, the negative plate, lower potential, is

$$W = -\Delta U = -q\Delta V. \tag{4.1.7}$$

The potential difference between points A and B is

$$-\Delta V = -(V_B - V_A) = V_A - V_B = V_{AB}. \tag{4.1.8}$$

Entering this into the expression for work yields

$$W = qV_{AB}. \tag{4.1.9}$$

Work is $W = \vec{F} \cdot \vec{d} = Fd \cos\theta$: here $\cos\theta = 1$, since the path is parallel to the field. Thus, $W = Fd$. Since $F = qE$ we see that $W = qEd$.

Substituting this expression for work into the previous equation gives

$$qEd = qV_{AB}. \tag{4.1.10}$$

The charge cancels, so we obtain for the voltage between points A and B

$$\left.\begin{aligned} V_{AB} &= Ed \\ E &= \frac{V_{AB}}{d} \end{aligned}\right\} \text{(uniform } E\text{-field only)} \tag{4.1.11}$$

where d is the distance from A to B, or the distance between the plates in Figure 4.1.1. Note that this equation implies that the units for electric field are volts per meter. We already know the units for electric field are newtons per coulomb; thus, the following relation among units is valid:

$$1\ \mathrm{N/C} = 1\ \mathrm{V/m}. \tag{4.1.12}$$

Furthermore, we may extend this to the integral form. Substituting Equation 4.1.3 into our definition for the potential difference between points A and B, we obtain

$$V_{AB} = V_B - V_A = -\int_R^B \vec{E}\cdot d\vec{l} + \int_R^A \vec{E}\cdot d\vec{l} \tag{4.1.13}$$

which simplifies to

$$V_B - V_A = -\int_A^B \vec{E}\cdot d\vec{l}. \tag{4.1.14}$$

As a demonstration, from this we may calculate the potential difference between two points (A and B) equidistant from a point charge q at the origin, as shown in Figure 4.1.2.

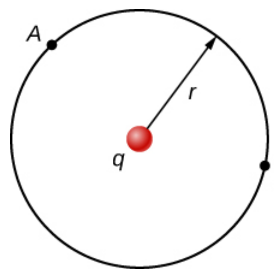


Figure 4.1.2: The arc for calculating the potential difference between two points that are equidistant from a point charge at the origin.

To do this, we integrate around an arc of the circle of constant radius r between A and B, which means we let $d\vec{l} = r\hat{\varphi}d\varphi$, while using $\vec{E} = \frac{kq}{r^2}\hat{r}$. Thus,

$$\Delta V = V_B - V_A = -\int_A^B \vec{E}\cdot d\vec{l}. \tag{4.1.15}$$

for this system becomes

$$V_B - V_A = -\int_A^B \frac{kq}{r^2}\hat{r}\cdot r\hat{\varphi}d\varphi. \tag{4.1.16}$$

However, $\hat{r}\cdot\hat{\varphi}$ and therefore

$$V_B - V_A = 0. \tag{4.1.17}$$

This result, that there is no difference in potential along a constant radius from a point charge, will come in handy when we map potentials.

Example $4.1.1A$: What Is the Highest Voltage Possible between Two Plates?

Dry air can support a maximum electric field strength of about $3.0\times10^6\ \mathrm{V/m}$. Above that value, the field creates enough ionization in the air to make the air a conductor. This allows a discharge or spark that reduces the field. What, then, is the maximum voltage between two parallel conducting plates separated by 2.5 cm of dry air?

Strategy

We are given the maximum electric field E between the plates and the distance d between them. We can use the equation $V_{AB} = Ed$ to calculate the maximum voltage.

Solution

The potential difference or voltage between the plates is

$$V_{AB} = Ed. \tag{4.1.18}$$

Entering the given values for E and d gives

$$V_{AB} = (3.0 \times 10^6 \,\text{V/m})(0.025 \,\text{m}) = 7.5 \times 10^4 \,\text{V} \tag{4.1.19}$$

or

$$V_{AB} = 75 \,\text{kV}. \tag{4.1.20}$$

(The answer is quoted to only two digits, since the maximum field strength is approximate.)

Significance

One of the implications of this result is that it takes about 75 kV to make a spark jump across a 2.5-cm (1-in.) gap, or 150 kV for a 5-cm spark. This limits the voltages that can exist between conductors, perhaps on a power transmission line. A smaller voltage can cause a spark if there are spines on the surface, since sharp points have larger field strengths than smooth surfaces. Humid air breaks down at a lower field strength, meaning that a smaller voltage will make a spark jump through humid air. The largest voltages can be built up with static electricity on dry days (Figure $4.1.3$).

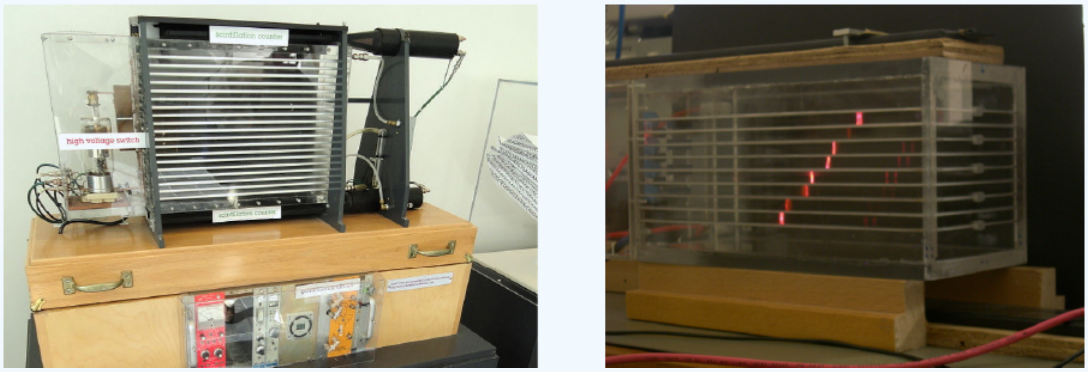


Figure $4.1.3$: A spark chamber is used to trace the paths of high-energy particles. Ionization created by the particles as they pass through the gas between the plates allows a spark to jump. The sparks are perpendicular to the plates, following electric field lines between them. The potential difference between adjacent plates is not high enough to cause sparks without the ionization produced by particles from accelerator experiments (or cosmic rays). This form of detector is now archaic and no longer in use except for demonstration purposes. (credit b: modification of work by Jack Collins)

Example $4.1.1B$: Field and Force inside an Electron Gun

An electron gun has parallel plates separated by 4.00 cm and gives electrons 25.0 keV of energy. (a) What is the electric field strength between the plates? (b) What force would this field exert on a piece of plastic with a $0.500-\mu\text{C}$ charge that gets between the plates?

Strategy

Since the voltage and plate separation are given, the electric field strength can be calculated directly from the expression $E = \frac{V_{AB}}{d}$. Once we know the electric field strength, we can find the force on a charge by using $\vec{F} = q\vec{E}$. Since the electric field is in only one direction, we can write this equation in terms of the magnitudes, $F = qE$.

Solution

a. The expression for the magnitude of the electric field between two uniform metal plates is

$$E = \frac{V_{AB}}{d}. \tag{4.1.21}$$

Since the electron is a single charge and is given 25.0 keV of energy, the potential difference must be 25.0 kV. Entering this value for V_{AB} and the plate separation of 0.0400 m, we obtain

$$E = \frac{25.0 \,\text{kV}}{0.0400 \,\text{m}} = 6.25 \times 10^5 \,\text{V/m}. \tag{4.1.22}$$

b. The magnitude of the force on a charge in an electric field is obtained from the equation

$$F = qE. \tag{4.1.23}$$

Substituting known values gives

$$F = (0.500 \times 10^{-6}\,\mathrm{C})(6.25 \times 10^{5}\,\mathrm{V/m}) = 0.313\,\mathrm{N}. \tag{4.1.24}$$

Significance

Note that the units are newtons, since $1\ \mathrm{V/m} = 1\ \mathrm{N/C}$. Because the electric field is uniform between the plates, the force on the charge is the same no matter where the charge is located between the plates.

Example 4.1.1C: Calculating Potential of a Point Charge

Given a point charge $q = +2.0\mathrm{nC}$ at the origin, calculate the potential difference between point P_1 a distance $a = 4.0\ \mathrm{cm}$ from q, and P_2 a distance $b = 12.0\ \mathrm{cm}$ from q, where the two points have an angle of $\varphi = 24^\circ$ between them (Figure 4.1.4).

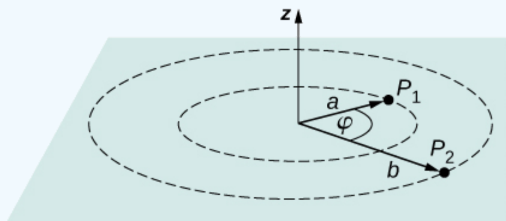


Figure 4.1.4: Find the difference in potential between P_1 and P_2.

Strategy

Do this in two steps. The first step is to use $V_B - V_A = -\int_A^B \vec{E} \cdot d\vec{l}$ and let $A = a = 4.0\ cm$ and $B = b = 12.0\ \mathrm{cm},$ with $d\vec{l} = d\vec{r} = \hat{r}\,dr$ and $\vec{E} = \frac{kq}{r^2}\hat{r}$. Then perform the integral. The second step is to integrate $V_B - V_A = -\int_A^B \vec{E} \cdot d\vec{l}$ around an arc of constant radius r, which means we let $d\vec{l} = r\vec{\varphi}d\varphi$ with limits $0 \leq \varphi \leq 24^\circ$, still using $\vec{E} = \frac{kq}{r^2}\hat{r}$.

Then add the two results together.

Solution

For the first part, $V_B - V_A = -\int_A^B \vec{E} \cdot d\vec{l}$ for this system becomes $V_b - V_a = -\int_a^b \frac{kq}{r^2}\hat{r} \cdot \hat{r}\,dr$ which computes to

$$\Delta V = -\int_a^b \frac{kq}{r^2} dr = kq\left[\frac{1}{a} - \frac{1}{b}\right]$$

$$= (8.99 \times 10^9\,\mathrm{N \cdot m^2/C^2})(2.0 \times 10^{-9}\,\mathrm{C})\left[\frac{1}{0.040\ \mathrm{m}} - \frac{1}{0.12\ \mathrm{m}}\right] = 300\ \mathrm{V}.$$

For the second step, $V_B - V_A = -\int_A^B \vec{E} \cdot d\vec{l}$ becomes $\Delta V = -\int_0^{24\pi/180} \frac{kq}{r^2}\hat{r} \cdot r\hat{\varphi}d\varphi$, but $\hat{r} \cdot \hat{\varphi} = 0$ and therefore $\Delta V = 0$. Adding the two parts together, we get 300 V.

Significance

We have demonstrated the use of the integral form of the potential difference to obtain a numerical result. Notice that, in this particular system, we could have also used the formula for the potential due to a point charge at the two points and simply taken the difference.

Exercise 4.1.4

From the examples, how does the energy of a lightning strike vary with the height of the clouds from the ground? Consider the cloud-ground system to be two parallel plates.

Answer

Given a fixed maximum electric field strength, the potential at which a strike occurs increases with increasing height above the ground. Hence, each electron will carry more energy. Determining if there is an effect on the total number of electrons lies in the future.

- **7.2: Electric Potential Energy** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.
- **7.3: Electric Potential and Potential Difference** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.

4.2: Electric Field from Electric Potential

Learning Objectives

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

- Explain how to calculate the electric field in a system from the given potential
- Calculate the electric field in a given direction from a given potential
- Calculate the electric field throughout space from a given potential

Recall that we were able, in certain systems, to calculate the potential by integrating over the electric field. As you may already suspect, this means that we may calculate the electric field by taking derivatives of the potential, although going from a scalar to a vector quantity introduces some interesting wrinkles. We frequently need $\vec{E}$ to calculate the force in a system; since it is often simpler to calculate the potential directly, there are systems in which it is useful to calculate V and then derive $\vec{E}$ from it.

In general, regardless of whether the electric field is uniform, it points in the direction of decreasing potential, because the force on a positive charge is in the direction of $\vec{E}$ and also in the direction of lower potential V. Furthermore, the magnitude of $\vec{E}$ equals the rate of decrease of V with distance. The faster V decreases over distance, the greater the electric field. This gives us the following result.

Relationship between Voltage and Uniform Electric Field

In equation form, the relationship between voltage and the s-component of a uniform electric field is

$$E_s = -\frac{\Delta V}{\Delta s} \tag{4.2.1}$$

where Δs is the distance over which the change in potential ΔV takes place. The minus sign tells us that E_s points in the direction of decreasing potential. The electric field is said to be the gradient (as in grade or slope) of the electric potential.

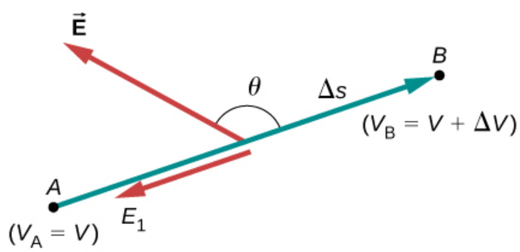


Figure 4.2.1: The electric field component along the displacement Δs is given by $E_s = -\Delta V/\Delta s$ in general (in the figure $E_1 = -\Delta V/\Delta x_1$ if the vector $\overrightarrow{AB}$ lies along the x_1-axis). Note that A and B are assumed to be so close together that the field is constant along Δs.

For continually changing potentials, ΔV and Δs become infinitesimals, and we need differential calculus to determine the electric field. As shown in Figure 4.2.1, if we treat the distance Δs as very small so that the electric field is essentially constant over it, we find that

$$E_s = -\frac{dV}{ds}. \tag{4.2.2}$$

Therefore, the electric field components in the Cartesian directions are given by

$$E_x = -\frac{\partial V}{\partial x},\ E_y = -\frac{\partial V}{\partial y},\ E_z = -\frac{\partial V}{\partial z}. \tag{4.2.3}$$

This allows us to define the “grad” or “del” vector operator, which allows us to compute the gradient in one step. In Cartesian coordinates, it takes the form

$$\vec{\nabla} = \hat{i}\frac{\partial}{\partial x} + \hat{j}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}. \tag{4.2.4}$$

With this notation, we can calculate the electric field from the potential with

$$\vec{E} = -\vec{\nabla} V, \tag{4.2.5}$$

a process we call **calculating the gradient of the potential**.

If we have a system with either cylindrical or spherical symmetry, we only need to use the del operator in the appropriate coordinates:

$$\vec{\nabla}_{cyl} = \underbrace{\hat{r}\frac{\partial}{\partial r} + \hat{\varphi}\frac{1}{r}\frac{\partial}{\partial \varphi} + \hat{z}\frac{\partial}{\partial z}}_{\text{Cylindrical}} \tag{4.2.6}$$

$$\vec{\nabla}_{sph} = \underbrace{\hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\varphi}\frac{1}{r\, sin\, \theta}\frac{\partial}{\partial \varphi}}_{\text{Spherical}} \tag{4.2.7}$$

Example 4.2.1: Electric Field of a Point Charge

Calculate the electric field of a point charge from the potential.

Strategy

The potential is known to be $V = k\frac{q}{r}$, which has a spherical symmetry. Therefore, we use the spherical del operator (Equation 4.2.7) into Equation 4.2.5:

$$\vec{E} = -\vec{\nabla}_{sph} V.$$

Solution

Performing this calculation gives us

$$\begin{aligned} \vec{E} &= -\left(\hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\varphi}\frac{1}{1\ \sin\, \theta}\frac{\partial}{\partial \varphi}\right) k\frac{q}{r} \\ &= -k\left(\hat{r}\frac{\partial}{\partial r}\frac{1}{r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta}\frac{1}{r} + \hat{\varphi}\frac{1}{1\ \sin\, \theta}\frac{\partial}{\partial \varphi}\frac{1}{r}\right). \end{aligned}$$

This equation simplifies to

$$\vec{E} = -kq\left(\hat{r}\frac{-1}{r^2} + \hat{\theta}0 + \hat{\varphi}0\right) = k\frac{q}{r^2}\hat{r}$$

as expected.

Significance

We not only obtained the equation for the electric field of a point particle that we've seen before, we also have a demonstration that $\vec{E}$ points in the direction of decreasing potential, as shown in Figure 4.2.2.

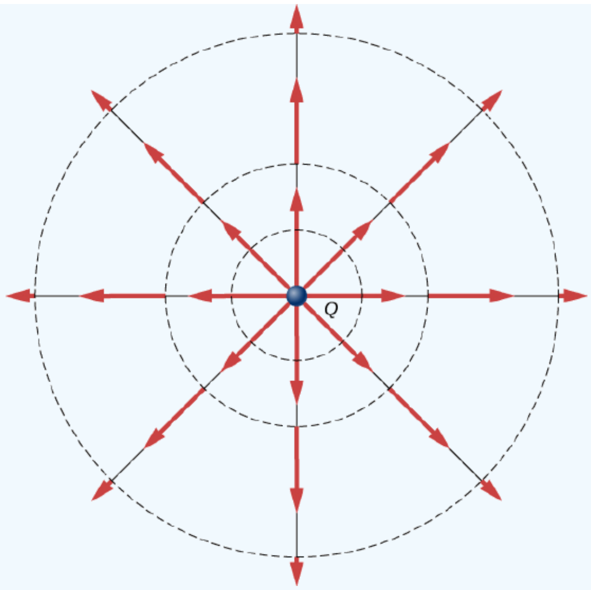


Figure 4.2.2: Electric field vectors inside and outside a uniformly charged sphere.

✔ Example 4.2.2: Electric Field of a Ring of Charge

Use the potential found previously to calculate the electric field along the axis of a ring of charge (Figure 4.2.3).

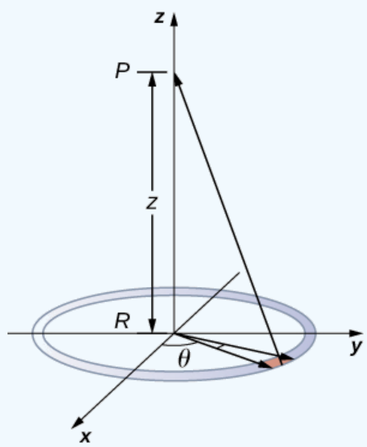


Figure 4.2.3: We want to calculate the electric field from the electric potential due to a ring charge.

Strategy

In this case, we are only interested in one dimension, the **z**-axis. Therefore, we use

$$E_z = -\frac{\partial V}{\partial z}$$

with the potential

$$V = k\frac{q_{tot}}{\sqrt{z^2+R^2}}$$

found previously.

Solution

Taking the derivative of the potential yields

$$\begin{aligned} E_z &= -\frac{\partial}{\partial z}\frac{kq_{tot}}{\sqrt{z^2+R^2}} \\ &= k\frac{q_{tot}z}{(z^2+R^2)^{3/2}}. \end{aligned}$$

Significance

Again, this matches the equation for the electric field found previously. It also demonstrates a system in which using the full del operator is not necessary.

Exercise 4.2.1

Which coordinate system would you use to calculate the electric field of a dipole?

Answer

Any, but cylindrical is closest to the symmetry of a dipole.

4.3: Equipotential Curves and Surfaces

Learning Objectives

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

- Define equipotential lines and and equipotential surfaces.
- Explain the relationships between equipotential lines and electric field lines.
- Map equipotential lines for one or two point charges.

Equipotentials

We can represent electric potentials (voltages) pictorially, just as we drew pictures to illustrate electric fields. This is not surprising, since the two concepts are related. Consider Figure $4.3.1$, which shows an isolated positive point charge and its electric field lines, which radiate out from a positive charge and terminate on negative charges. We use red arrows to represent the magnitude and direction of the electric field, and we use black lines to represent places where the electric potential is constant. These are called **equipotential surfaces** in three dimensions, or **equipotential lines** in two dimensions. The term **equipotential** is also used as a noun, referring to an equipotential line or surface. The potential for a point charge is the same anywhere on an imaginary sphere of radius r surrounding the charge. This is true because the potential for a point charge is given by $V = kq/r$ and thus has the same value at any point that is a given distance r from the charge. An equipotential sphere is a circle in the two-dimensional view of Figure $4.3.1$. Because the electric field lines point radially away from the charge, they are perpendicular to the equipotential lines.

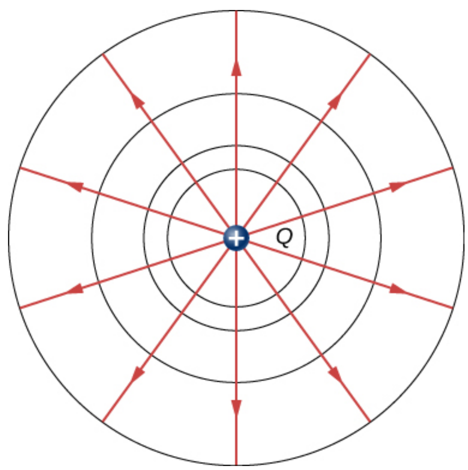


Figure 4.3.1: An isolated point charge Q with its electric field lines with red arrows and equipotential lines in black. The potential is the same along each equipotential line, meaning that no work is required to move a charge anywhere along one of those lines. Work is needed to move a charge from one equipotential line to another. Equipotential lines are perpendicular to electric field lines in every case. For a three-dimensional visualization, see http://falstad.com/vector3de/ (select Field selection: point charge, Display: Equipotentials or Display: Field Lines, and "No Slicing"; adjust potential or field line density, as desired).

It is important to note that *equipotential lines are always perpendicular to electric field lines*. No work is required to move a charge along an equipotential, since $\Delta V = 0$. Thus, the work is

$$W = -\Delta U = -q\Delta V = 0. \tag{4.3.1}$$

Work is zero if the direction of the force is perpendicular to the displacement. Force is in the same direction as E, so motion along an equipotential must be perpendicular to E. More precisely, work is related to the electric field by

$$\begin{aligned} W &= \vec{F} \cdot \vec{d} \\ &= q\vec{E} \cdot \vec{d} \\ &= qEd \cos \theta \\ &= 0. \end{aligned} \tag{4.3.2}$$

Note that in Equation 4.3.2, E and F symbolize the magnitudes of the electric field and force, respectively. Neither q nor E is zero and d is also not zero. So cos θ must be 0, meaning θ must be 90°. In other words, motion along an equipotential is perpendicular to $\vec{E}$.

In addition, the strength of the field is related to the density of the equipotentials, assuming that they have equal changes in potential. To see this, recall from Electric Field from Electric Potential that the component of the electric field in the s-direction is

$$E_s=-\frac{dV}{ds} \tag{4.3.3}$$

For small spatial intervals, the derivative can be approximated by a ratio of finite differences

$$E_s \approx -\frac{\Delta V}{\Delta s} \tag{4.3.4}$$

If ΔV is the same for adjacent equipotentials, then Equation 4.3.4 implies that the electric field strength will be higher for more closely-spaced equipotentials where Δs is smaller and lower where the spacing is higher, as illustrated in Figure 4.3.2. The negative sign means that the electric field will point in the direction of decreasing potential.

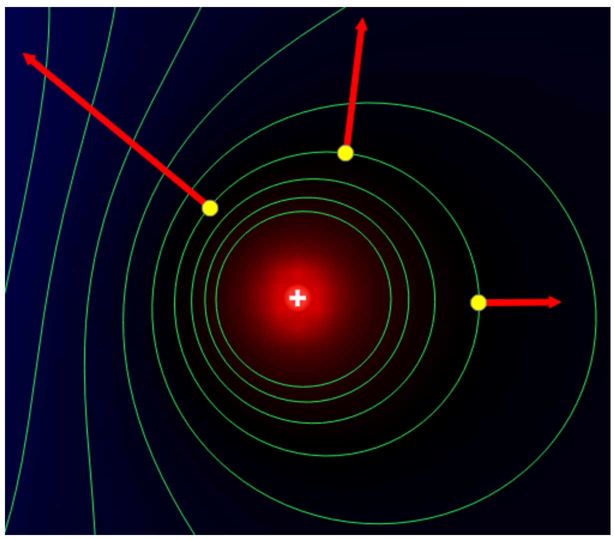


Figure 4.3.2: Relationship between equipotentials (green) and electric field (red arrows). The electric field direction is always perpendicular to the equipotentials and points toward decreasing potential. The electric field magnitude is inversely proportional to the spacing of equipotentials of equal intervals (Ronald Kumon, CC-BY-SA 4.0)

Equipotentials of Point-Charge Systems

Figure 4.3.3 shows the electric field and equipotential lines for two equal and opposite charges. Given the electric field lines, the equipotential lines can be drawn simply by making them perpendicular to the electric field lines.

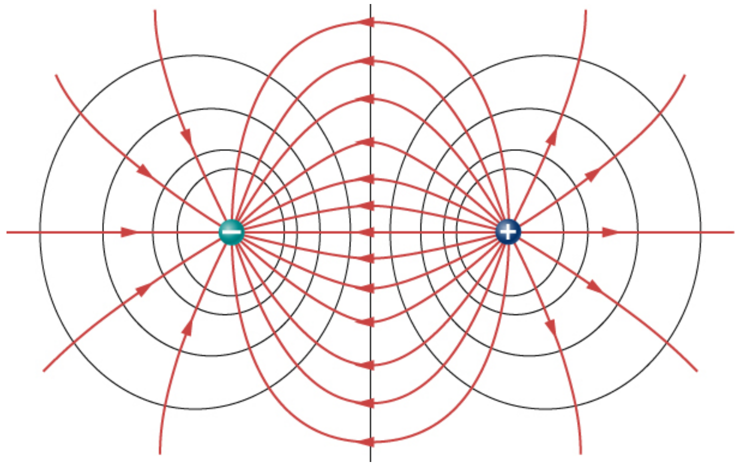


Figure 4.3.3: The electric field lines (red arrows) and equipotential lines (black) for two equal but opposite charges. The equipotential lines can be drawn by making them perpendicular to the electric field lines, if those are known. Note that the potential is greatest (most positive) near the positive charge and least (most negative) near the negative charge. For a three-dimensional version, see http://falstad.com/vector3de/ (select Field selection: dipole, Display: Equipotentials or Display: Field Lines, and "No Slicing"; adjust settings for charge separation, and potential or field line density, as desired).

Conversely, given the equipotential lines, as in Figure $4.3.4a$, the electric field lines can be drawn by making them perpendicular to the equipotentials, as in Figure $4.3.4b$.

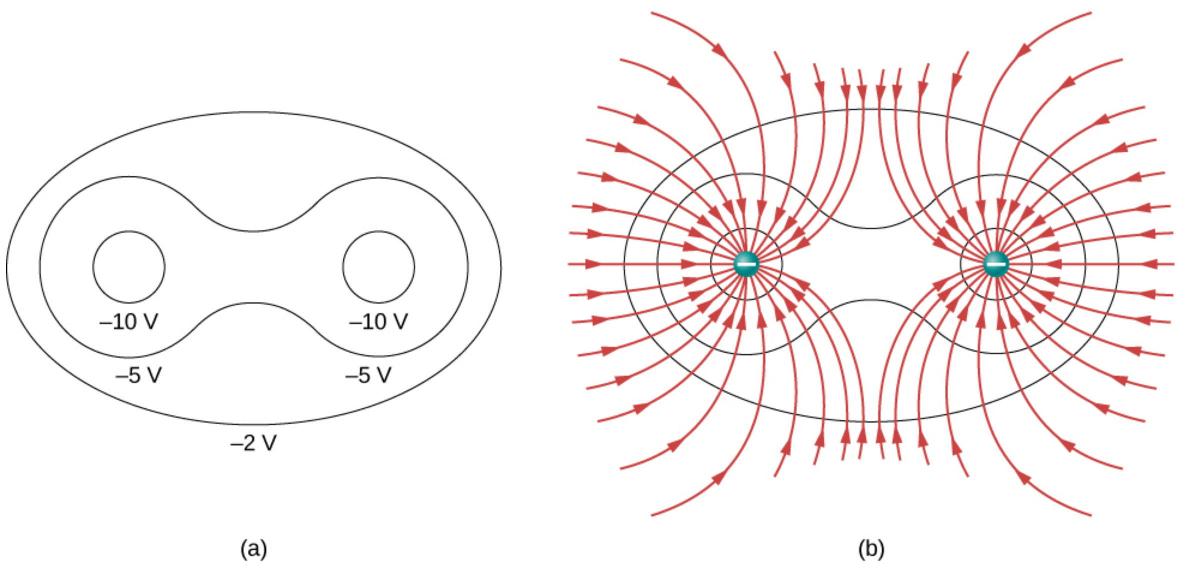


Figure 4.3.4: (a) These equipotential lines might be measured with a voltmeter in a laboratory experiment. (b) The corresponding electric field lines are found by drawing them perpendicular to the equipotentials. Note that these fields are consistent with two equal negative charges. For a three-dimensional version, see http://falstad.com/vector3de/ (select Field selection: point charge double, Display: Equipotentials or Display: Field Lines, and "No Slicing"; adjust settings for charge separation, and potential or field line density, as desired).

Electric equipotential lines are sometimes compared to equipotential isolines (lines of constant elevation) for gravity on hills . If the hill has any extent at the same slope, the isolines along that extent would be parallel to each other. Furthermore, in regions of constant slope, the isolines would be evenly spaced. An example of real topographic lines is shown in Figure $4.3.5$.

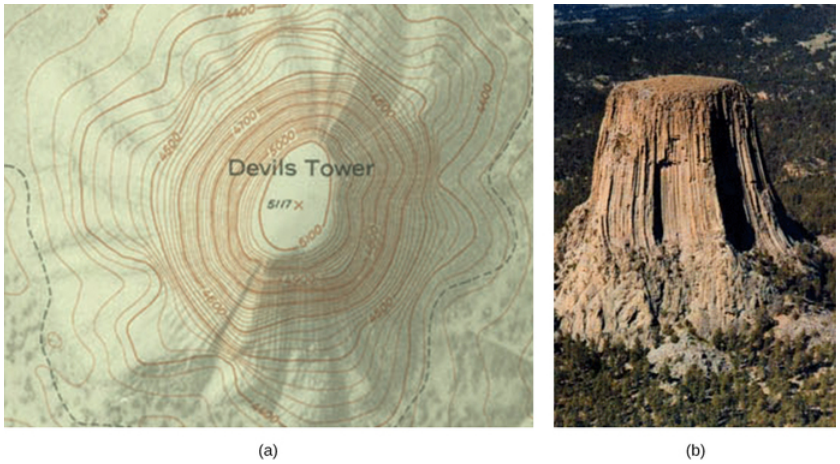


Figure 4.3.5. (a) A topographical map of Devil's Tower, Wyoming. Lines that are close together indicate very steep terrain. (b) A perspective photo of Devil's Tower shows just how steep its sides are. Notice the top of the tower has the same shape as the center of the topographical map

To improve your intuition, we show a three-dimensional variant of the potential in a system with two opposing charges. Figure 4.3.6 displays a three-dimensional map of electric potential, where lines on the map are for equipotential surfaces. The hill is at the positive charge, and the trough is at the negative charge. The potential is zero far away from the charges. Note that the cut off at a particular potential implies that the charges are on conducting spheres with a finite radius.

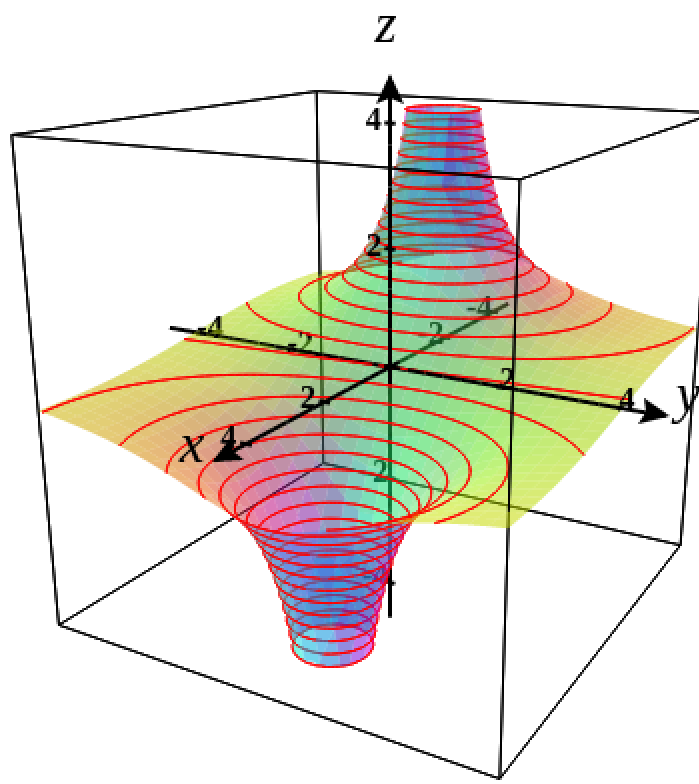


Figure 4.3.6: Electric potential map of two opposite charges of equal magnitude on conducting spheres. The potential is negative near the negative charge and positive near the positive charge. This dynamic image is powered by CalcPlot3D.

A two-dimensional map of the cross-sectional plane that contains both charges is shown in Figure $4.3.7$. The line that is equidistant from the two opposite charges corresponds to zero potential, since at the points on the line, the positive potential from the positive charge cancels the negative potential from the negative charge. Equipotential lines in the cross-sectional plane are closed loops, which are not necessarily circles, since at each point, the net potential is the sum of the potentials from each charge.

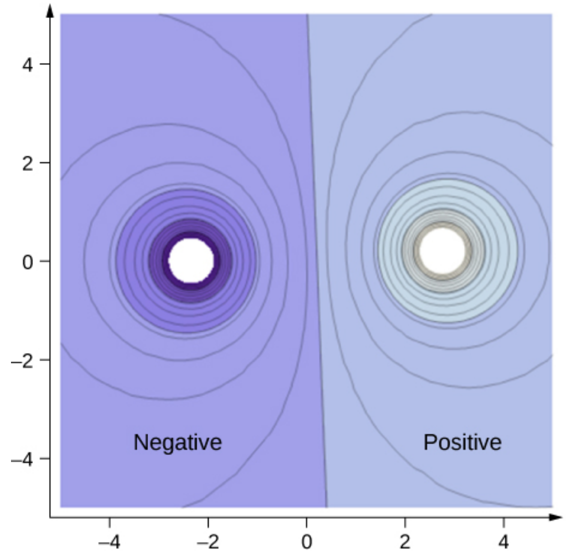


Figure $4.3.7$: A cross-section of the electric potential map of two opposite charges of equal magnitude. The potential is negative near the negative charge and positive near the positive charge.

You can draw equipotential lines using computer simulations. Try to reproduce the equipotential curves for the dipole and other charge configurations using the following simulation.

PhET Simulation: Charges and Fields

In this use of the "Charges and Fields" simulation, you will investigate equipotentials and the relationship between equipotentials and the electric field, starting with the example of an electric dipole.

Instructions:

1. Uncheck "Electric Field" and check "Voltage" and "Grid" in the checkboxes in the upper right corner. (When "Voltage" is selected, the strength of the voltage is indicated by red where the voltage is positive, blue where the voltage is negative. The brightness of the color indicates the magnitude of the voltage.)
2. Drag a positive charge and a negative charge into the box. Place them on the same horizontal line 4 major grid units apart (20 minor grid units apart)
3. Drag the purple voltage meter into the box. Measure the voltage at various points around the charges.
4. Put the voltage meter at the midpoint between the two charges. Click on the "pencil" button on the meter to draw the equipotential line corresponding to the voltage on the meter.
5. Move the voltage meter along the line connecting the two charges until the voltage changes by 2 V from its starting value. Click on the "pencil" button to draw another equipotential. Continue this process at 2 V intervals on either side of the midpoint until you have at least 11 equipotentials drawn. Check the "Values" box to show the potential values on each equipotential curve.
6. Drag an electric field sensor into the box and move it around an equipotential curve. Does the electric field vector have the same magnitude (length) at every point on the equipotential curve? If not, where is the electric field the strongest (longest vector)? What is the density of equipotential curves near this location? Where is the electric field the weakest? What is the density of equipotential curves near this location?
7. Finally, check the "Electric Field" box. How do the directions of the arrows relate to the equipotential curves?
8. Remove the positive charge by dragging it back to the box at the bottom of the simulation windows, and then replace it with a negative charge b. How do the equipotentials and electric field vectors change?

Source: Charges and Fields

Credit: Simulation by PhET Interactive Simulations, University of Colorado Boulder, licensed under CC-BY-4.0.

4.4: Conductors in Electrostatic Equilibrium

Learning Objectives

- List the six properties of a conductor in electrostatic equilibrium.
- Explain why no electric field may exist inside a conductor in electrostatic equilibrium.
- Explain why electric potential is constant throughout a conductor in electrostatic equilibrium.
- Explain how surface curvature affects surface charge and electric field in a conductor in electrostatic equilibrium.
- Describe the electric field and surface charge density in an empty cavity in a conductor in electrostatic equilibrium.

Recall that when a conductor is in **electrostatic equilibrium**, all the charges are stationary. Conductors in electrostatic equilibrium have several interesting properties:

Property 4.4.1: Electric Field Inside Conductors

The electric field is zero inside a conductor.

Proof: Suppose that there was a nonzero electric field inside the conductor. We know then that an electric force would be exerted on the charge. This force would cause the charges to accelerate. However, moving charges violate the assumption that the conductor is in electrostatic equilibrium. Thus, the electric field $\vec{E}_{in}=0$ inside the conductor must be zero.

Property 4.4.2: Electric Potential Inside Conductors

The electric potential is a constant everywhere inside and on the surface of the conductor.

Proof: Consider a path in the conductor between any two Points A and B. From Electric Potential from Electric Field, we know that the change of potential between the points is given by

$$\Delta V_{BA}=V_B-V_A=\int_A^B \vec{E}_{in}\cdot d\vec{l}=\int_A^B 0\cdot d\vec{l}=0.$$

Hence $V_B-V-A=0$ or $V_B=V_A$. But because the choice of Points A and B were arbitrary, it must be that the potential is everywhere the same throughout the conductor, and the surface of the conductor must also be an equipotential. (Note: This statement does not claim that the potential must be zero; it can be any constant value.)

Property 4.4.3: Electric Field on Surface of Conductors

The exterior electric field is perpendicular to the surface of the conductor, beginning or ending on charges on the surface.

Proof: Because the surface is an equipotential, it follows that

$$\Delta V_{BA}=V_B-V_A=\int_A^B \vec{E}_{in}\cdot d\vec{l}=0.$$

The only way the integral can always be zero is if $\vec{E}_{in}\cdot d\vec{l}=|\vec{E}||d\vec{l}|\cos\theta=0$. However, this relationship can only hold in general if $\theta=0$. In other words, the electric field must be everywhere perpendicular to the displacements along the surface of the conductor (Fig. 4.4.1. (Alternatively, if the electric field were not perpendicular, then there would be force on the charge on the surface that would cause it to accelerate, thereby violating the assumption of electrostatic equilibrium.)

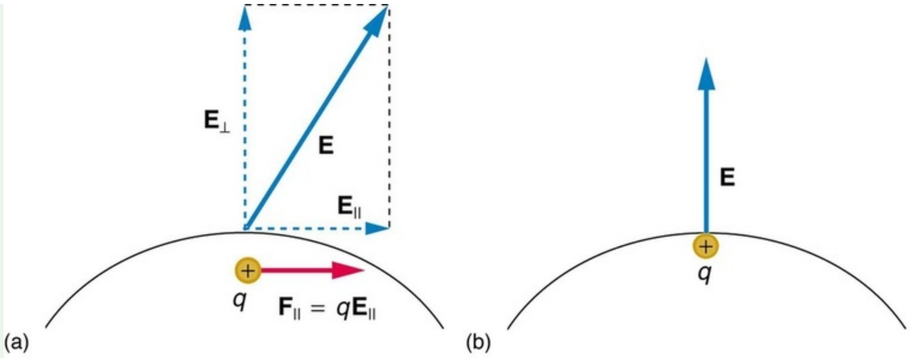


Figure 4.4.1: When an electric field $\vec{E} = \mathbf{E}$ is applied to a conductor, free charges inside the conductor move until the field is perpendicular to the surface. (a) The electric field is a vector quantity, with both parallel and perpendicular components. The parallel component ($\vec{E}_{||} = \mathbf{E}_{||}$) exerts a force ($\vec{F}_{||} = \mathbf{F}_{||}$) on the free charge q, which moves the charge until $\vec{F} = \mathbf{F} = 0$. (b) The resulting field is perpendicular to the surface. The free charge has been brought to the conductor's surface, leaving electrostatic forces in equilibrium. [Image from OpenStax College Physics Figure 18.26].

Property 4.4.4: Excess Charge on Conductors

Any free excess charge resides entirely on the surface or surfaces of a conductor.

Plausibility Argument: Because all the excess charges are free to move in the conductor, they will repel each other as much as possible. Having the charge on the surface will achieve this result. (A mathematical proof is provided in the section Conductors in Electrostatic Equilibrium via Gauss's Law.)

Remarkably all of these properties will hold true simultaneously for any conductor regardless of its shape. As an example, Fig. 4.4.2 shows the result of placing a neutral conductor in an originally uniform electric field. The field becomes stronger near the conductor but entirely disappears inside it.

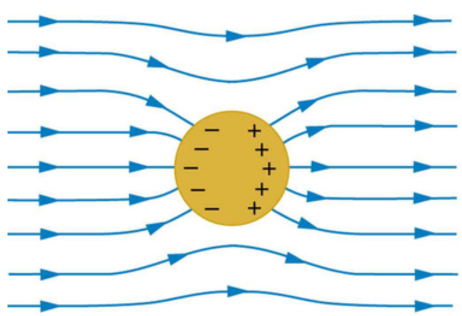


Figure 4.4.2: This illustration shows a spherical conductor in static equilibrium with an originally uniform electric field. Free charges move within the conductor, polarizing it, until the electric field lines are perpendicular to the surface. The field lines end on excess negative charge on one section of the surface and begin again on excess positive charge on the opposite side. No electric field exists inside the conductor, since free charges in the conductor would continue moving in response to any field until it was neutralized.

Charged Parallel Plates

One of the most important cases of conductors in electrostatic equilibrium is the familiar parallel conducting plates shown in Figure 4.4.3. Between the plates, the equipotentials are evenly spaced and parallel. The same field could be maintained by placing conducting plates at the equipotential lines at the potentials shown.

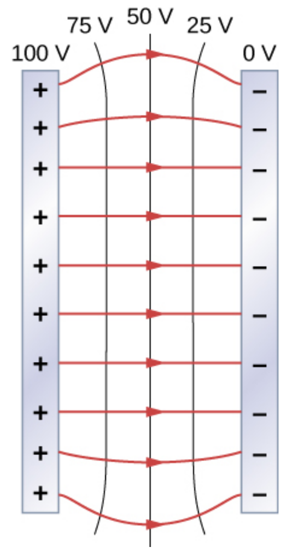


Figure 4.4.3: The electric field and equipotential lines between two metal plates. Note that the electric field is perpendicular to the equipotentials and hence normal to the plates at their surface as well as in the center of the region between them.

Based on what we know about electric potenial and electric field, we can also solve some quantitative problems about equipotentials, as shown in Examples 4.4.1 and 4.4.2.

Example 4.4.1: Calculating Equipotential Lines

You have seen the equipotential lines of a point charge in Figure 4.4.1. How do we calculate them? For example, if we have a $+10\text{-nC}$ charge at the origin, what are the equipotential surfaces at which the potential is (a) 100 V, (b) 50 V, (c) 20 V, and (d) 10 V?

Strategy

Set the equation for the potential of a point charge equal to a constant and solve for the remaining variable(s). Then calculate values as needed.

Solution

In $V = k\frac{q}{r}$, let V be a constant. The only remaining variable is r; hence, $r = k\frac{q}{V} = \text{constant}$. Thus, the equipotential surfaces are spheres about the origin. Their locations are:

a. $r = k\frac{q}{V} = \left(8.99 \times 10^9\,\mathrm{N \cdot m^2/C^2}\right)\frac{(10 \times 10^{-9}\,\mathrm{C})}{100\,\mathrm{V}} = 0.90\,\mathrm{m}$;
b. $r = k\frac{q}{V} = \left(8.99 \times 10^9\,\mathrm{N \cdot m^2/C^2}\right)\frac{(10 \times 10^{-9}\,\mathrm{C})}{50\,\mathrm{V}} = 1.8\,\mathrm{m}$;
c. $r = k\frac{q}{V} = \left(8.99 \times 10^9\,\mathrm{N \cdot m^2/C^2}\right)\frac{(10 \times 10^{-9}\,\mathrm{C})}{20\,\mathrm{V}} = 4.5\,\mathrm{m}$;
d. $r = k\frac{q}{V} = \left(8.99 \times 10^9\,\mathrm{N \cdot m^2/C^2}\right)\frac{(10 \times 10^{-9}\,\mathrm{C})}{10\,\mathrm{V}} = 9.0\,\mathrm{m}$.

Significance

This means that equipotential surfaces around a point charge are spheres of constant radius, as shown earlier, with well-defined locations.

Example 4.4.2:Potential Difference betwwen Oppositely Charged Parallel Plates

Two large conducting plates carry equal and opposite charges, with a surface charge density σ of magnitude $6.81 \times 10^{-7}\,\mathrm{C/m}$, as shown in Figure 4.4.4. The separation between the plates is $l = 6.50\,\mathrm{mm}$.

a. What is the electric field between the plates?

b. What is the potential difference between the plates?
c. What is the distance between equipotential planes which differ by 100 V?

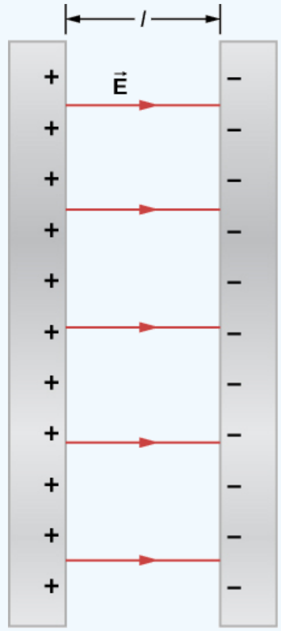


Figure 4.4.4: The electric field between oppositely charged parallel plates. A portion is released at the positive plate.

Strategy

1. Since the plates are described as “large” and the distance between them is not, we will approximate each of them as an infinite plane.
2. Use $\Delta V_{AB} = -\int_A^B \vec{E} \cdot d\vec{l}$.
3. Since the electric field is constant, find the ratio of 100 V to the total potential difference; then calculate this fraction of the distance.

Solution

a. The electric field is directed from the positive to the negative plate as shown in the figure, and its magnitude is given by (Common Models of Electric Potential)

$$\begin{aligned} E &= \frac{\sigma}{\epsilon_0} \\ &= \frac{6.81 \times 10^{-7}\ \mathrm{C/m^2}}{8.85 \times 10^{-12}\ \mathrm{C^2/N \cdot m^2}} \\ &= 7.69 \times 10^4\ \mathrm{V/m}. \end{aligned}$$

b. To find the potential difference ΔV between the plates, we use a path from the negative to the positive plate that is directed against the field. The displacement vector $d\vec{l}$ and the electric field $\vec{E}$ are antiparallel so $\vec{E} \cdot d\vec{l} = -E\,dl$. The potential difference between the positive plate and the negative plate is then

$$\begin{aligned} \Delta V &= -\int E \cdot dl \\ &= E \int dl \\ &= El \\ &= (7.69 \times 10^4\ \mathrm{V/m})(6.50 \times 10^{-3}\ \mathrm{m}) \\ &= 500\ \mathrm{V} \end{aligned}$$

c. The total potential difference is 500 V, so 1/5 of the distance between the plates will be the distance between 100-V potential differences. The distance between the plates is 6.5 mm, so there will be 1.3 mm between 100-V potential differences.

Significance

You have now seen a numerical calculation of the locations of equipotentials between two charged parallel plates.

Exercise 4.4.1

What are the equipotential surfaces for an infinite line charge?

Answer

infinite cylinders of constant radius, with the line charge as the axis

Falstad Simulation: 3D Electrostatic Field Simulation

Explore the 3-D Electrostatic Field Simulation (Java applet from Falstad.com). Start your investigation with a pair of parallel plates of equal but opposite charge.

Instructions:

1. Set the field selection to "charged plate pair."
2. Set the slice to "Show Y Slice."
3. Set the display to "Field Vectors." Describe the electric field-vector diagram. (Is it is similar to the diagram in Fig. 4.4.6?)
4. Set the display to "Field Lines." Is the field-line diagram consistent with the field-vector diagram?
5. Set the display to "Equipotentials." Are the equipotentials consistent with the field vectors?
6. Vary the sheet size and sheet separation. Can you create a uniform field between the plates?

Next, explore other geometries we have discussed (finite line, charged line, charged ring, conducting plate) and investigate their electric field and equipotential diagrams.

Distribution of Charges on Curved Conductors

In Example $4.4.1$ with a point charge, we found that the equipotential surfaces were in the form of spheres, with the point charge at the center. Given that a conducting sphere in electrostatic equilibrium is a spherical equipotential surface, we should expect that we could replace one of the surfaces in Example $4.4.2$ with a conducting sphere and have an identical solution outside the sphere. We also know that the electric field must be zero inside the conductor. The electric field of the sphere may therefore be written as

$$\begin{aligned} E &= 0 && (r<R), \\ E &= \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r} && (r\geq R). \end{aligned} \tag{4.4.1}$$

To find the electric potential inside and outside the sphere, note that for $r \geq R$, the potential must then be the same as that of an isolated point charge q located at $r=0$,

$$V(r)=\frac{1}{4\pi\epsilon_0}\frac{q}{r}, (r\geq R) \tag{4.4.2}$$

simply due to the similarity of the electric field.

For $r<R,\ E=0$, so $V(r)$ is constant in this region. Since $V(R)=q/4\pi\epsilon_0 R$,

$$V(r)=\frac{1}{4\pi\epsilon_0}\frac{q}{R}(r<R). \tag{4.4.3}$$

With this background, we can now state an additional property of a conductor at electrostatic equilibrium.

Property 4.4.5: Surface Charge Density on Conductors

The surface charge density is inversely proportional to the local radius of curvature on the surface of a conductor.

Proof: Our goal will be to show that

$$\frac{\sigma_1}{\sigma_2} = \frac{R_2}{R_1},$$

for two conducting spheres of radii R_1 and R_2, with surface charge densities σ_1 and σ_2 respectively, that are connected by a thin wire, as shown in Figure 4.4.5. The spheres are sufficiently separated so that each can be treated as if it were isolated (aside from the wire). Note that the connection by the wire means that this entire system must be an equipotential.

Figure 4.4.5: Two conducting spheres are connected by a thin conducting wire.

We have just seen that the electrical potential at the surface of an isolated, charged conducting sphere of radius R is

$$V = \frac{1}{4\pi\epsilon_0}\frac{q}{R}.$$

Now, the spheres are connected by a conductor and are therefore at the same potential; hence

$$\frac{1}{4\pi\epsilon_0}\frac{q_1}{R_1} = \frac{1}{4\pi r\epsilon_0}\frac{q_2}{R_2},$$

and

$$\frac{q_1}{R_1} = \frac{q_2}{R_2}.$$

The net charge on a conducting sphere and its surface charge density are related by $q = \sigma(4\pi R^2)$. Substituting this equation into the previous one, we find

$$\sigma_1 R_1 = \sigma_2 R_2.$$

from which it follows that

$$\frac{\sigma_1}{\sigma_2} = \frac{R_2}{R_1},$$

showing that the surface charge density is inversely proportional to the radius of curvature of the sphere.

Obviously, two spheres connected by a thin wire do not constitute a typical conductor with a variable radius of curvature, but this result still holds true even in the more general case. The equation indicates that where the radius of curvature is large (points B and D in 4.4.10), σ and E are small. Similarly, the charges tend to be denser where the curvature of the surface is greater, as demonstrated by the charge distribution on oddly shaped metal (Figure 4.4.6). The surface charge density is higher at locations with a small radius of curvature than at locations with a large radius of curvature.

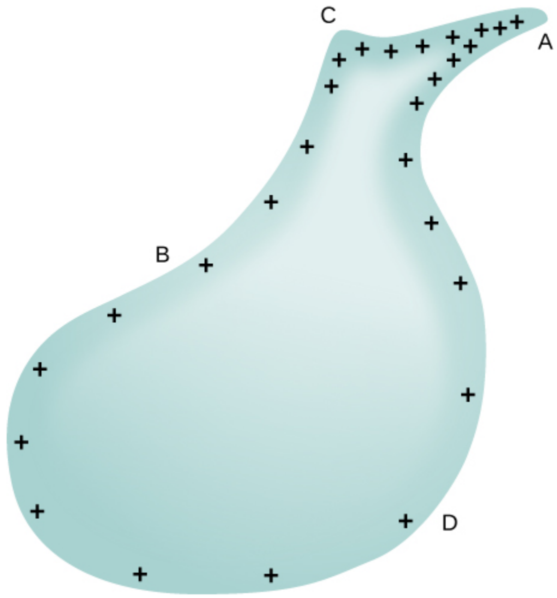


Figure 4.4.6: The surface charge density and the electric field of a conductor are greater at regions with smaller radii of curvature.

Electric Fields on Curved Surfaces

Having established that the surface charge density varies on the surface of a charged object based on the local curvature of the surface, one might expect that the electric field magnitude will also vary, and this is indeed true.

Property 4.4.6: Electric Field Magnitude on the Surface of a Conductor

The electric field magnitude on the curved surface of a conductor is inversely proportional to the radius curvature of the surface.

To see how and why this happens, consider the charged conductor in Fig. 4.4.7. Qualitatively, the electrostatic repulsion of like charges is most effective in moving them apart on the flattest surface, and so they become least concentrated there. This is because the forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surfaces are different. The component parallel to the surface is greatest on the flattest surface and, hence, more effective in moving the charge (Fig. 4.4.7*a*). The result is that the electric field is higher at sharper corners (Fig. 4.4.7*b*). In fact, it is possible to show that the local electric field $E = \sigma/\epsilon_0$ (see Conductors in Electrostatic Equilibrium Via Gauss's Law).

The same effect is produced on a conductor by an externally applied electric field (4.4.7*c*). Since the field lines must be perpendicular to the surface, more of them are concentrated on the most curved parts.

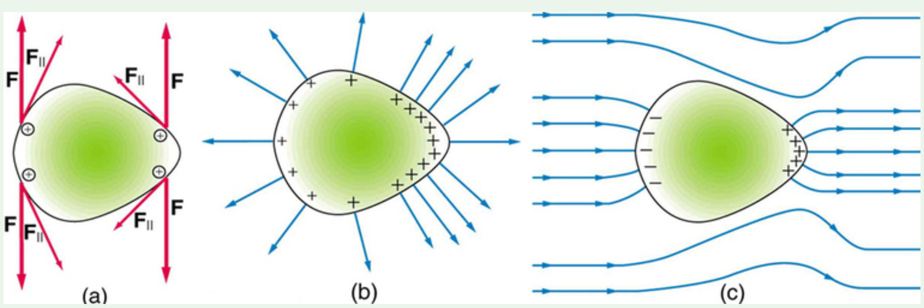


Figure 4.4.7: Excess charge on a nonuniform conductor becomes most concentrated at the location of greatest curvature. (a) The forces between identical pairs of charges at either end of the conductor are identical, but the components of the forces parallel to the surface are different. It is $\vec{F}_{||}$ that moves the charges apart once they have reached the surface. (b) $\vec{F}_{||}$ is smallest at the more pointed end, the charges are left closer together, producing the electric field shown. (c) An uncharged conductor in an originally uniform electric field is polarized, with the most concentrated charge at its most pointed end.

Cavities in Conductors

Suppose we have a conductor with a cavity inside it. If we place the conductor into an external electric field, we know that the conductor will become polarized, as in Fig. 4.4.2. Will the empty cavity become polarized also? The answer is no. Because there is no electric field inside the conductor and no charge in the cavity, there will also be no electric field in the cavity and, therefore, no charge on the surface of the cavity (Fig. 4.4.8b; see Conductors in Electrostatic Equilibrium Via Gauss's Law for additional justification for this claim).

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 4.4.8a). However, the electric field inside the conductor itself will remain zero.

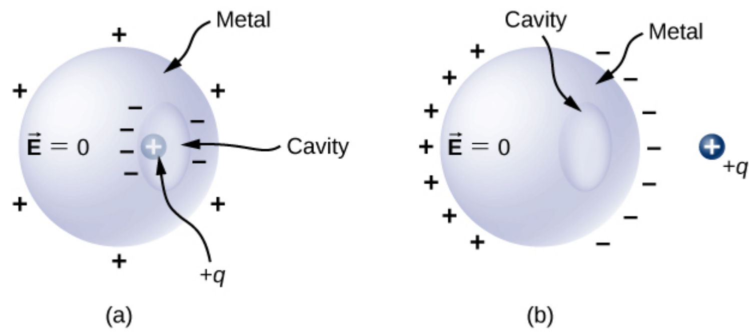


Figure 4.4.8: (a) A charge inside a cavity in a metal. The distribution of charges at the outer surface does not depend on how the charges are distributed at the inner surface, since the $\vec{E}$-field inside the body of the metal is zero. That magnitude of the charge on the outer surface does depend 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.

4.5: Applications of Electric Potential and Conductors in Electrostatic Equilibrium

Learning Objectives

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

- Define grounding and bonding.
- Identify some effects of electricity on the human body and their implications for electrical safety.
- Describe the electric field surrounding Earth.
- Explain how a lightning rod works.
- Define screening, and explain how a Faraday cage works.

Applications of Electric Potential and Field

There are numerous practical applications of electric potential and electric field in the practice of Amateur Radio and other wireless technology. We highlight a few important examples in this section.

Grounding and Bonding

With respect to electricity, a **ground** is considered a reference point in an electrical circuit from which voltages are measured [1]. The term "ground" is used because, it is common to make physical connections to the Earth, often by a metal stake placed in ground (Fig. 4.5.1), via a conductor. From Conductors in Electrostatic Equilibrium, we know that a conductor will have a constant potential. If multiple connections are made to ground via conductors, then they will all share the same potential. It is conventional to set the potential to zero volts for convenience. The process of connecting a circuit to ground is called **grounding**. **Bonding** is the process of connecting together multiple electrical devices with a conductor, usually a thick wire, to maintain a constant potential among the devices, and is often done before connecting the devices to ground.

Figure 4.5.1: Example of grounding. The electrical circuits from a building are connected by a conductor (green and yellow wire) into a metal stake into the ground. [3]

Proper grounding and bonding are important for electrical safety. For example, grounding the metal case of an electrical appliance ensures that it is at zero volts relative to Earth. This practice can minimize the chance of electric shock hazards when a fault occurs in the wiring. Grounding can also help to minimize potential differences between devices during sudden changes in voltage in the electrical distribution system. During lighting strikes, it can also help to dissipate charge, routing away from valuable equipment and into the Earth. While these processes sound simple enough, there can be subtleties when implementing them in practice. You should follow the grounding requirements in your local electrical codes, particularly for Amateur Radio towers and antennas. (See Ref [2] for a comprehensive discussion of best practices for grounding and bonding for Amateur Radio.)

Electrophysiology

Electrophysiology is a branch of physiology that studies the electrical properties of cells and tissue. The human body contains several systems that use electrical signals. For example, the heart relies on electrical signals to maintain its rhythm [4]. The movement of electrical signals causes the chambers of the heart to contract and relax. The equipotential lines around the heart and surrounding areas are useful ways of monitoring the structure and functions of the heart. An **electrocardiogram (ECG)** measures the small electric signals being generated during the activity of the heart. When a person has a heart attack, the movement of these electrical signals may be disturbed. When a person has irregular heartbeat, an artificial pacemaker and a defibrillator can be used to initiate the rhythm of electrical signals. Electrical signals also play an important role in the nervous system. Like cardiac tissue, nerves are also capable of transmitting electrochemical nerve impulses called **action potentials** (voltages) along the body of the nerve cells [5] (Fig. $4.5.2$).

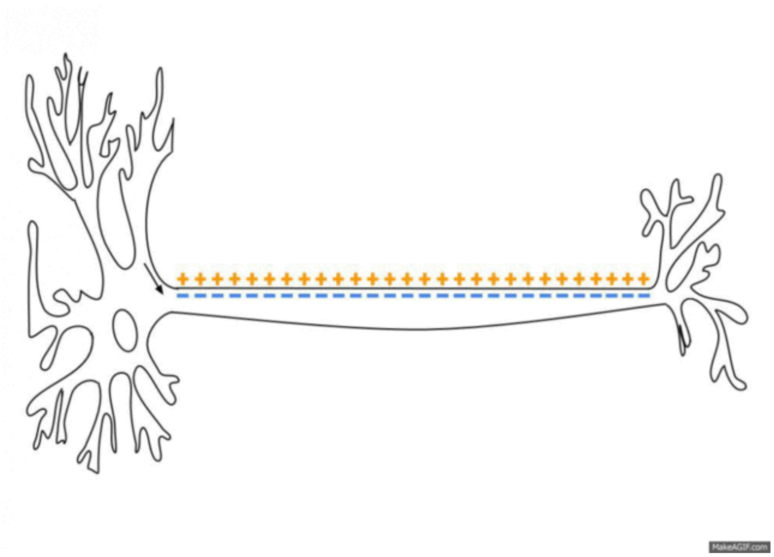


Figure $4.5.2$: Electrical impulses in a nerve. As a nerve impulse travels down the axon, there is a change in polarity across the membrane. The Na+ and K+ gated ion channels open and close in response to a signal from another neuron. At the beginning of action potential, the Na+ gates open and Na+ moves into the axon. This is depolarization. Repolarization occurs when the K+ gates open and K+ moves outside the axon. This creates a change in polarity between the outside and inside of the cell . The impulse continuously travels down the axon in one direction only, through the axon terminal and to other neurons. [6]

Human electrophysiology has implications for electrical safety. As we will see, electrical charges will move from regions of high electrical potential to regions of low potential (ground). When touching electrical circuits, you want to ensure you do not become the path to ground! One good practice is only to touch circuits at one point. (Keep your other hand in a pocket!) In particular, avoid allowing charge to move in a path across the heart, for example, from one hand to the other. Electric shocks can also cause nervous system responses, including involuntary muscle contractions. In the worst cases, these contractions are so strong that they may prevent you from releasing from the charged object, making it difficult to remove yourself from the hazard! As will be discussed later, electrical shocks may also cause heating or burning of tissue. Other common safety practices include removing metal jewelry, working in dry conditions, and being aware of overhead wires.

Earth's Electric Field

A near uniform electric field of approximately 100 to 150 N/C, directed downward, surrounds Earth, with the magnitude increasing slightly as we get closer to the surface. What causes the electric field? At around 100 km above the surface of Earth, we have a layer of charged particles, called the **ionosphere**. The ionosphere is predominantly caused by ultraviolet radiation from the sun. The ionosphere is responsible for a range of interesting phenomena, including the electric field surrounding Earth and the bending of radio waves. In fair weather, the ionosphere is positive and the Earth largely negative, maintaining the electric field (Figure $4.5.2a$) [7] .

In storm conditions, clouds form, and localized electric fields can be larger and even reversed in direction (Figure $4.5.2b$). The exact charge distributions depend on the local conditions, and variations of Figure $4.5.2b$ are possible. If the electric field is sufficiently large, the insulating properties of the surrounding material experience **electrical breakdown**, and it becomes conducting. For air, this occurs at around 3×10^6 N/C. Air ionizes ions, electrons recombine, and we can get effects like **lightning** [8, 9] and **corona discharge** [10].

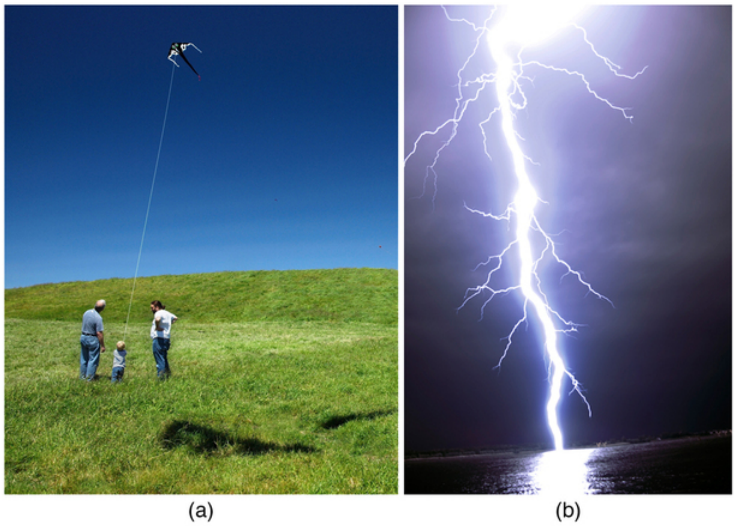


Figure 4.5.2: Earth's electric field. (a) Fair weather field. Earth and the ionosphere (a layer of charged particles) are both conductors. They produce a uniform electric field of about 100 to 150 N/C. (credit: D. H. Parks) (b) Storm fields. In the presence of storm clouds, the local electric fields can be larger. At very high fields, the insulating properties of the air break down, and lightning can occur. (credit: Jan-Joost Verhoef)

Applications of Conductors

Lightning Rods

As discussed in Conductors in Electrostatic Equilibrium, a charged conductor will have a higher electric field and surface charge density where it has higher curvature. A practical application of this phenomenon is the **lightning rod**, which is simply a grounded metal rod with a sharp end pointing upward (Fig. $4.5.3$, $4.5.4a$). As positive charge accumulates in the ground due to a negatively charged cloud overhead, the electric field around the sharp point gets very large. When the field reaches a value of approximately $3.0 \times 10^6\ \mathrm{N/C}$ (the **dielectric strength** of the air), the free ions in the air are accelerated to such high energies that their collisions with air molecules actually ionize the molecules. The resulting free electrons in the air then flow through the rod to Earth, thereby neutralizing some of the positive charge. This keeps the electric field between the cloud and the ground from getting large enough to produce a lightning bolt in the region around the rod.

Lightning protection is an important part of the safety considerations for Amateur Radio operators, who often have tall antennas or towers to improve their ability to transmit and receive radio signals. As the height of an antenna or tower increases, so does its probability of being struck. The risk level also depends on the prevalence of thunderstorms in the vicinity. This property of conductors must also be considered when performing the grounding of a lightning rod. The connections from the rod should be as short and direct as possible and should not connect sharp angles that could result in the electric charges arcing through the air prior to reaching ground.

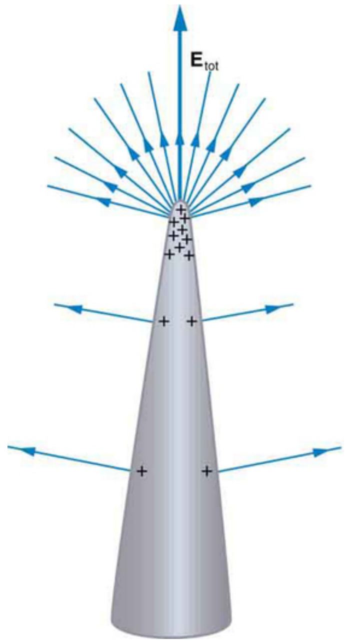


Figure 4.5.3: A very pointed conductor has a large charge concentration at the point. The electric field is very strong at the point and can exert a force large enough to transfer charge on or off the conductor. Lightning rods are used to prevent the buildup of large excess charges on structures and, thus, are pointed.

Lightning Mapping

It is possible to create maps of lightning activity in near real-time. Each time a lightning strike occurs, very low frequency (VLF) radio waves (3 to 30 kHz) are emitted. By detecting these signals at multiple receiving stations at known locations, it is possible to triangulate the location of the lightning strike [11]. See the maps at Blitzortung.org [12] for examples in your region.

Of course, we sometimes wish to prevent the transfer of charge rather than to facilitate it. In that case, the conductor should be very smooth and have as large a radius of curvature as possible (Figure 4.5.4*b*). For example, smooth surfaces are used on high-voltage transmission lines to avoid leakage of charge into the air.

Figure 4.5.4: (a) A lightning rod is pointed to facilitate the transfer of charge. (credit: Romaine, Wikimedia Commons) (b) This Van de Graaff generator has a smooth surface with a large radius of curvature to prevent the transfer of charge and allow a large voltage to be generated. The mutual repulsion of like charges is evident in the person's hair while touching the metal sphere. (credit: Jon 'ShakataGaNai' Davis/Wikimedia Commons).

Screening

We also learned in Conductors in Electrostatic Equilibrium that no electric field exists inside a cavity in a conductor. This effect can be used practically to perform **screening** or **shielding** of objects from electric fields. For example, a **Faraday cage** is a metal shield that encloses a volume (Fig. 4.5.5) [13]. The Faraday cage is commonly used to prohibit stray electrical fields in the environment from interfering with sensitive measurements, such as the electrical signals inside a nerve cell. Perhaps the most common example of a (partial) Faraday cage is a typical microwave oven, which will have 5 complete metal sides and one metal screen on the oven's door to contain the microwaves in the oven and prevent exposure to nearby users.

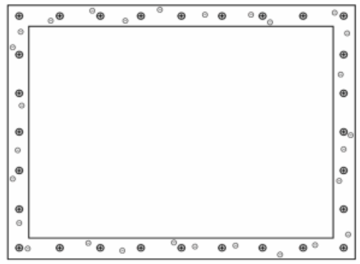


Figure 4.5.5: Illustration of the screening effect of the Faraday cage. In the presence of an external electric field, the electrical charges will reside on the outside surface of this shield and arrange themselves such that is no electrical field inside. [14]

By similar reasoning, the body of your car can serve as a partial Faraday cage. As a result, if you are driving a car during an electrical storm, it is best to stay inside the car. Even if the car is in the vicinity of a lightning strike, the effect of the strike will be felt on the outside of the car. You will be unaffected, provided you remain totally inside and not in contact with the body of vehicle. This protective effect is also true if an active ("hot") electrical wire breaks in a storm or an accident and falls on the car.

Amateur radio operators will use screening to try to protect their radio equipment from interference from surrounding devices. Some cables, like the coaxial cables used in radio and cable television, also use screening to minimize the effects of surrounding electrical noise. The American Radio Relay League headquarters contains a Faraday cage used to test radio equipment without outside interference. However, the screening effects can also be detrimental to radio reception. For example, it can sometimes be difficult to get good reception inside a vehicle using a handheld VHF transceiver with a directly-mounted antenna (see photo in Amateur Radio Equipment Basics). Instead, the antenna should be placed on the exterior of the vehicle and then connected to the radio via a transmission cable.

References

1. Wikipedia contributors. Ground (electricity) [Internet]. Wikipedia, The Free Encyclopedia.
2. Silver, HW (N0AX). Grounding and bonding for the radio amateur: Good practices for electrical safety, lightning protection, and RF management, 2nd edition. Newington, CT: American Radio Relay League; 2021.
3. Wikimedia Commons contributors. File:HomeEarthRodAustralia1.jpg [Internet]. Wikimedia Commons. (Ali K., CC-BY-SA 3.0)
4. Wikipedia contributors. Cardiac conduction system [Internet]. Wikipedia, The Free Encyclopedia.
5. Wikipedia contributors. Nerve [Internet]. Wikipedia, The Free Encyclopedia.
6. Wikimedia Commons contributors. File:Action Potential.gif [Internet]. Wikimedia Commons. (Laurentaylorj, CC-BY-SA 3.0)
7. Wikipedia contributors. Atmospheric electricity [Internet]. Wikipedia, The Free Encyclopedia.
8. Wikipedia contributors. Lightning [Internet]. Wikipedia, The Free Encyclopedia.
9. JetStream - An Online School for Weather. Lightning: introduction to lightning. [Internet] National Oceanic and Atmospheric Administration.
10. Wikipedia contributors. Corona discharge [Internet]. Wikipedia, The Free Encyclopedia.
11. Lightningmaps.org. Lightning detection [Internet]. Lightningmaps.org.
12. Blitzortung.org. Overview map [Internet]. Blitzortung.org.
13. Wikipedia contributors. Faraday cage [Internet]. Wikipedia, The Free Encyclopedia.
14. Wikimedia Commons contributors. File:Faraday cage.gif [Internet]. Wikimedia Commons. (Stanisław Skowron, public domain)

Technician Exam Questions

Relevant exam questions include: T9A07, T0A02, T0A09, T0B01, T0B10, T0B11

- **7.6: Equipotential Surfaces and Conductors** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.
- **18.7: Conductors and Electric Fields in Static Equilibrium** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/college-physics.

4.6: Potential and Field Relationships (Summary)

Key Terms

bonding	connecting multiple electrical devices together with a conductor to make them have a common electric potential
electric breakdown	the condition when a sufficient electric field exists to enable an insulator to conduct charge
electric potential	potential energy per unit charge
electric potential difference	the change in potential energy of a charge q moved between two points, divided by the charge.
electrophysiology	branch of physiology that studies the electrical properties of cells and tissue
equipotential curve	two-dimensional representation of an equipotential surface
equipotential surface	surface (usually in three dimensions) on which all points are at the same potential
Faraday cage	an enclosure used for screening electric field from a region of space
grounding	process of attaching a conductor to the earth to ensure that there is no potential difference between it and Earth
lightning rod	a sharp-tipped conductor used to draw charge from the air and minimize the probability of a lightning strike
screening	the use of a cavity in a conductor to minimize or eliminate the electric field in the cavity

Key Equations

Potential difference between two points in terms of the integral of the dot product of the electric field with displacement along a path	$\Delta V_{AB} = V_B - V_A = -\int_A^B \vec{E} \cdot \vec{dl}$
Potential difference in a uniform electric field over a distance	$V_{AB} = Ed$
Electric field as gradient of potential	$\vec{E} = -\vec{\nabla} V$
Electric field components in Cartesian coordinates	$E_x = -\frac{\partial V}{\partial x}, E_y = -\frac{\partial V}{\partial y}, E_z = -\frac{\partial V}{\partial z}$
Del operator in Cartesian coordinates	$\vec{\nabla} = \hat{i}\frac{\partial}{\partial x} + \hat{j}\frac{\partial}{\partial y} + \hat{k}\frac{\partial}{\partial z}$
Electric field components in Cartesian coordinates	$E_x = -\frac{\partial V}{\partial x}, E_y = -\frac{\partial V}{\partial y}, E_z = -\frac{\partial V}{\partial z}$
Del operator in cylindrical coordinates	$\vec{\nabla} = \hat{r}\frac{\partial}{\partial r} + \hat{\varphi}\frac{1}{r}\frac{\partial}{\partial \varphi} + \hat{z}\frac{\partial}{\partial z}$
Del operator in spherical coordinates	$\vec{\nabla} = \hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\varphi}\frac{1}{r\sin\theta}\frac{\partial}{\partial \varphi}$

Summary

Electric Potential from Electric Field

- The potential difference between points can be computed from the negative line integral of the electric field over a path between the two points.

Electric Field from Electric Potential

- Just as we may calculate the negative integral over the electric field to calculate the potential difference, we may take the negative derivative of the potential to calculate the electric field.
- This may be done for individual components of the electric field, or we may calculate the entire electric field vector with the gradient operator.

Equipotential Curves and Surfaces

- An equipotential surface is the collection of points in space that are all at the same potential. Equipotential lines are the two-dimensional representation of equipotential surfaces.
- Equipotential surfaces are always perpendicular to electric field lines.

Conductors in Electrostatic Equilibrium

- Conductors in static equilibrium are equipotential surfaces.
- The electric field is zero inside a conductor and, if charged, everywhere perpendicular to its surface.
- The electric potential is constant throughout a conductor.
- The electric field magnitude and surface charge density is inversely proportional to the radius of curvature of the surface.

Applications of Electric Potential and Conductors in Electrostatic Equilibrium

- Grounding and bonding use conductors to maintain a constant potential throughout a set of connected devices.
- The Earth has a natural electric field that is directed toward the earth's surface.
- The human body has cells and tissues that use changes in electric potential to achieve their function. Muscles and nerves are examples.
- Lightning rods can draw current from the air and minimize the probability of lightning strikes.
- Screening of electric fields can be accomplished through the use of enclosed conductors (Faraday cage).

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).

4.7: Potential and Field Relationships (Exercises)

Conceptual Questions

Electric Field from Electric Potential

17. If the electric field is zero throughout a region, must the electric potential also be zero in that region?

18. Explain why knowledge of $\vec{E}(x,y,z)$ is not sufficient to determine $V(x,y,z)$. What about the other way around?

Equipotential Curves and Surfaces

19. If two points are at the same potential, are there any electric field lines connecting them?

20. Suppose you have a map of equipotential surfaces spaced 1.0 V apart. What do the distances between the surfaces in a particular region tell you about the strength of the $\vec{E}$ in that region?

Conductors in Electrostatic Equilibrium

21. Is the electric potential necessarily constant over the surface of a conductor?

22. Under electrostatic conditions, the excess charge on a conductor resides on its surface. Does this mean that all of the conduction electrons in a conductor are on the surface?

23. Can a positively charged conductor be at a negative potential? Explain.

24. Can equipotential surfaces intersect?

Applications of Electric Potential and Conductors in Electrostatic Equilibrium

25. Why are the metal support rods for satellite network dishes generally grounded?

Problems

Electric Field from Electric Potential

55. Throughout a region, equipotential surfaces are given by $z = constant$. The surfaces are equally spaced with $V = 100V$ for $z = 0.00m, V = 200V$ for $z = 0.50m, V = 300V$ for $z = 1.00m$. What is the electric field in this region?

56. In a particular region, the electric potential is given by $V = -xy^2z + 4xy$. What is the electric field in this region?

57. Calculate the electric field of an infinite line charge, throughout space.

Equipotential Curves and Surfaces

Conductors in Electrostatic Equilibrium

58. Two very large metal plates are placed 2.0 cm apart, with a potential difference of 12 V between them. Consider one plate to be at 12 V, and the other at 0 V. (a) Sketch the equipotential surfaces for 0, 4, 8, and 12 V.

(b) Next sketch in some electric field lines, and confirm that they are perpendicular to the equipotential lines.

59. A very large sheet of insulating material has had an excess of electrons placed on it to a surface charge density of $-3.00nC/m^2$.

(a) As the distance from the sheet increases, does the potential increase or decrease? Can you explain why without any calculations? Does the location of your reference point matter?

(b) What is the shape of the equipotential surfaces?

(c) What is the spacing between surfaces that differ by 1.00 V?

60. A metallic sphere of radius 2.0 cm is charged with $+5.0-\mu C$ charge, which spreads on the surface of the sphere uniformly. The metallic sphere stands on an insulated stand and is surrounded by a larger metallic spherical shell, of inner radius 5.0 cm and outer radius 6.0 cm. Now, a charge of $-5.0-\mu C$ is placed on the inside of the spherical shell, which spreads out uniformly on the inside surface of the shell. If potential is zero at infinity, what is the potential of

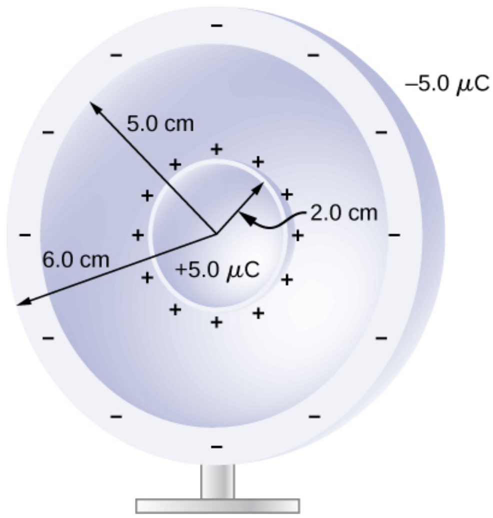
(a) the spherical shell,

(b) the sphere,

(c) the space between the two,

(d) inside the sphere, and

(e) outside the shell?



61. Two large charged plates of charge density $\pm 30\mu C/m^2$ face each other at a separation of 5.0 mm.

(a) Find the electric potential everywhere.

(b) An electron is released from rest at the negative plate; with what speed will it strike the positive plate?

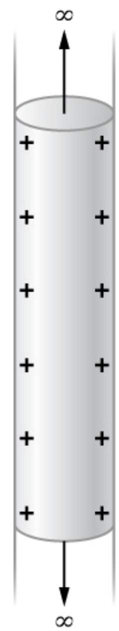
62. A long cylinder of aluminum of radius R meters is charged so that it has a uniform charge per unit length on its surface of λ.

(a) Find the electric field inside and outside the cylinder.

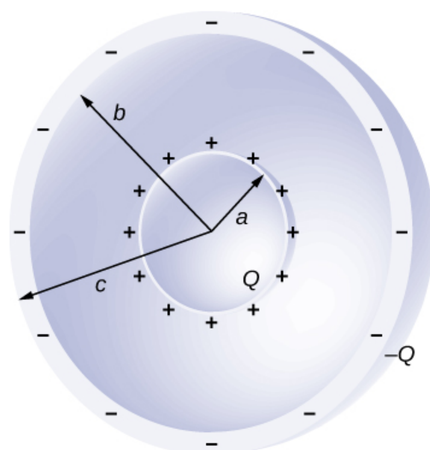
(b) Find the electric potential inside and outside the cylinder. (c) Plot electric field and electric potential as a function of distance from the center of the rod.

63. Two parallel plates 10 cm on a side are given equal and opposite charges of magnitude $5.0 \times 10^{-9} C$.The plates are 1.5 mm apart. What is the potential difference between the plates?

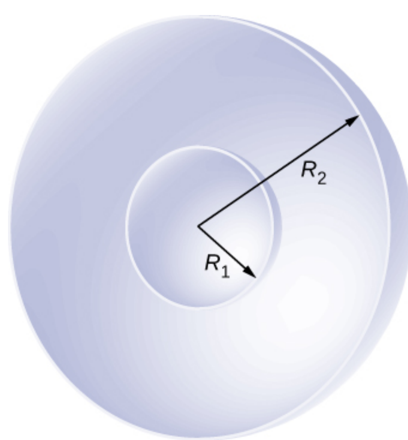
64. The surface charge density on a long straight metallic pipe is σ. What is the electric potential outside and inside the pipe? Assume the pipe has a diameter of **2a**.



65. Concentric conducting spherical shells carry charges **Q** and **–Q**, respectively. The inner shell has negligible thickness. What is the potential difference between the shells?



66. Shown below are two concentric spherical shells of negligible thicknesses and radii R_1 and R_2. 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 potential in the regions (a) $r < R_1$, (b) $R_1 < r < R_2$, and (c) $r > R_2$?



67. 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**, what is the potential difference between the two conductors?

Applications of Electric Potential and Conductors in Electrostatic Equilibrium

68. (a) What is the electric field 5.00 m from the center of the terminal of a Van de Graaff with a 3.00-mC charge, noting that the field is equivalent to that of a point charge at the center of the terminal?

(b) At this distance, what force does the field exert on a $2.00-\mu C$ charge on the Van de Graaff's belt?

69. (a) What is the direction and magnitude of an electric field that supports the weight of a free electron near the surface of Earth?

(b) Discuss what the small value for this field implies regarding the relative strength of the gravitational and electrostatic forces.

70. A simple and common technique for accelerating electrons is shown in Figure $4.7.1$, where there is a uniform electric field between two plates. Electrons are released, usually from a hot filament, near the negative plate, and there is a small hole in the positive plate that allows the electrons to continue moving.

(a) Calculate the acceleration of the electron if the field strength is $2.50 \times 10^4 N/C$.

(b) Explain why the electron will not be pulled back to the positive plate once it moves through the hole.

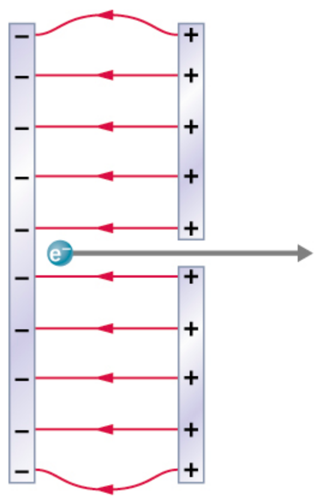


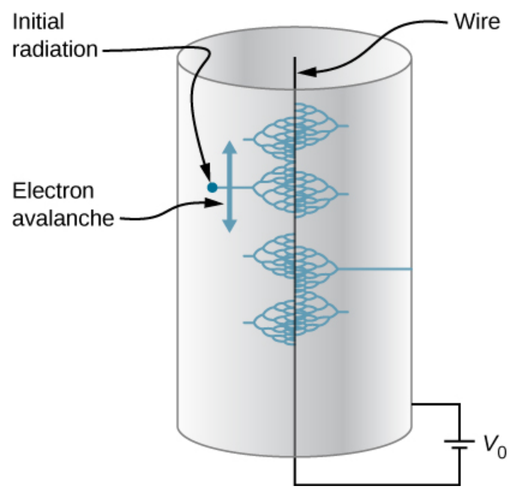
Figure $4.7.1$: Parallel conducting plates with opposite charges on them create a relatively uniform electric field used to accelerate electrons to the right. Those that go through the hole can be used to make a TV or computer screen glow or to produce X-rays.

71. In a Geiger counter, a thin metallic wire at the center of a metallic tube is kept at a high voltage with respect to the metal tube. Ionizing radiation entering the tube knocks electrons off gas molecules or sides of the tube that then accelerate towards the center wire, knocking off even more electrons. This process eventually leads to an avalanche that is detectable as a current. A particular Geiger counter has a tube of radius **R** and the inner wire of radius a is at a potential of V_0 volts with respect to the outer metal tube. Consider a point P at a distance s from the center wire and far away from the ends.

(a) Find a formula for the electric field at a point **P** inside using the infinite wire approximation.

(b) Find a formula for the electric potential at a point **P** inside.

(c) Use $V_0 = 900V, a = 3.00mm, R = 2.00cm$, and find the value of the electric field at a point 1.00 cm from the center.



72. The practical limit to an electric field in air is about $3.00 \times 10^6 N/C$. Above this strength, sparking takes place because air begins to ionize.

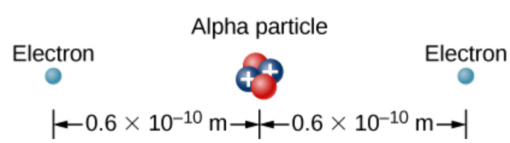
(a) At this electric field strength, how far would a proton travel before hitting the speed of light (ignore relativistic effects)?

(b) Is it practical to leave air in particle accelerators?

73. To form a helium atom, an alpha particle that contains two protons and two neutrons is fixed at one location, and two electrons are brought in from far away, one at a time. The first electron is placed at $0.600 \times 10^{-10} m$ from the alpha particle and held there while the second electron is brought to $0.600 \times 10^{-10} m$ from the alpha particle on the other side from the first electron. See the final configuration below.

(a) How much work is done in each step?

(b) What is the electrostatic energy of the alpha particle and two electrons in the final configuration?



74. Find the electrostatic energy of eight equal charges $(+3 \mu C)$ each fixed at the corners of a cube of side 2 cm.

75. The probability of fusion occurring is greatly enhanced when appropriate nuclei are brought close together, but mutual Coulomb repulsion must be overcome. This can be done using the kinetic energy of high-temperature gas ions or by accelerating the nuclei toward one another.

(a) Calculate the potential energy of two singly charged nuclei separated by $1.00 \times 10^{-12} m$.

(b) At what temperature will atoms of a gas have an average kinetic energy equal to this needed electrical potential energy?

76. A bare helium nucleus has two positive charges and a mass of $6.64 \times 10^{-27} kg$.

(a) Calculate its kinetic energy in joules at 2.00 of the speed of light.

(b) What is this in electron-volts?

(c) What voltage would be needed to obtain this energy?

77. An electron enters a region between two large parallel plates made of aluminum separated by a distance of 2.0 cm and kept at a potential difference of 200 V. The electron enters through a small hole in the negative plate and moves toward the positive plate. At the time the electron is near the negative plate, its speed is $4.0 \times 10^5 m/s$. Assume the electric field between the plates to be uniform, and find the speed of electron at

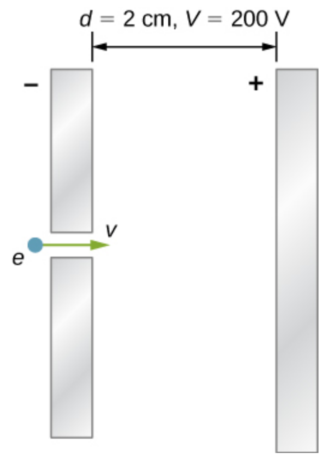
(a) 0.10 cm,

(b) 0.50 cm,

(c) 1.0 cm, and

(d) 1.5 cm from the negative plate, and

(e) immediately before it hits the positive plate.



78. How far apart are two conducting plates that have an electric field strength of $4.50 \times 10^3 V/m$ between them, if their potential difference is 15.0 kV?

79. (a) Will the electric field strength between two parallel conducting plates exceed the breakdown strength of dry air, which is $3.00 \times 10^6 V/m$, if the plates are separated by 2.00 mm and a potential difference of $5.0 \times 10^3 V$ is applied?

(b) How close together can the plates be with this applied voltage?

80. Membrane walls of living cells have surprisingly large electric fields across them due to separation of ions. What is the voltage across an 8.00-nm-thick membrane if the electric field strength across it is 5.50 MV/m? You may assume a uniform electric field.

81. A double charged ion is accelerated to an energy of 32.0 keV by the electric field between two parallel conducting plates separated by 2.00 cm. What is the electric field strength between the plates?

82. The temperature near the center of the Sun is thought to be 15 million degrees Celsius ($1.5 \times 10^{7\circ} C$) (or kelvin). Through what voltage must a singly charged ion be accelerated to have the same energy as the average kinetic energy of ions at this temperature?

83. A lightning bolt strikes a tree, moving 20.0 C of charge through a potential difference of $1.00 \times 1^0 2 MV$.

(a) What energy was dissipated?

(b) What mass of water could be raised from **15°C** to the boiling point and then boiled by this energy?

(c) Discuss the damage that could be caused to the tree by the expansion of the boiling steam.

84. What is the potential $0.530 \times 10^{-10} m$ from a proton (the average distance between the proton and electron in a hydrogen atom)?

85. (a) A sphere has a surface uniformly charged with 1.00 C. At what distance from its center is the potential 5.00 MV? (b) What does your answer imply about the practical aspect of isolating such a large charge?

86. What are the sign and magnitude of a point charge that produces a potential of –2.00 V at a distance of 1.00 mm?

87. In one of the classic nuclear physics experiments at the beginning of the twentieth century, an alpha particle was accelerated toward a gold nucleus, and its path was substantially deflected by the Coulomb interaction. If the energy of the doubly charged alpha nucleus was 5.00 MeV, how close to the gold nucleus (79 protons) could it come before being deflected?

Additional Problems

88. A 12.0-V battery-operated bottle warmer heats 50.0 g of glass, $2.50 \times 10^2 g$ of baby formula, and $2.00 \times 10^2 g$ of aluminum from **20.0°C** to **90.0°C**.

(a) How much charge is moved by the battery?

(b) How many electrons per second flow if it takes 5.00 min to warm the formula? (Hint: Assume that the specific heat of baby formula is about the same as the specific heat of water.)

89. A battery-operated car uses a 12.0-V system. Find the charge the batteries must be able to move in order to accelerate the 750 kg car from rest to 25.0 m/s, make it climb a $2.00 \times 10^2 - m$ high hill, and finally cause it to travel at a constant 25.0 m/s while climbing with $5.00 \times 10^2 - N$ force for an hour.

90. (a) Find the voltage near a 10.0 cm diameter metal sphere that has 8.00 C of excess positive charge on it.

(b) What is unreasonable about this result?

(c) Which assumptions are responsible?

91. A uniformly charged half-ring of radius 10 cm is placed on a nonconducting table. It is found that 3.0 cm above the center of the half-ring the potential is –3.0 V with respect to zero potential at infinity. How much charge is in the half-ring?

92. A glass ring of radius 5.0 cm is painted with a charged paint such that the charge density around the ring varies continuously given by the following function of the polar angle $\theta, \lambda = (3.0 \times 10^{-6} C/m) cos^2\theta.$. Find the potential at a point 15 cm above the center.

93. A CD disk of radius $(R = 3.0cm)$ is sprayed with a charged paint so that the charge varies continually with radial distance rfrom the center in the following manner: $\sigma = -(6.0C/m)r/R$ Find the potential at a point 4 cm above the center.

94. (a) What is the final speed of an electron accelerated from rest through a voltage of 25.0 MV by a negatively charged Van de Graff terminal? (b) What is unreasonable about this result? (c) Which assumptions are responsible?

95. A large metal plate is charged uniformly to a density of $\sigma = 2.0 \times 10^{-9} C/m^2$. How far apart are the equipotential surfaces that represent a potential difference of 25 V?

96. Your friend gets really excited by the idea of making a lightning rod or maybe just a sparking toy by connecting two spheres as shown in Figure 7.39, and making R_2 so small that the electric field is greater than the dielectric strength of air, just from the usual 150 V/m electric field near the surface of the Earth. If R_1 is 10 cm, how small does R_2 need to be, and does this seem practical? (**Hint**: recall the calculation for electric field at the surface of a conductor from Gauss's Law.)

97. (a) Find $x >> L$ limit of the potential of a finite uniformly charged rod and show that it coincides with that of a point charge formula. (b) Why would you expect this result?

98. A small spherical pith ball of radius 0.50 cm is painted with a silver paint and then $-10\mu C$ of charge is placed on it. The charged pith ball is put at the center of a gold spherical shell of inner radius 2.0 cm and outer radius 2.2 cm.

(a) Find the electric potential of the gold shell with respect to zero potential at infinity.

(b) How much charge should you put on the gold shell if you want to make its potential 100 V?

99. 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,

(a) what is the potential difference between the plates?

(b) What is the potential difference between the positive plate and a point 1.25 cm from it that is between the plates?

100. A point charge of $q = 5.0 \times 10^{-8} C$ 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 potential at

(a) $r = 4.0cm$,

(b) $r = 8.0cm$,

(c) $r = 12.0cm$.

101. Earth has a net charge that produces an electric field of approximately 150 N/C downward at its surface.

(a) What is the magnitude and sign of the excess charge, noting the electric field of a conducting sphere is equivalent to a point charge at its center?

(b) What acceleration will the field produce on a free electron near Earth's surface?

(c) What mass object with a single extra electron will have its weight supported by this field?

102. Point charges of $25.0\mu C$ and $45.0\mu C$ are placed 0.500 m apart.

(a) At what point along the line between them is the electric field zero?

(b) What is the electric field halfway between them?

103. What can you say about two charges q_1 and q_2, if the electric field one-fourth of the way from q_1 to q_2 is zero?

104. Calculate the angular velocity ω of an electron orbiting a proton in the hydrogen atom, given the radius of the orbit is $0.530 \times 10^{-10} m$. You may assume that the proton is stationary and the centripetal force is supplied by Coulomb attraction.

105. An electron has an initial velocity of $5.00 \times 10^6 m/s$ in a uniform $2.00 \times 10^5 - N/C$ electric field. The field accelerates the electron in the direction opposite to its initial velocity.

(a) What is the direction of the electric field?

(b) How far does the electron travel before coming to rest?

(c) How long does it take the electron to come to rest?

(d) What is the electron's velocity when it returns to its starting point?

Challenge Problems

106. Three Na^+ and three Cl^- ions are placed alternately and equally spaced around a circle of radius 50 nm. Find the electrostatic energy stored.

107. Look up (presumably online, or by dismantling an old device and making measurements) the magnitude of the potential deflection plates (and the space between them) in an ink jet printer. Then look up the speed with which the ink comes out the nozzle. Can you calculate the typical mass of an ink drop?

108. Use the electric field of a finite sphere with constant volume charge density to calculate the electric potential, throughout space. Then check your results by calculating the electric field from the potential.

109. Calculate the electric field of a dipole throughout space from the potential.

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).

4.8: Potential and Field Relationships (Answers)

Conceptual Questions

17. No. It will be constant, but not necessarily zero.

19. no

21. No; it might not be at electrostatic equilibrium.

23. Yes. It depends on where the zero reference for potential is. (Though this might be unusual.)

25. So that lightning striking them goes into the ground instead of the television equipment.

Problems

59. a. increases; the constant (negative) electric field has this effect, the reference point only matters for magnitude; b. they are planes parallel to the sheet; c. 0.006 m

61. a. from the previous chapter, the electric field has magnitude $\frac{\sigma}{\varepsilon_0}$ in the region between the plates and zero outside; defining the negatively charged plate to be at the origin and zero potential, with the positively charged plate located at **+5mm** in the **z**-direction, V=1.7×104VV=1.7×104Vso the potential is 0 for $z<0, 1.7\times10^4V(\frac{z}{5mm})$ for $0\le z\le 5mm, 1.7\times10^4V$ for $z>5mm$;

b. $qV=\frac{1}{2}mv^2\rightarrow v=7.7\times10^7m/s$

63. $V=85V$

65. In the region $a\le r\le b, \vec{E}=\frac{kQ}{r^2}\hat{r}$, and **E** is zero elsewhere; hence, the potential difference is $V=kQ(\frac{1}{a}-\frac{1}{b})$.

67. From previous results $V_P-V_R=-2k\lambda ln\frac{s_P}{s_R}$., note that b is a very convenient location to define the zero level of potential: $\Delta V=-2k\frac{Q}{L}ln\frac{a}{b}$.

69. a. $F=5.58\times10^{-11}N/C$;The electric field is towards the surface of Earth.

b. The coulomb force is much stronger than gravity.

71. We know from the Gauss's law chapter that the electric field for an infinite line charge is $\vec{E}_P=2k\lambda\frac{1}{s}\hat{s}$, and from earlier in this chapter that the potential of a wire-cylinder system of this sort is $V_P=-2k\lambda ln\frac{s_P}{R}$ by integration. We are not given λ, but we are given a fixed V_0; thus, we know that $V_0=-2k\lambda ln\frac{a}{R}$ and hence $\lambda=-\frac{V_0}{2kln(\frac{a}{R})}$. We may substitute this back in to find a. $\vec{E}_P=-\frac{V_0}{ln(\frac{a}{R})}\frac{1}{s}\hat{s}$;

b. $V_P=V_0\frac{ln(\frac{s_P}{R})}{ln(\frac{a}{R})}$;

c. $4.74\times10^4N/C$

73. a. $U_1=7.68\times10^{-18}J, U_2=5.76\times10^{-18}J$;

b. $U_1+U_2=-1.34\times10^{-17}J$

75. a. \(\displaystyle U=2.30×10^{−16}J;

b. $\overline{K}=\frac{3}{2}kT\rightarrow T=1.11\times10^7$

77. a. $1.9\times10^6m/s$;

b. $4.2 \times 10^6 m/s$;

c. $5.9 \times 10^6 m/s$;

d. $7.3 \times 10^6 m/s$;

e. $8.4 \times 10^6 m/s$

79. a. $E = 2.5 \times 10^6 V/m < 3 \times 10^6 V/m$ No, the field strength is smaller than the breakdown strength for air.

b. $d = 1.7mm$

81. $K_f = qV_{AB} = qEd \to E = 8.00 \times 10^5 V/m$

83. a. Energy=$2.00 \times 10^9 J$;

b. $Q = m(c\Delta T + L_\nabla)\ m = 766kg$;

c. The expansion of the steam upon boiling can literally blow the tree apart.

85. a. $V = \frac{kQ}{r} \to r = 1.80km$;

b. A 1-C charge is a very large amount of charge; a sphere of 1.80 km is impractical.

87. The alpha particle approaches the gold nucleus until its original energy is converted to potential energy. $5.00MeV = 8.00 \times 10^{-13} J$, so $E_0 = \frac{qkQ}{r} \to r = 4.54 \times 10^{-14} m$

(Size of gold nucleus is about $7 \times 10^{-15} m$).

Additional Problems

89. $E_{tot} = 4.67 \times 10^7 J\ E_{tot} = qV \to q = \frac{E_{tot}}{V} = 3.89 \times 10^6 C$

91. $V_P = k\frac{q_{tot}}{\sqrt{z^2+R^2}} \to q_{tot} = -3.5 \times 10^{-11} C$

93. $V_P = -2.2GV$

95. Recall from the previous chapter that the electric field $E_P = \frac{\sigma_0}{2\varepsilon_0}$ is uniform throughout space, and that for uniform fields we have $E = -\frac{\Delta V}{\Delta z}$ for the relation. Thus, we get $\frac{\sigma}{2\varepsilon_0} = \frac{\Delta V}{\Delta z} \to \Delta z = 0.22m$ for the distance between 25-V equipotentials.

97. a. Take the result from Example 7.13, divide both the numerator and the denominator by x, take the limit of that, and then apply a Taylor expansion to the resulting log to get: $V_P \approx k\lambda\frac{L}{x}$;

b. which is the result we expect, because at great distances, this should look like a point charge of $q = \lambda L$

99. a. $V = 9.0 \times 10^3 V$;

b. $-9.0 \times 10^3 V(\frac{1.25cm}{2.0cm}) = -5.7 \times 10^3 V$

101. a. $E = \frac{KQ}{r^2} \to Q = -6.76 \times 10^5 C$;

b. $F = ma = qE \to a = \frac{qE}{m} = 2.63 \times 10^{13} m/s^2 (upwards)$;

c. $F = -mg = qE \to m = \frac{-qE}{g} = 2.45 \times 10^{-18} kg$

103. If the electric field is zero ¼ from the way of q_1 and q_2, then we know from $E = k\frac{Q}{r^2}$ that $|E_1| = |E_2| \to \frac{Kq_1}{x^2} = \frac{Kq^2}{(3x)^2}$ so that $\frac{q_2}{q_1} = \frac{(3x)^2}{x^2} = 9$; the charge q_2 is 9 times larger than q_1.

105. a. The field is in the direction of the electron's initial velocity.

b. $v^2 = v_0^2 + 2ax \to x = -\frac{v_0^2}{2a}(v=0)$. Also, $F = ma = qE \to a = \frac{qE}{m}$,$x = 3.56 \times 10^{-4} m$;

c. $v_2 = v_0 + at \to t = -\frac{v_0 m}{qE}(v=0), \therefore t = 1.42 \times 10^{-10} s$;

d. $v = -(\frac{2qEx}{m})^{1/2} - 5.00 \times 10^6 m/s$ (opposite its initial velocity)

Challenge Problems

107. Answers will vary. This appears to be proprietary information, and ridiculously difficult to find. Speeds will be 20 m/s or less, and there are claims of 10^{-7} grams for the mass of a drop.

109. Apply $\vec{E} = -\vec{\nabla}V$ with $\vec{\nabla} = \hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\varphi}\frac{1}{rsin\theta}\frac{\partial}{\partial \varphi}$ to the potential calculated earlier, $V_P = k\frac{\vec{p}\cdot\hat{r}}{r^2}$ with $\vec{p} = q\vec{d}$, and assume that the axis of the dipole is aligned with the **z**-axis of the coordinate system. Thus, the potential is $V_P = k\frac{q\vec{d}\cdot\hat{r}}{r^2} = k\frac{qdcos\theta}{r^2}$.

$$\vec{E} = 2kqd(\frac{cos\theta}{r^3})\hat{r} + kqd(\frac{sin\theta}{r^3})\hat{\theta}$$

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).

CHAPTER OVERVIEW

5: Electric Current and Resistance

5.1: Introduction

In this chapter, we study the electrical **current**, which defined as the flow of charge through a material. We also examine a characteristic of materials known as its resistance. Resistance is a measure of how much a material impedes the flow of charge, and it will be shown that the resistance depends on temperature. In general, a good conductor, such as copper, gold, or silver, has very low resistance.

One way we can tell if current is flowing through a material is if it gets warm. In many cases, these losses of electrical energy to thermal energy are undesirable. But in an electric toaster or space heater they can be exactly what is needed: a compact, non-combustible source of heat (Fig. 5.1.1).

Figure 5.1.1: Current flowing through resistive elements in a toaster cause them to give off heat and light. [1]

References

1. Wikimedia Commons contributors. File:Toaster Filaments.JPG [Internet]. Wikimedia Commons. (Nick Carson, CC-BY 3.0)

5.2: Electric Current

LEARNING OBJECTIVES

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

- Describe an electrical current
- Define the unit of electrical current
- Explain the direction of current flow

Up to now, we have considered primarily static charges. When charges did move, they were accelerated in response to an electrical field created by a voltage difference. The charges lost potential energy and gained kinetic energy as they traveled through a potential difference where the electrical field did work on the charge.

Although charges do not require a material to flow through, the majority of this chapter deals with understanding the movement of charges through a material. The rate at which the charges flow past a location—that is, the amount of charge per unit time—is known as the **electrical current**. When charges flow through a medium, the current depends on the voltage applied, the material through which the charges flow, and the state of the material. Of particular interest is the motion of charges in a conducting wire. In previous chapters, charges were accelerated due to the force provided by an electrical field, losing potential energy and gaining kinetic energy. In this chapter, we discuss the situation of the force provided by an electrical field in a conductor, where charges lose kinetic energy to the material reaching a constant velocity, known as the "**drift velocity**." This is analogous to an object falling through the atmosphere and losing kinetic energy to the air, reaching a constant terminal velocity.

If you have ever taken a course in first aid or safety, you may have heard that in the event of electric shock, it is the current, not the voltage, which is the important factor on the severity of the shock and the amount of damage to the human body. Current is measured in units called amperes; you may have noticed that circuit breakers in your home and fuses in your car are rated in amps (or amperes). But what is the ampere and what does it measure?

Defining Current and the Ampere

Electrical current is defined to be the rate at which charge flows. When there is a large current present, such as that used to run a refrigerator, a large amount of charge moves through the wire in a small amount of time. If the current is small, such as that used to operate a handheld calculator, a small amount of charge moves through the circuit over a long period of time.

Electrical Current

The average electrical current I is the rate at which charge flows,

$$I_{ave} = \frac{\Delta Q}{\Delta t}, \tag{5.2.1}$$

where ΔQ is the amount of net charge passing through a given cross-sectional area in time Δt (Figure $5.2.1$). The SI unit for current is the **ampere** (A), named for the French physicist André-Marie Ampère (1775–1836). Since $I = \frac{\Delta Q}{\Delta t}$, we see that an ampere is defined as one coulomb of charge passing through a given area per second:

$$1A \equiv 1\frac{C}{s}. \tag{5.2.2}$$

The instantaneous electrical current, or simply the **electrical current**, is the time derivative of the charge that flows and is found by taking the limit of the average electrical current as $\Delta t \to 0$.

$$I = \lim_{\Delta t \to 0} \frac{\Delta Q}{\Delta t} = \frac{dQ}{dt}. \tag{5.2.3}$$

Most electrical appliances are rated in amperes (or amps) required for proper operation, as are fuses and circuit breakers.

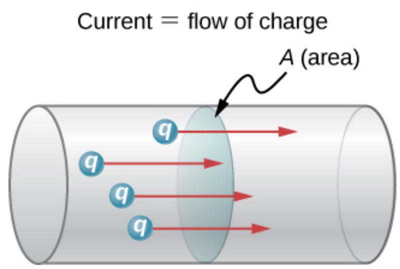


Figure 5.2.1: The rate of flow of charge is current. An ampere is the flow of one coulomb of charge through an area in one second. A current of one amp would result from 6.25×10^{18} electrons flowing through the area A each second.

Calculating the Average Current

The main purpose of a battery in a car or truck is to run the electric **starter motor**, which starts the engine. The operation of starting the vehicle requires a large current to be supplied by the battery. Once the engine starts, a device called an alternator takes over supplying the electric power required for running the vehicle and for charging the battery.

a. What is the average current involved when a truck battery sets in motion 720 C of charge in 4.00 s while starting an engine?
b. How long does it take 1.00 C of charge to flow from the battery?

Strategy

We can use the definition of the average current in Equation 5.2.1 to find the average current in part (a), since charge and time are given. For part (b), once we know the average current, we use Equation 5.2.1 to find the time required for 1.00 C of charge to flow from the battery.

Solution

a. Entering the given values for charge and time into the definition of current gives

$$\begin{aligned} I &= \frac{\Delta Q}{\Delta t} \\ &= \frac{720\, C}{4.00\, s} \\ &= 180\, C/s \\ &= 180\, A. \end{aligned}$$

b. Solving the relationship $I = \dfrac{\Delta Q}{\Delta t}$ for time Δt and entering the known values for charge and current gives

$$\begin{aligned} \Delta t &= \frac{\Delta Q}{I} \\ &= \frac{1.00\, C}{180\, C/s} \\ &= 5.56 \times 10^{-3} s \\ &= 5.56\, ms. \end{aligned}$$

Significance

a. This large value for current illustrates the fact that a large charge is moved in a small amount of time. The currents in these "starter motors" are fairly large to overcome the inertia of the engine.
b. A high current requires a short time to supply a large amount of charge. This large current is needed to supply the large amount of energy needed to start the engine.

Calculating Instantaneous Currents

Consider a charge moving through a cross-section of a wire where the charge is modeled as $Q(t) = Q_M(1 - e^{-t/\tau})$. Here, Q_M is the charge after a long period of time, as time approaches infinity, with units of coulombs, and τ is a time constant with units of seconds (Figure $5.2.2$). What is the current through the wire?

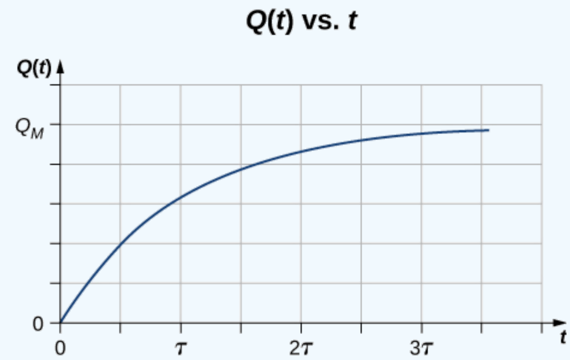


Figure $5.2.2$: A graph of the charge moving through a cross-section of a wire over time.

Strategy

The current through the cross-section can be found from $I = \frac{dQ}{dt}$. Notice from the figure that the charge increases to Q_M and the derivative decreases, approaching zero, as time increases (Figure $5.2.2$).

Solution

The derivative can be found using $\frac{d}{dx}e_u = e^u \frac{du}{dx}$.

$$\begin{aligned} I &= \frac{dQ}{dt} \\ &= \frac{d}{dt}\left[Q_M\left(1 - e^{-t/\tau}\right)\right] \\ &= \frac{Q_M}{\tau}e^{-t/\tau}. \end{aligned}$$

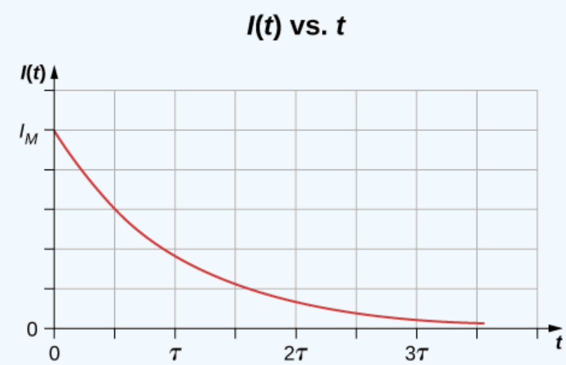


Figure $5.2.3$: A graph of the current flowing through the wire over time.

Significance

The current through the wire in question decreases exponentially, as shown in Figure $5.2.3$. In later chapters, it will be shown that a time-dependent current appears when a capacitor charges or discharges through a resistor. Recall that a capacitor is a device that stores charge. You will learn about the resistor in Model of Conduction in Metals.

Exercise $5.2.1A$

Handheld calculators often use small solar cells to supply the energy required to complete the calculations needed to complete your next physics exam. The current needed to run your calculator can be as small as 0.30 mA. How long would it take for 1.00 C of charge to flow from the solar cells? Can solar cells be used, instead of batteries, to start traditional internal combustion engines presently used in most cars and trucks?

Answer

The time for 1.00 C of charge to flow would be

$$\Delta t=\frac{\Delta Q}{I}=\frac{1.00\,C}{0.300\times 10^{-3}\,C/s}=3.33\times 10^{3}\ s. \tag{5.2.4}$$

This is slightly less than an hour. This is quite different from the 5.55 ms for the truck battery. The calculator takes a very small amount of energy to operate, unlike the truck's starter motor. There are several reasons that vehicles use batteries and not solar cells. Aside from the obvious fact that a light source to run the solar cells for a car or truck is not always available, the large amount of current needed to start the engine cannot easily be supplied by present-day solar cells. Solar cells can possibly be used to charge the batteries. Charging the battery requires a small amount of energy when compared to the energy required to run the engine and the other accessories such as the heater and air conditioner. Present day solar-powered cars are powered by solar panels, which may power an electric motor, instead of an internal combustion engine.

Exercise $5.2.1B$

Circuit breakers in a home are rated in amperes, normally in a range from 10 amps to 30 amps, and are used to protect the residents from harm and their appliances from damage due to large currents. A single 15-amp circuit breaker may be used to protect several outlets in the living room, whereas a single 20-amp circuit breaker may be used to protect the refrigerator in the kitchen. What can you deduce from this about current used by the various appliances?

Answer

The total current needed by all the appliances in the living room (a few lamps, a television, and your laptop) draw less current and require less power than the refrigerator.

Current in a Circuit

In the previous paragraphs, we defined the current as the charge that flows through a cross-sectional area per unit time. In order for charge to flow through an appliance, such as the headlight shown in Figure $5.2.4$, there must be a complete path (or **circuit**) from the positive terminal to the negative terminal. Consider a simple circuit of a car battery, a switch, a headlight lamp, and wires that provide a current path between the components. In order for the lamp to light, there must be a complete path for current flow. In other words, a charge must be able to leave the positive terminal of the battery, travel through the component, and back to the negative terminal of the battery. The switch is there to control the circuit. Part (a) of the figure shows the simple circuit of a car battery, a switch, a conducting path, and a headlight lamp. Also shown is the **schematic** of the circuit [part (b)]. A schematic is a graphical representation of a circuit and is very useful in visualizing the main features of a circuit. Schematics use standardized symbols to represent the components in a circuits and solid lines to represent the wires connecting the components. The battery is shown as a series of long and short lines, representing the historic voltaic pile. The lamp is shown as a circle with a loop inside, representing the filament of an incandescent bulb. The switch is shown as two points with a conducting bar to connect the two points and the wires connecting the components are shown as solid lines. The schematic in part (c) shows the direction of current flow when the switch is closed.

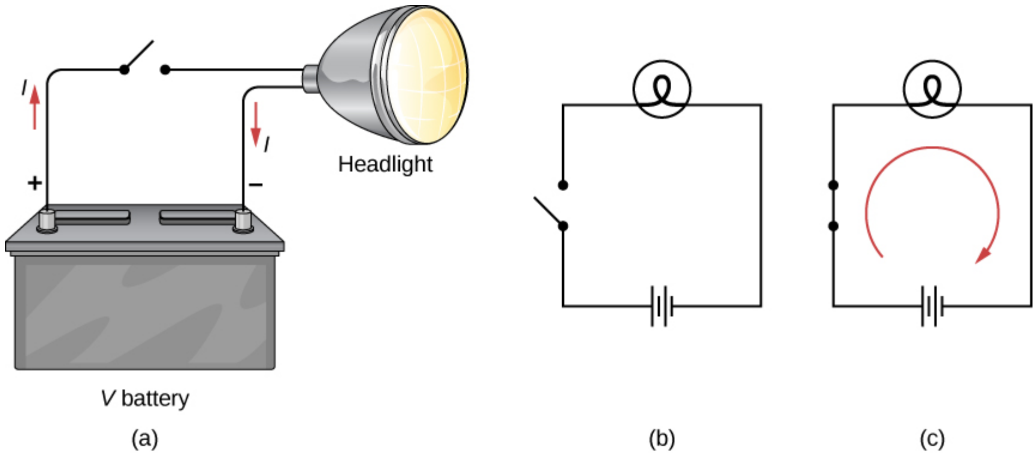


Figure 5.2.4: (a) A simple electric circuit of a headlight (lamp), a battery, and a switch. When the switch is closed, an uninterrupted path for current to flow through is supplied by conducting wires connecting a load to the terminals of a battery. (b) In this schematic, the battery is represented by parallel lines, which resemble plates in the original design of a battery. The longer lines indicate the positive terminal. The conducting wires are shown as solid lines. The switch is shown, in the open position, as two terminals with a line representing a conducting bar that can make contact between the two terminals. The lamp is represented by a circle encompassing a filament, as would be seen in an incandescent light bulb. (c) When the switch is closed, the circuit is complete and current flows from the positive terminal to the negative terminal of the battery.

When the switch is closed in Figure 5.2.4c, there is a complete path for charges to flow, from the positive terminal of the battery, through the switch, then through the headlight and back to the negative terminal of the battery. Note that the direction of current flow is from positive to negative. The direction of conventional current is always represented in the direction that positive charge would flow, from the positive terminal to the negative terminal.

The conventional current flows from the positive terminal to the negative terminal, but depending on the actual situation, positive charges, negative charges, or both may move. In metal wires, for example, current is carried by electrons—that is, negative charges move. In ionic solutions, such as salt water, both positive and negative charges move. This is also true in nerve cells. A Van de Graaff generator, used for nuclear research, can produce a current of pure positive charges, such as protons. In the Tevatron Accelerator at Fermilab, before it was shut down in 2011, beams of protons and antiprotons traveling in opposite directions were collided. The protons are positive and therefore their current is in the same direction as they travel. The antiprotons are negativity charged and thus their current is in the opposite direction that the actual particles travel.

A closer look at the current flowing through a wire is shown in Figure 5.2.5. The figure illustrates the movement of charged particles that compose a current. The fact that conventional current is taken to be in the direction that positive charge would flow can be traced back to American scientist and statesman Benjamin Franklin in the 1700s. Having no knowledge of the particles that make up the atom (namely the proton, electron, and neutron), Franklin believed that electrical current flowed from a material that had more of an “electrical fluid” and to a material that had less of this “electrical fluid.” He coined the term **positive** for the material that had more of this electrical fluid and **negative** for the material that lacked the electrical fluid. He surmised that current would flow from the material with more electrical fluid—the positive material—to the negative material, which has less electrical fluid. Franklin called this direction of current a positive current flow. This was pretty advanced thinking for a man who knew nothing about the atom.

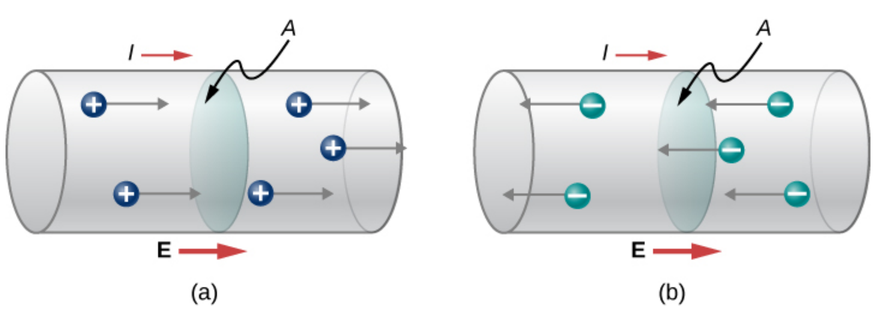


Figure 5.2.5: Current **I** is the rate at which charge moves through an area **A**, such as the cross-section of a wire. Conventional current is defined to move in the direction of the electrical field. (a) Positive charges move in the direction of the electrical field, which is the same direction as conventional current. (b) Negative charges move in the direction opposite to the electrical field. Conventional current is in the direction opposite to the movement of negative charge. The flow of electrons is sometimes referred to as electronic flow.

We now know that a material is positive if it has a greater number of protons than electrons, and it is negative if it has a greater number of electrons than protons. In a conducting metal, the current flow is due primarily to electrons flowing from the negative material to the positive material, but for historical reasons, we consider the positive current flow and the current is shown to flow from the positive terminal of the battery to the negative terminal.

It is important to realize that an electrical field is present in conductors and is responsible for producing the current (Figure 5.2.5). In previous chapters, we considered the static electrical case, where charges in a conductor quickly redistribute themselves on the surface of the conductor in order to cancel out the external electrical field and restore equilibrium. In the case of an electrical circuit, the charges are prevented from ever reaching equilibrium by an external source of electric potential, such as a battery. The energy needed to move the charge is supplied by the electric potential from the battery.

Although the electrical field is responsible for the motion of the charges in the conductor, the work done on the charges by the electrical field does not increase the kinetic energy of the charges. We will show that the electrical field is responsible for keeping the electric charges moving at a "drift velocity."

5.3: Basic Model of Conduction in Metals

LEARNING OBJECTIVES

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

- Define the drift velocity of charges moving through a metal
- Define the vector current density
- Describe the operation of an incandescent lamp

When electrons move through a conducting wire, they do not move at a constant velocity, that is, the electrons do not move in a straight line at a constant speed. Rather, they interact with and collide with atoms and other free electrons in the conductor. Thus, the electrons move in a zig-zag fashion and drift through the wire. We should also note that even though it is convenient to discuss the direction of current, current is a scalar quantity. When discussing the velocity of charges in a current, it is more appropriate to discuss the current density. We will come back to this idea at the end of this section.

Drift Velocity

Electrical signals move very rapidly. Telephone conversations carried by currents in wires cover large distances without noticeable delays. Lights come on as soon as a light switch is moved to the 'on' position. Most electrical signals carried by currents travel at speeds on the order of $10^8 m/s$ a significant fraction of the speed of light. Interestingly, the individual charges that make up the current move much slower on average, typically drifting at speeds on the order of $10^{-4} m/s$. How do we reconcile these two speeds, and what does it tell us about standard conductors?

The high speed of electrical signals results from the fact that the force between charges acts rapidly at a distance. Thus, when a free charge is forced into a wire, as in Figure $5.3.1$, the incoming charge pushes other charges ahead of it due to the repulsive force between like charges. These moving charges push on charges farther down the line. The density of charge in a system cannot easily be increased, so the signal is passed on rapidly. The resulting electrical shock wave moves through the system at nearly the speed of light. To be precise, this fast-moving signal, or shock wave, is a rapidly propagating change in the electrical field.

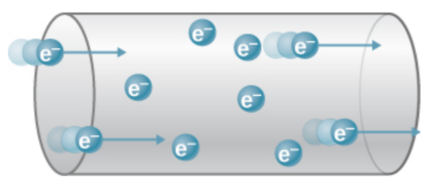


Figure $5.3.1$: When charged particles are forced into this volume of a conductor, an equal number are quickly forced to leave. The repulsion between like charges makes it difficult to increase the number of charges in a volume. Thus, as one charge enters, another leaves almost immediately, carrying the signal rapidly forward.

Good conductors have large numbers of free charges. In metals, the free charges are free electrons. (In fact, good electrical conductors are often good heat conductors too, because large numbers of free electrons can transport thermal energy as well as carry electrical current.) Figure $5.3.2$ shows how free electrons move through an ordinary conductor. The distance that an individual electron can move between collisions with atoms or other electrons is quite small. The electron paths thus appear nearly random, like the motion of atoms in a gas. But there is an electrical field in the conductor that causes the electrons to drift in the direction shown (opposite to the field, since they are negative). The **drift velocity** $\vec{v}_d$ is the average velocity of the free charges. Drift velocity is quite small, since there are so many free charges. If we have an estimate of the density of free electrons in a conductor, we can calculate the drift velocity for a given current. The larger the density, the lower the velocity required for a given current.

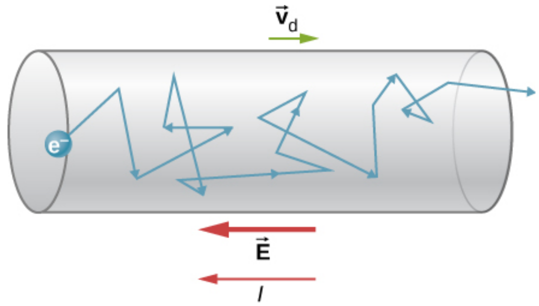


Figure $5.3.2$: Free electrons moving in a conductor make many collisions with other electrons and other particles. A typical path of one electron is shown. The average velocity of the free charges is called the drift velocity $\vec{v}_d$ and for electrons, it is in the direction opposite to the electrical field. The collisions normally transfer energy to the conductor, requiring a constant supply of energy to maintain a steady current.

Free-electron collisions transfer energy to the atoms of the conductor. The electrical field does work in moving the electrons through a distance, but that work does not increase the kinetic energy (nor speed) of the electrons. The work is transferred to the conductor's atoms, often increasing temperature. Thus, a continuous power input is required to keep a current flowing. (An exception is superconductors, for reasons we shall explore in a later chapter. Superconductors can have a steady current without a continual supply of energy—a great energy savings.) For a conductor that is not a superconductor, the supply of energy can be useful, as in an incandescent light bulb filament (Figure $5.3.3$). The supply of energy is necessary to increase the temperature of the tungsten filament, so that the filament glows.

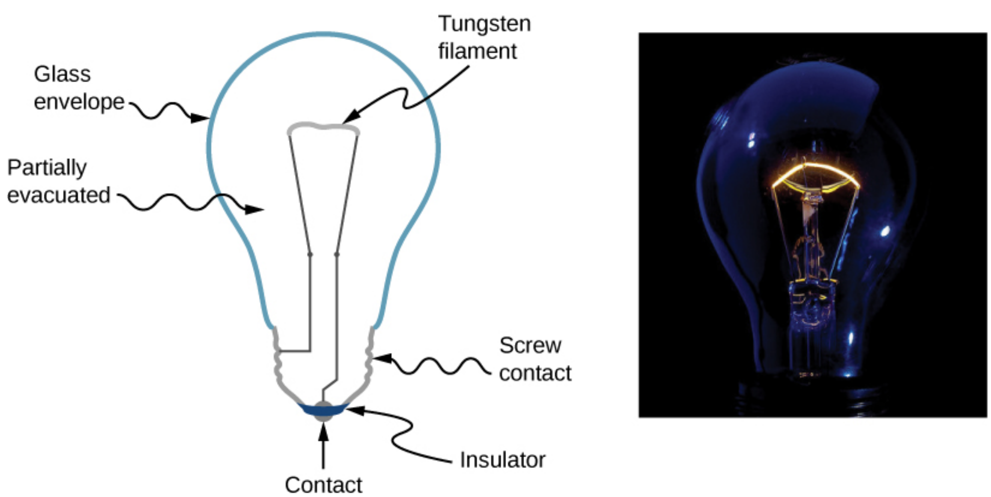


Figure $5.3.3$: The incandescent lamp is a simple design. A tungsten filament is placed in a partially evacuated glass envelope. One end of the filament is attached to the screw base, which is made out of a conducting material. The second end of the filament is attached to a second contact in the base of the bulb. The two contacts are separated by an insulating material. Current flows through the filament, and the temperature of the filament becomes large enough to cause the filament to glow and produce light. However, these bulbs are not very energy efficient, as evident from the heat coming from the bulb. In the year 2012, the United States, along with many other countries, began to phase out incandescent lamps in favor of more energy-efficient lamps, such as light-emitting diode (LED) lamps and compact fluorescent lamps (CFL) (credit right: modification of work by Serge Saint).

We can obtain an expression for the relationship between current and drift velocity by considering the number of free charges in a segment of wire, as illustrated in Figure $5.3.4$. The number of free charges per unit volume, or the number density of free charges, is given the symbol n where

$$n = \frac{\text{number of charges}}{\text{volume}}.$$

The value of n depends on the material. The shaded segment has a volume $Av_d\,dt$, so that the number of free charges in the volume is $nAv_d\,dt$. The charge dQ in this segment is thus $qnAv_d\,dt$, where q is the amount of charge on each carrier. (The magnitude of the charge of electrons is $q = 1.60 \times 10^{-19}\,C$.) Current is charge moved per unit time; thus, if all the original charges move out of this segment in time **dt**, the current is

$$I = \frac{dQ}{dt} = qnAv_d.$$

Rearranging terms gives

$$v_d = \frac{I}{nqA}$$

where

- v_d is the drift velocity,
- $\boldsymbol{n}$ is the free charge density,
- $\boldsymbol{A}$ is the cross-sectional area of the wire, and
- I is the current through the wire.

The carriers of the current each have charge **q** and move with a drift velocity of magnitude v_d.

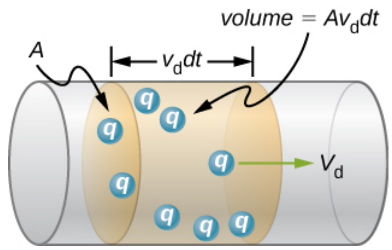


Figure $5.3.4$: All the charges in the shaded volume of this wire move out in a time **dt**, having a drift velocity of magnitude v_d.

Note that simple drift velocity is not the entire story. The speed of an electron is sometimes much greater than its drift velocity. In addition, not all of the electrons in a conductor can move freely, and those that do move might move somewhat faster or slower than the drift velocity. So what do we mean by free electrons?

Atoms in a metallic conductor are packed in the form of a lattice structure. Some electrons are far enough away from the atomic nuclei that they do not experience the attraction of the nuclei as strongly as the inner electrons do. These are the free electrons. They are not bound to a single atom but can instead move freely among the atoms in a "sea" of electrons. When an electrical field is applied, these free electrons respond by accelerating. As they move, they collide with the atoms in the lattice and with other electrons, generating thermal energy, and the conductor gets warmer. In an insulator, the organization of the atoms and the structure do not allow for such free electrons.

As you know, electric power is usually supplied to equipment and appliances through round wires made of a conducting material (copper, aluminum, silver, or gold) that are stranded or solid. The diameter of the wire determines the current-carrying capacity—the larger the diameter, the greater the current-carrying capacity. Even though the current-carrying capacity is determined by the diameter, wire is not normally characterized by the diameter directly. Instead, wire is commonly sold in a unit known as "gauge." Wires are manufactured by passing the material through circular forms called "drawing dies." In order to make thinner wires, manufacturers draw the wires through multiple dies of successively thinner diameter. Historically, the gauge of the wire was related to the number of drawing processes required to manufacture the wire. For this reason, the larger the gauge, the smaller the diameter. In the United States, the American Wire Gauge (AWG) was developed to standardize the system. Household wiring commonly consists of 10-gauge (2.588-mm diameter) to 14-gauge (1.628-mm diameter) wire. A device used to measure the gauge of wire is shown in Figure $5.3.5$.

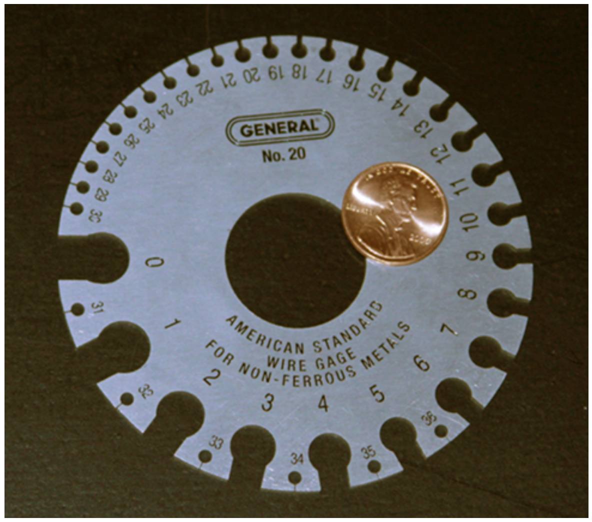


Figure 5.3.5: A device for measuring the gauge of electrical wire. As you can see, higher gauge numbers indicate thinner wires.

Example 5.3.1: Calculating Drift Velocity in a Common Wire

Calculate the drift velocity of electrons in a copper wire with a diameter of 2.053 mm (12-gauge) carrying a 20.0-A current, given that there is one free electron per copper atom. (Household wiring often contains 12-gauge copper wire, and the maximum current allowed in such wire is usually 20.0 A.) The density of copper is $8.80 \times 10^3 kg/m^3$ and the atomic mass of copper is 63.54 g/mol.

Strategy

We can calculate the drift velocity using the equation $I = nqAv_d$. The current is $I = 20.00\,A$ and $q = 1.60 \times 10{-}19C$ is the charge of an electron. We can calculate the area of a cross-section of the wire using the formula $A = \pi r^2$, where r is one-half the diameter. The given diameter is 2.053 mm, so **r** is 1.0265 mm. We are given the density of copper, $8.80 \times 10^3 kg/m^3$, and the atomic mass of copper is $63.54\,g/mol$. We can use these two quantities along with Avogadro's number, $6.02 \times 10^{23}\,atoms/mol$, to determine **n**, the number of free electrons per cubic meter.

Solution

First, we calculate the density of free electrons in copper. There is one free electron per copper atom. Therefore, the number of free electrons is the same as the number of copper atoms per m^3. We can now find **n** as follows:

$$\begin{aligned} n &= \frac{1\,e^-}{atom} \times \frac{6.02 \times 10^{23} atoms}{mol} \times \frac{1\,mol}{63.54\,g} \times \frac{1000\,g}{kg} \times \frac{8.80 \times 10^3 kg}{1\,m^3} \\ &= 8.34 \times 10^{28} e^-/m^3. \end{aligned}$$

The cross-sectional area of the wire is

$$\begin{aligned} A &= \pi r^2 \\ &= \pi \left(\frac{2.05 \times 10^{-3} m}{2} \right)^2 \\ &= 3.30 \times 10^{-6} m^2. \end{aligned}$$

Rearranging $I = nqAv_d$ to isolate drift velocity gives

$$\begin{aligned} v_d &= \frac{I}{nqA} \\ &= \frac{20.00\,A}{(8.34 \times 10^{28}/m^3)(-1.60 \times 10^{-19} C)(3.30 \times 10^{-6} m^2)} \\ &= -4.54 \times 10^{-4} m/s. \end{aligned}$$

Significance

The minus sign indicates that the negative charges are moving in the direction opposite to conventional current. The small value for drift velocity (on the order of $10^{-4} m/s$ confirms that the signal moves on the order of 10^{12} times faster (about $10^{8} m/s$ than the charges that carry it.

? Exercise 5.3.1

In Example 5.3.1, the drift velocity was calculated for a 2.053-mm diameter (12-gauge) copper wire carrying a 20-amp current. Would the drift velocity change for a 1.628-mm diameter (14-gauge) wire carrying the same 20-amp current?

Answer

The diameter of the 14-gauge wire is smaller than the diameter of the 12-gauge wire. Since the drift velocity is inversely proportional to the cross-sectional area, the drift velocity in the 14-gauge wire is larger than the drift velocity in the 12-gauge wire carrying the same current. The number of electrons per cubic meter will remain constant.

Current Density

Although it is often convenient to attach a negative or positive sign to indicate the overall direction of motion of the charges, current is a scalar quantity, $I = \frac{dQ}{dt}$. It is often necessary to discuss the details of the motion of the charge, instead of discussing the overall motion of the charges. In such cases, it is necessary to discuss the current density, $\vec{J}$, a vector quantity. The **current density** is the flow of charge through an infinitesimal area, divided by the area. The current density must take into account the local magnitude and direction of the charge flow, which varies from point to point. The unit of current density is ampere per meter squared, and the direction is defined as the direction of net flow of positive charges through the area.

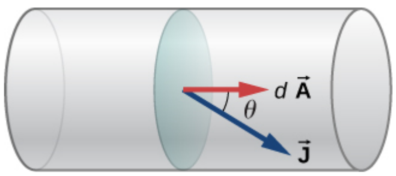


Figure 5.3.6: The current density $\vec{J}$ is defined as the current passing through an infinitesimal cross-sectional area divided by the area. The direction of the current density is the direction of the net flow of positive charges and the magnitude is equal to the current divided by the infinitesimal area.

The relationship between the current and the current density can be seen in Figure 5.3.6. The differential current flow through the area $d\vec{A}$ is found as

$$dI = \vec{J} \cdot d\vec{A} = JdA \cos \theta,$$

where θ is the angle between the area and the current density. The total current passing through area $d\vec{A}$ can be found by integrating over the area,

$$I = \iint_{area} \vec{J} \cdot d\vec{A}.$$

Consider the magnitude of the current density, which is the current divided by the area:

$$J = \frac{I}{A} = \frac{n|q|Av_d}{A} = n|q|v_d.$$

Thus, the current density is $\vec{J} = nq\vec{v}_d$. If **q** is positive, $\vec{v}_d$ is in the same direction as the electrical field $\vec{E}$. If **q** is negative, $\vec{v}_d$ is in the opposite direction of $\vec{E}$. Either way, the direction of the current density $\vec{J}$ is in the direction of the electrical field $\vec{E}$.

✓ Example 5.3.2: Calculating the Current Density in a Wire

The current supplied to a lamp with a 100-W light bulb is 0.87 amps. The lamp is wired using a copper wire with diameter 2.588 mm (10-gauge). Find the magnitude of the current density.

Strategy

The current density is the current moving through an infinitesimal cross-sectional area divided by the area. We can calculate the magnitude of the current density using $J = \frac{I}{A}$. The current is given as 0.87 A. The cross-sectional area can be calculated to be $A = 5.26\, mm^2$.

Solution

Calculate the current density using the given current $I = 0.87\, A$ and the area, found to be $A = 5.26\, mm^2$.

$$J = \frac{I}{A} = \frac{0.87\, A}{5.26 \times 10^{-6} m^2} = 1.65 \times 10^5 \frac{A}{m^2}.$$

Significance

The current density in a conducting wire depends on the current through the conducting wire and the cross-sectional area of the wire. For a given current, as the diameter of the wire increases, the charge density decreases.

? Exercise 5.3.2

The current density is proportional to the current and inversely proportional to the area. If the current density in a conducting wire increases, what would happen to the drift velocity of the charges in the wire?

Answer

The current density in a conducting wire increases due to an increase in current. The drift velocity is inversely proportional to the current $\left(v_d = \frac{nqA}{I}\right)$, so the drift velocity would decrease.

What is the significance of the current density? The current density is proportional to the current, and the current is the number of charges that pass through a cross-sectional area per second. The charges move through the conductor, accelerated by the electric force provided by the electrical field. The electrical field is created when a voltage is applied across the conductor. In Ohm's Law, we will use this relationship between the current density and the electrical field to examine the relationship between the current through a conductor and the voltage applied.

5.4: Resistivity and Resistance

Learning Objectives

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

- Differentiate between resistance and resistivity
- Define the term conductivity
- Describe the electrical component known as a resistor
- State the relationship between resistance of a resistor and its length, cross-sectional area, and resistivity
- State the relationship between resistivity and temperature

What drives current? We can think of various devices—such as batteries, generators, wall outlets, and so on—that are necessary to maintain a current. All such devices create a potential difference and are referred to as voltage sources. When a voltage source is connected to a conductor, it applies a potential difference **V** that creates an electrical field. The electrical field, in turn, exerts force on free charges, causing current. The amount of current depends not only on the magnitude of the voltage, but also on the characteristics of the material that the current is flowing through. The material can resist the flow of the charges, and the measure of how much a material resists the flow of charges is known as the **resistivity**. This resistivity is crudely analogous to the friction between two materials that resists motion.

Resistivity

When a voltage is applied to a conductor, an electrical field $\vec{E}$ is created, and charges in the conductor feel a force due to the electrical field. The current density $\vec{J}$ that results depends on the electrical field and the properties of the material. This dependence can be very complex. In some materials, including metals at a given temperature, the current density is approximately proportional to the electrical field. In these cases, the current density can be modeled as

$$\vec{J}=\sigma\vec{E},$$

where σ is the **electrical conductivity**. The electrical conductivity is analogous to thermal conductivity and is a measure of a material's ability to conduct or transmit electricity. Conductors have a higher electrical conductivity than insulators. Since the electrical conductivity is $\sigma=J/E$, the units are

$$\sigma=\frac{|J|}{|E|}=\frac{A/m^2}{V/m}=\frac{A}{V\cdot m}.$$

Here, we define a unit named the **ohm** with the Greek symbol uppercase omega, Ω. The unit is named after Georg Simon Ohm, whom we will discuss later in this chapter. The Ω is used to avoid confusion with the number 0. One ohm equals one volt per amp: $1\,\Omega=1\,V/A$. The units of electrical conductivity are therefore $(\Omega\cdot m)^{-1}$.

Conductivity is an intrinsic property of a material. Another intrinsic property of a material is the resistivity, or electrical **resistivity**. The resistivity of a material is a measure of how strongly a material opposes the flow of electrical current. The symbol for resistivity is the lowercase Greek letter rho, ρ, and resistivity is the reciprocal of electrical conductivity:

$$\rho=\frac{1}{\sigma}.$$

The unit of resistivity in SI units is the ohm-meter $(\Omega\cdot m$. We can define the resistivity in terms of the electrical field and the current density.

$$\rho=\frac{E}{J}.$$

The greater the resistivity, the larger the field needed to produce a given current density. The lower the resistivity, the larger the current density produced by a given electrical field. Good conductors have a high conductivity and low resistivity. Good insulators have a low conductivity and a high resistivity. Table $5.4.1$ lists resistivity and conductivity values for various materials.

Table $5.4.1$: Resistivities and Conductivities of Various Materials at 20 °C[1] Values depend strongly on amounts and types of impurities.

Material	Conductivity, σ $(\Omega \cdot m)^{-1}$	Resistivity, ρ $(\Omega \cdot m)$	Temperature Coefficient α $(^{o}C)^{-1}$
Conductors			
Silver	6.29×10^{7}	1.59×10^{-8}	0.0038
Copper	5.95×10^{7}	1.68×10^{-8}	0.0039
Gold	4.10×10^{7}	2.44×10^{-8}	0.0034
Aluminum	3.77×10^{7}	2.65×10^{-8}	0.0039
Tungsten	1.79×10^{7}	5.60×10^{-8}	0.0045
Iron	1.03×10^{7}	9.71×10^{-8}	0.0065
Platinum	0.94×10^{7}	10.60×10^{-8}	0.0039
Steel	0.50×10^{7}	20.00×10^{-8}	
Lead	0.45×10^{7}	22.00×10^{-8}	
Manganin (Cu, Mn. Ni alloy)	0.21×10^{7}	48.20×10^{-8}	0.000002
Constantan (Cu, Ni alloy)	0.20×10^{7}	49.00×10^{-8}	0.00003
Mercury	0.10×10^{7}	98.00×10^{-8}	0.0009
Nichrome (Ni, Fe, Cr alloy)	0.10×10^{7}	100.00×10^{-8}	0.0004
Semiconductors [1]			
Carbon (pure)	2.86×10^{4}	3.50×10^{-5}	-0.0005
Carbon	$(2.86 - 1.67) \times 10^{-6}$	$(3.5 - 60) \times 10^{-5}$	-0.0005
Germanium (pure)		600×10^{-3}	-0.048
Germanium		$(1 - 600) \times 10^{-3}$	-0.050
Silicon (pure)		2300	-0.075
Silicon		0.1 - 2300	-0.07
Insulators			
Amber	2.00×10^{-15}	5×10^{14}	
Glass	$10^{-9} - 19^{-14}$	$10^{9} - 10^{14}$	
Lucite	$< 10^{-13}$	$> 10^{13}$	
Mica	$10^{-11} - 10^{-15}$	$10^{11} - 10^{15}$	
Quartz (fused)	1.33×10^{-18}	75×10^{16}	
Rubber (hard)	$10^{-13} - 10^{-16}$	$10^{13} - 10^{16}$	
Sulfur	10^{-15}	10^{15}	
Teflon^{TM}	$< 10^{-13}$	$> 10^{13}$	
Wood	$10^{-8} - 10^{-11}$	$10^{8} - 10^{11}$	

The materials listed in the table are separated into categories of conductors, semiconductors, and insulators, based on broad groupings of resistivity. Conductors have the smallest resistivity, and insulators have the largest; semiconductors have intermediate resistivity. Conductors have varying but large, free charge densities, whereas most charges in insulators are bound to atoms and are not free to move. Semiconductors are intermediate, having far fewer free charges than conductors, but having properties that make the number of free charges depend strongly on the type and amount of impurities in the semiconductor. These unique properties of semiconductors are put to use in modern electronics, as we will explore in later chapters.

Example 5.4.1: Current Density, Resistance, and Electrical field for a Current-Carrying Wire

Calculate the current density, resistance, and electrical field of a 5-m length of copper wire with a diameter of 2.053 mm (12-gauge) carrying a current of $I - 10\,mA$.

Strategy

We can calculate the current density by first finding the cross-sectional area of the wire, which is $A = 3.31\,mm^2$, and the definition of current density $J = \frac{I}{A}$. The resistance can be found using the length of the wire $L = 5.00\,m$, the area, and the resistivity of copper $\rho = 1.68 \times 10^{-8}\,\Omega \cdot m$, where $R = \rho\frac{L}{A}$. The resistivity and current density can be used to find the electrical field.

Solution

First, we calculate the current density:

$$\begin{aligned} J &= \frac{I}{A} \\ &= \frac{10 \times 10^{-3}\,A}{3.31 \times 10^{-6}\,m^2} \\ &= 3.02 \times 10^{3}\,\frac{A}{m^2}. \end{aligned}$$

The resistance of the wire is

$$\begin{aligned} R &= \rho\frac{L}{A} \\ &= (1.68 \times 10^{-8}\,\Omega \cdot m)\frac{5.00\,m}{3.31 \times 10^{-6}\,m^2} \\ &= 0.025\,\Omega. \end{aligned}$$

Finally, we can find the electrical field:

$$\begin{aligned} E &= \rho J \\ &= 1.68 \times 10^{-8}\,\Omega \cdot m\left(3.02 \times 10^{3}\,\frac{A}{m^2}\right) \\ &= 5.07 \times 10^{-5}\,\frac{V}{m}. \end{aligned}$$

Significance

From these results, it is not surprising that copper is used for wires for carrying current because the resistance is quite small. Note that the current density and electrical field are independent of the length of the wire, but the voltage depends on the length.

Exercise 5.4.1

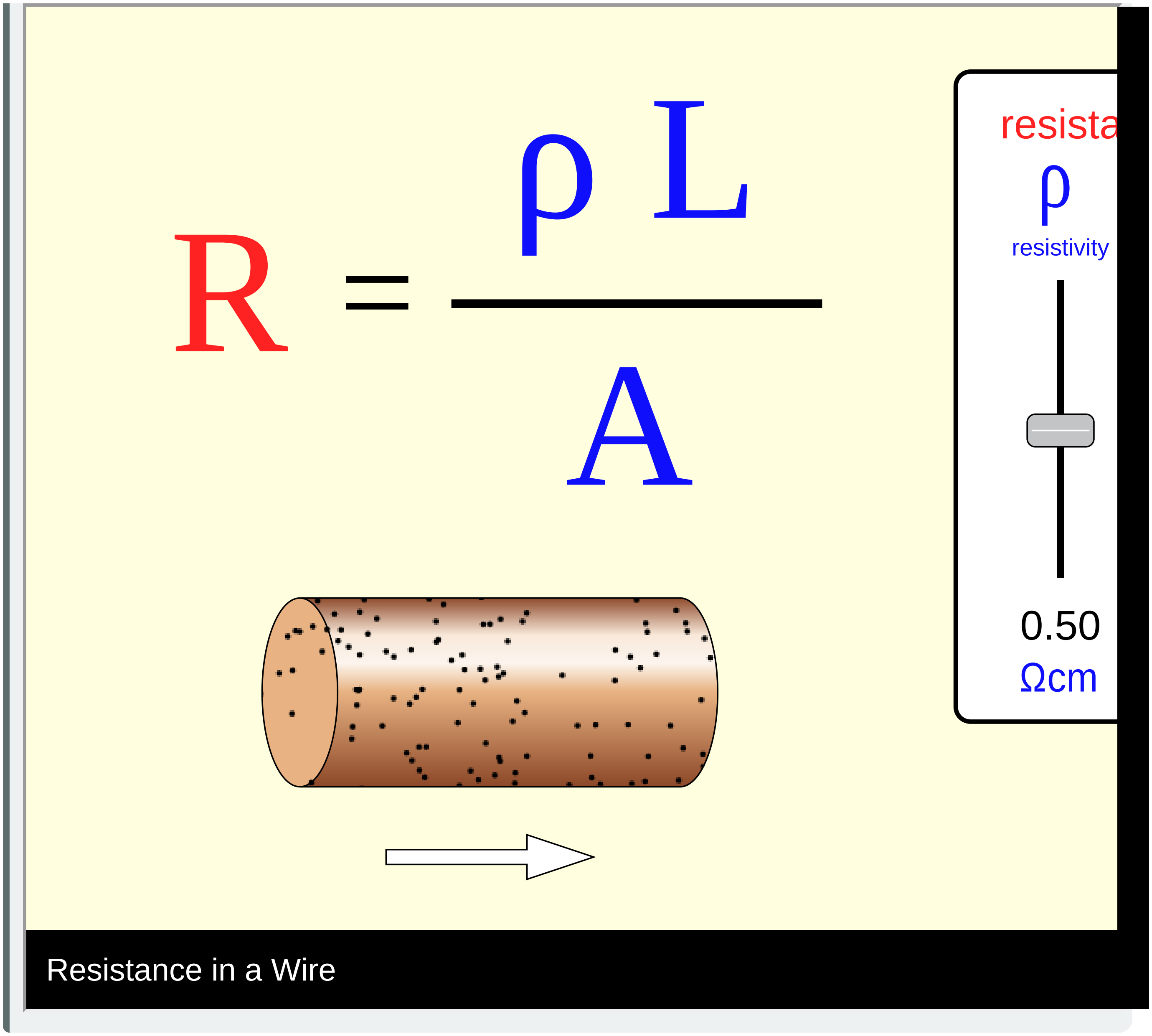
Copper wires use routinely used for extension cords and house wiring for several reasons. Copper has the highest electrical conductivity rating, and therefore the lowest resistivity rating, of all nonprecious metals. Also important is the tensile strength, where the tensile strength is a measure of the force required to pull an object to the point where it breaks. The tensile strength of a material is the maximum amount of tensile stress it can take before breaking. Copper has a high tensile strength, $2 \times 10^{8}\,\frac{N}{m^2}$. A third important characteristic is ductility. Ductility is a measure of a material's ability to be drawn into wires and a measure of the flexibility of the material, and copper has a high ductility. Summarizing, for a conductor to be a suitable candidate for making wire, there are at least three important characteristics: low resistivity, high tensile strength, and high ductility. What other materials are used for wiring and what are the advantages and disadvantages?

Answer

Silver, gold, and aluminum are all used for making wires. All four materials have a high conductivity, silver having the highest. All four can easily be drawn into wires and have a high tensile strength, though not as high as copper. The obvious disadvantage of gold and silver is the cost, but silver and gold wires are used for special applications, such as speaker wires. Gold does not oxidize, making better connections between components. Aluminum wires do have their drawbacks. Aluminum has a higher resistivity than copper, so a larger diameter is needed to match the resistance per length of copper wires, but aluminum is cheaper than copper, so this is not a major drawback. Aluminum wires do not have as high of a ductility and tensile strength as copper, but the ductility and tensile strength is within acceptable levels. There are a few concerns that must be addressed in using aluminum and care must be used when making connections. Aluminum has a higher rate of thermal expansion than copper, which can lead to loose connections and a possible fire hazard. The oxidation of aluminum does not conduct and can cause problems. Special techniques must be used when using aluminum wires and components, such as electrical outlets, must be designed to accept aluminum wires.

PhET

View this interactive simulation to see what the effects of the cross-sectional area, the length, and the resistivity of a wire are on the resistance of a conductor. Adjust the variables using slide bars and see if the resistance becomes smaller or larger.



Temperature Dependence of Resistivity

Looking back at Table 5.4.1, you will see a column labeled "Temperature Coefficient." The resistivity of some materials has a strong temperature dependence. In some materials, such as copper, the resistivity increases with increasing temperature. In fact, in most conducting metals, the resistivity increases with increasing temperature. The increasing temperature causes increased vibrations of the atoms in the lattice structure of the metals, which impede the motion of the electrons. In other materials, such as carbon, the resistivity decreases with increasing temperature. In many materials, the dependence is approximately linear and can be modeled using a linear equation:

$$\rho \approx \rho_0[1+\alpha(T-T_0)],$$

where ρ is the resistivity of the material at temperature **T**, α is the temperature coefficient of the material, and ρ_0 is the resistivity at T_0, usually taken as $T_0 = 20.00^oC$.

Note also that the temperature coefficient α is negative for the semiconductors listed in Table 5.4.1, meaning that their resistivity decreases with increasing temperature. They become better conductors at higher temperature, because increased thermal agitation increases the number of free charges available to carry current. This property of decreasing ρ with temperature is also related to the type and amount of impurities present in the semiconductors.

Resistance

We now consider the resistance of a wire or component. The resistance is a measure of how difficult it is to pass current through a wire or component. Resistance depends on the resistivity. The resistivity is a characteristic of the material used to fabricate a wire or other electrical component, whereas the resistance is a characteristic of the wire or component.

To calculate the resistance, consider a section of conducting wire with cross-sectional area **A**, length **L**, and resistivity ρ. A battery is connected across the conductor, providing a potential difference ΔV across it (Figure $5.4.1$). The potential difference produces an electrical field that is proportional to the current density, according to $\vec{E} = \rho \vec{J}$.

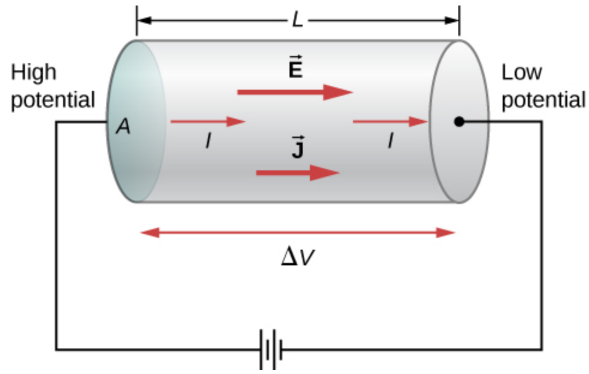


Figure $5.4.1$: A potential provided by a battery is applied to a segment of a conductor with a cross-sectional area A and a length L.

The magnitude of the electrical field across the segment of the conductor is equal to the voltage divided by the length, $E = V/L$, and the magnitude of the current density is equal to the current divided by the cross-sectional area, $J = I/A$. Using this information and recalling that the electrical field is proportional to the resistivity and the current density, we can see that the voltage is proportional to the current:

$$E = \rho J$$

$$\frac{V}{L} = \rho \frac{I}{A}$$

$$V = \left(\rho \frac{L}{A}\right) I.$$

> **Definition: Resistance**
>
> The ratio of the voltage to the current is defined as the **resistance** R:
>
> $$R \equiv \frac{V}{I}.$$
>
> The resistance of a cylindrical segment of a conductor is equal to the resistivity of the material times the length divided by the area:
>
> $$R \equiv \frac{V}{I} = \rho \frac{L}{A}.$$

The unit of resistance is the ohm, Ω. For a given voltage, the higher the resistance, the lower the current.

Resistors

A common component in electronic circuits is the resistor. The resistor can be used to reduce current flow or provide a voltage drop. Figure $5.4.2$ shows the symbols used for a resistor in schematic diagrams of a circuit. Two commonly used standards for circuit diagrams are provided by the American National Standard Institute (ANSI, pronounced “AN-see”) and the International Electrotechnical Commission (IEC). Both systems are commonly used. We use the ANSI standard in this text for its visual recognition, but we note that for larger, more complex circuits, the IEC standard may have a cleaner presentation, making it easier to read.

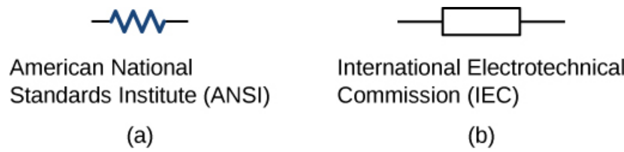


Figure 5.4.2: Symbols for a resistor used in circuit diagrams. (a) The ANSI symbol; (b) the IEC symbol.

Material and shape dependence of resistance

A resistor can be modeled as a cylinder with a cross-sectional area **A** and a length **L**, made of a material with a resistivity ρ (Figure 5.4.3). The resistance of the resistor is $R = \rho \frac{L}{A}$

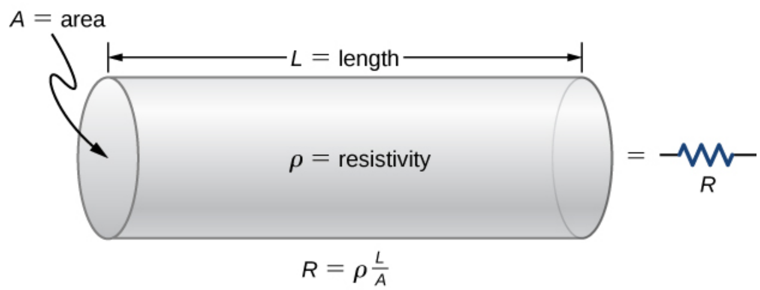
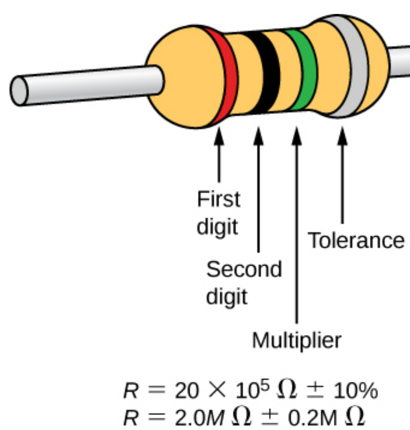


Figure 5.4.3: A model of a resistor as a uniform cylinder of length **L** and cross-sectional area **A**. Its resistance to the flow of current is analogous to the resistance posed by a pipe to fluid flow. The longer the cylinder, the greater its resistance. The larger its cross-sectional area **A**, the smaller its resistance.

The most common material used to make a resistor is carbon. A carbon track is wrapped around a ceramic core, and two copper leads are attached. A second type of resistor is the metal film resistor, which also has a ceramic core. The track is made from a metal oxide material, which has semiconductive properties similar to carbon. Again, copper leads are inserted into the ends of the resistor. The resistor is then painted and marked for identification. A resistor has four colored bands, as shown in Figure 5.4.4.



Color	Digit	Multiplier
Black	0	10^0
Brown	1	10^1
Red	2	10^2
Orange	3	10^3
Yellow	4	10^4
Green	5	10^5
Blue	6	10^6
Violet	7	10^7
Gray	8	10^8
White	9	10^9

Tolerance	
Gold	5%
Silver	10%

Figure 5.4.4: Many resistors resemble the figure shown above. The four bands are used to identify the resistor. The first two colored bands represent the first two digits of the resistance of the resistor. The third color is the multiplier. The fourth color represents the tolerance of the resistor. The resistor shown has a resistance of $20 \times 10^5\ \Omega \pm 10\%$

Resistances range over many orders of magnitude. Some ceramic insulators, such as those used to support power lines, have resistances of $10^{12}\ \Omega$ or more. A dry person may have a hand-to-foot resistance of $10^5\ \Omega$ whereas the resistance of the human heart is about $10^3\ \Omega$ A meter-long piece of large-diameter copper wire may have a resistance of $10^{-5}\ \Omega$, and superconductors have no resistance at all at low temperatures. As we have seen, resistance is related to the shape of an object and the material of which it is composed.

The resistance of an object also depends on temperature, since R_0 is directly proportional to ρ. For a cylinder, we know $R = \rho \frac{L}{A}$, so if **L** and **A** do not change greatly with temperature, **R** has the same temperature dependence as ρ. (Examination of the coefficients of linear expansion shows them to be about two orders of magnitude less than typical temperature coefficients of resistivity, so the effect of temperature on **L** and **A** is about two orders of magnitude less than on ρ.) Thus,

$$R = R_0(1+\alpha\Delta T) \tag{5.4.1}$$

is the temperature dependence of the resistance of an object, where R_0 is the original resistance (usually taken to be $T = 20.00^oC$ and **R** is the resistance after a temperature change ΔT. The color code gives the resistance of the resistor at a temperature of $T = 20.00^oC$.

Numerous thermometers are based on the effect of temperature on resistance (Figure $5.4.5$). One of the most common thermometers is based on the thermistor, a semiconductor crystal with a strong temperature dependence, the resistance of which is measured to obtain its temperature. The device is small, so that it quickly comes into thermal equilibrium with the part of a person it touches.

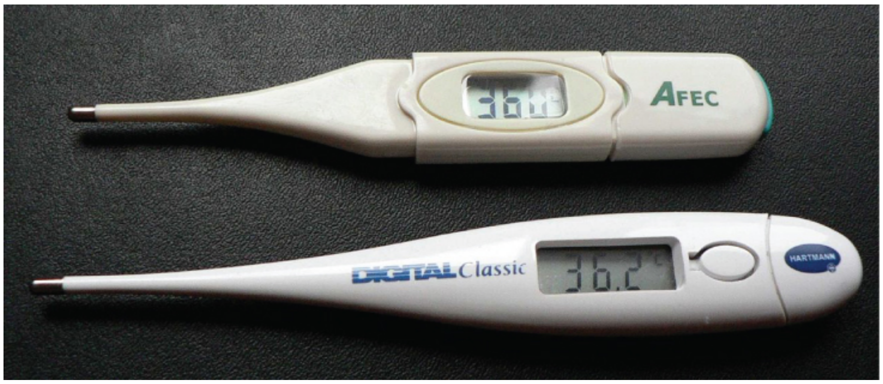


Figure $5.4.5$: These familiar thermometers are based on the automated measurement of a thermistor's temperature-dependent resistance.

Example $5.4.2$: Calculating Resistance

Although caution must be used in applying $\rho = \rho_0(1+\alpha\Delta T)$ and $R = R_0(1+\alpha\Delta T)$ for temperature changes greater than 100^oC, for tungsten, the equations work reasonably well for very large temperature changes. A tungsten filament at 20^oC has a resistance of $0.350\,\Omega$. What would the resistance be if the temperature is increased to 2850^oC?

Strategy

This is a straightforward application of Equation $5.4.1$, since the original resistance of the filament is given as $R_0 = 0.350\,\Omega$ and the temperature change is $\Delta T = 2830^oC$.

Solution

The resistance of the hotter filament R is obtained by entering known values into the above equation:

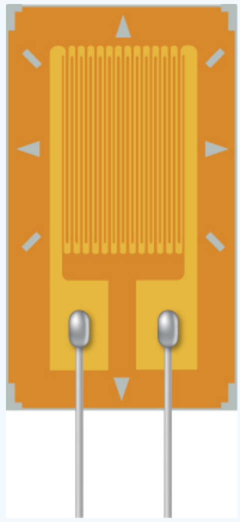
$$\begin{aligned} R &= R_0(1+\alpha\Delta T) \\ &= (0.350\,\Omega)\left(1+\left(\frac{4.5\times10^{-3}}{^oC}\right)(2830^oC)\right) \\ &= 4.8\,\Omega \end{aligned}$$

Significance

Notice that the resistance changes by more than a factor of 10 as the filament warms to the high temperature and the current through the filament depends on the resistance of the filament and the voltage applied. If the filament is used in an incandescent light bulb, the initial current through the filament when the bulb is first energized will be higher than the current after the filament reaches the operating temperature.

Exercise 5.4.2

A strain gauge is an electrical device to measure strain, as shown below. It consists of a flexible, insulating backing that supports a conduction foil pattern. The resistance of the foil changes as the backing is stretched. How does the strain gauge resistance change? Is the strain gauge affected by temperature changes?



Answer

The foil pattern stretches as the backing stretches, and the foil tracks become longer and thinner. Since the resistance is calculated as $R = \rho \frac{L}{A}$, the resistance increases as the foil tracks are stretched. When the temperature changes, so does the resistivity of the foil tracks, changing the resistance. One way to combat this is to use two strain gauges, one used as a reference and the other used to measure the strain. The two strain gauges are kept at a constant temperature

The Resistance of Coaxial Cable

Long cables can sometimes act like antennas, picking up electronic noise, which are signals from other equipment and appliances. Coaxial cables are used for many applications that require this noise to be eliminated. For example, they can be found in the home in cable TV connections or other audiovisual connections. Coaxial cables consist of an inner conductor of radius r_i surrounded by a second, outer concentric conductor with radius r_0 (Figure 5.4.6). The space between the two is normally filled with an insulator such as polyethylene plastic. A small amount of radial leakage current occurs between the two conductors. Determine the resistance of a coaxial cable of length **L**.

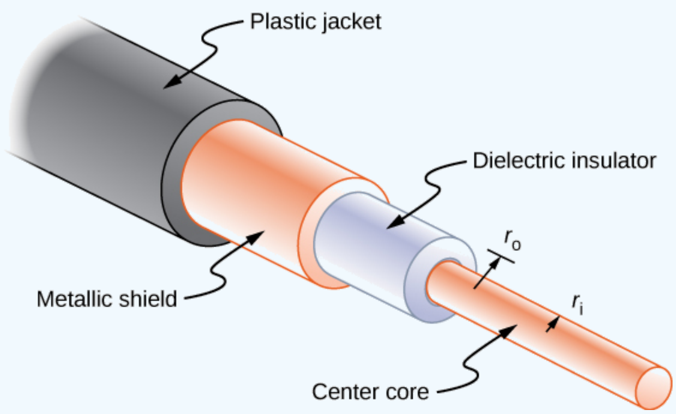


Figure 5.4.6: Coaxial cables consist of two concentric conductors separated by insulation. They are often used in cable TV or other audiovisual connections.

Strategy

We cannot use the equation $R = \rho \frac{L}{A}$ directly. Instead, we look at concentric cylindrical shells, with thickness **dr**, and integrate.

Solution

We first find an expression for dR and then integrate from r_i to r_0,

$$\begin{aligned} dR &= \frac{\rho}{A} dr \\ &= \frac{\rho}{2\pi r L} dr, \end{aligned}$$

Integrating both sides

$$\begin{aligned} R &= \int_{r_i}^{r_0} dR \\ &= \int_{r_i}^{r_0} \frac{\rho}{2\pi r L} dr \\ &= \frac{\rho}{2\pi L} \int_{r_i}^{r_0} \frac{1}{r} dr \\ &= \frac{\rho}{2\pi L} \ln \frac{r_0}{r_i}. \end{aligned}$$

Significance

The resistance of a coaxial cable depends on its length, the inner and outer radii, and the resistivity of the material separating the two conductors. Since this resistance is not infinite, a small leakage current occurs between the two conductors. This leakage current leads to the attenuation (or weakening) of the signal being sent through the cable.

Exercise 5.4.3

The resistance between the two conductors of a coaxial cable depends on the resistivity of the material separating the two conductors, the length of the cable and the inner and outer radius of the two conductor. If you are designing a coaxial cable, how does the resistance between the two conductors depend on these variables?

Answer

The longer the length, the smaller the resistance. The greater the resistivity, the higher the resistance. The larger the difference between the outer radius and the inner radius, that is, the greater the ratio between the two, the greater the resistance. If you are attempting to maximize the resistance, the choice of the values for these variables will depend on the application. For example, if the cable must be flexible, the choice of materials may be limited.

Phet: Battery-Resistor Circuit

View this simulation to see how the voltage applied and the resistance of the material the current flows through affects the current through the material. You can visualize the collisions of the electrons and the atoms of the material effect the temperature of the material.

-

5.5: Ohm's Law

Learning Objectives

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

- Describe Ohm’s law
- Recognize when Ohm’s law applies and when it does not

We have been discussing three electrical properties so far in this chapter: current, voltage, and resistance. It turns out that many materials exhibit a simple relationship among the values for these properties, known as Ohm’s law. Many other materials do not show this relationship, so despite being called Ohm’s law, it is not considered a law of nature, like Newton’s laws or the laws of thermodynamics. But it is very useful for calculations involving materials that do obey Ohm’s law.

Description of Ohm’s Law

The current that flows through most substances is directly proportional to the voltage V applied to it. The German physicist Georg Simon Ohm (1787–1854) was the first to demonstrate experimentally that the current in a metal wire is **directly proportional to the voltage applied**:

$$I \propto V.$$

This important relationship is the basis for **Ohm’s law**. It can be viewed as a cause-and-effect relationship, with voltage the cause and current the effect. This is an empirical law, which is to say that it is an experimentally observed phenomenon, like friction. Such a linear relationship doesn’t always occur. Any material, component, or device that obeys Ohm’s law, where the current through the device is proportional to the voltage applied, is known as an ohmic material or **ohmic** component. Any material or component that does not obey Ohm’s law is known as a **nonohmic** material or nonohmic component.

Ohm’s Experiment

In a paper published in 1827, Georg Ohm described an experiment in which he measured voltage across and current through various simple electrical circuits containing various lengths of wire. A similar experiment is shown in Figure $5.5.1$. This experiment is used to observe the current through a resistor that results from an applied voltage. In this simple circuit, a resistor is connected in series with a battery. The voltage is measured with a voltmeter, which must be placed across the resistor (in parallel with the resistor). The current is measured with an ammeter, which must be in line with the resistor (in series with the resistor).

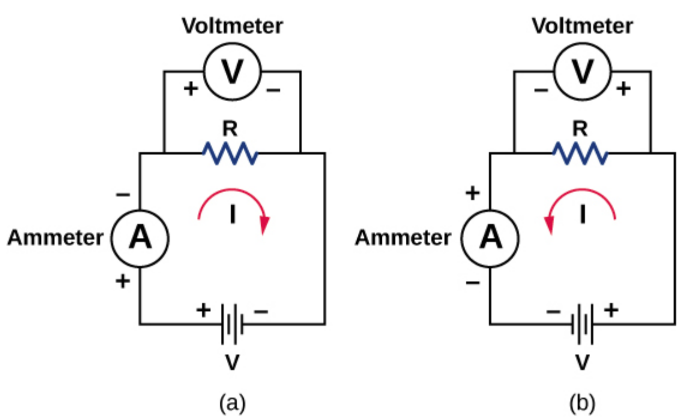


Figure $5.5.1$: The experimental set-up used to determine if a resistor is an ohmic or nonohmic device. (a) When the battery is attached, the current flows in the clockwise direction and the voltmeter and ammeter have positive readings. (b) When the leads of the battery are switched, the current flows in the counterclockwise direction and the voltmeter and ammeter have negative readings.

In this updated version of Ohm’s original experiment, several measurements of the current were made for several different voltages. When the battery was hooked up as in Figure $5.5.1a$, the current flowed in the clockwise direction and the readings of the voltmeter and ammeter were positive. Does the behavior of the current change if the current flowed in the opposite direction? To get the current to flow in the opposite direction, the leads of the battery can be switched. When the leads of the battery were switched, the readings of the voltmeter and ammeter readings were negative because the current flowed in the opposite direction, in this case, counterclockwise. Results of a similar experiment are shown in Figure 5.5.2.

I(A)	V(V)
−2.74	−10.00
−2.25	−9.00
−2.00	−8.00
−1.75	−7.00
−1.50	−6.00
−1.49	−5.00
−1.00	−4.00
−0.51	−3.00
−0.74	−2.00
−0.49	−1.00
+0.00	+0.00
+0.49	+1.00
+0.50	+2.00
+0.99	+3.00
+0.76	+4.00
+1.01	+5.00
+1.74	+6.00
+1.75	+7.00
+2.00	+8.00
+2.49	+9.00
+2.50	+10.00

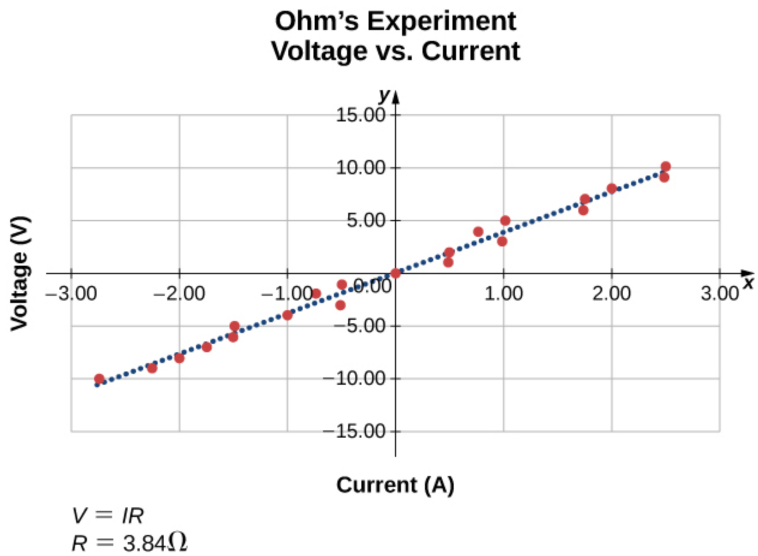


Figure 5.5.1: A resistor is placed in a circuit with a battery. The voltage applied varies from −10.00 V to +10.00 V, increased by 1.00-V increments. A plot shows values of the voltage versus the current typical of what a casual experimenter might find.

In this experiment, the voltage applied across the resistor varies from −10.00 to +10.00 V, by increments of 1.00 V. The current through the resistor and the voltage across the resistor are measured. A plot is made of the voltage versus the current, and the result is approximately linear. The slope of the line is the resistance, or the voltage divided by the current. This result is known as **Ohm's law**:

$$V = IR \tag{5.5.1}$$

where V is the voltage measured in volts across the object in question, I is the current measured through the object in amps, and R is the resistance in units of ohms. As stated previously, any device that shows a linear relationship between the voltage and the current is known as an ohmic device. A resistor is therefore an ohmic device.

Example 5.5.1: Measuring Resistance

A carbon resistor at room temperature (20^oC) is attached to a 9.00-V battery and the current measured through the resistor is 3.00 mA.

a. What is the resistance of the resistor measured in ohms?
b. If the temperature of the resistor is increased to 60^oC by heating the resistor, what is the current through the resistor?

Strategy

a. The resistance can be found using Ohm's law. Ohm's law states that $V = IR$, so the resistance can be found using $R = V/I$.
b. First, the resistance is temperature dependent so the new resistance after the resistor has been heated can be found using $R = R_0(1 + \alpha\Delta T)$. The current can be found using Ohm's law in the form $I = V/R$.

Solution

a. Using Ohm's law and solving for the resistance yields the resistance at room temperature:

$$R = \frac{V}{I} = \frac{9.00\,V}{3.00\times 10^{-3}\,A} = 3.00\times 10^3\ \Omega = 3.00 k\Omega$$

b. The resistance at 60^oC can be found using $R = R_0(1 + \alpha\Delta T)$ where the temperature coefficient for carbon is $\alpha = -0.0005$.

$$R = R_0(1 + \alpha\Delta T) = 3.00\times 10^3(1 - 0.0005(60^oC - 20^oC)) = 2.94\ k\Omega.$$

The current through the heated resistor is

$$I = \frac{V}{R} = \frac{9.00\,V}{2.94 \times 10^3\,\Omega} = 3.06 \times 10^{-3}\,A = 3.06\,mA.$$

Significance

A change in temperature of 40^oC resulted in a 2.00% change in current. This may not seem like a very great change, but changing electrical characteristics can have a strong effect on the circuits. For this reason, many electronic appliances, such as computers, contain fans to remove the heat dissipated by components in the electric circuits.

Exercise 5.5.1

The voltage supplied to your house varies as $V(t) = V_{max}\,sin\,(2\pi\,ft)$. If a resistor is connected across this voltage, will Ohm's law $V = IR$ still be valid?

Answer

Yes, Ohm's law is still valid. At every point in time the current is equal to $I(t) = V(t)/R$, so the current is also a function of time, $I(t) = \frac{V_{max}}{R}\,sin\,(2\pi\,ft)$.

Simulation: PhET

See how Ohm's law (Equation 5.5.1) relates to a simple circuit. Adjust the voltage and resistance, and see the current change according to Ohm's law. The sizes of the symbols in the equation change to match the circuit diagram.

Nonohmic devices do not exhibit a linear relationship between the voltage and the current. One such device is the semiconducting circuit element known as a **diode**. A diode is a circuit device that allows current flow in only one direction. A diagram of a simple circuit consisting of a battery, a diode, and a resistor is shown in Figure 5.5.3. Although we do not cover the theory of the diode in this section, the diode can be tested to see if it is an ohmic or a nonohmic device.

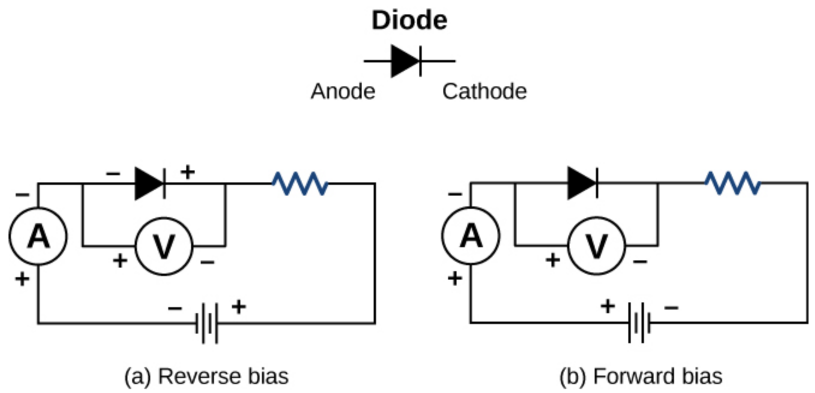


Figure 5.5.3: A diode is a semiconducting device that allows current flow only if the diode is forward biased, which means that the anode is positive and the cathode is negative.

A plot of current versus voltage is shown in Figure 5.5.4. Note that the behavior of the diode is shown as current versus voltage, whereas the resistor operation was shown as voltage versus current. A diode consists of an anode and a cathode. When the anode is at a negative potential and the cathode is at a positive potential, as shown in part (a), the diode is said to have reverse bias. With reverse bias, the diode has an extremely large resistance and there is very little current flow—essentially zero current—through the diode and the resistor. As the voltage applied to the circuit increases, the current remains essentially zero, until the voltage reaches the breakdown voltage and the diode conducts current. When the battery and the potential across the diode are reversed, making the anode positive and the cathode negative, the diode conducts and current flows through the diode if the voltage is greater than 0.7 V. The resistance of the diode is close to zero. (This is the reason for the resistor in the circuit; if it were not there, the current would become very large.) You can see from the graph in Figure 5.5.4 that the voltage and the current do not have a linear relationship. Thus, the diode is an example of a nonohmic device.

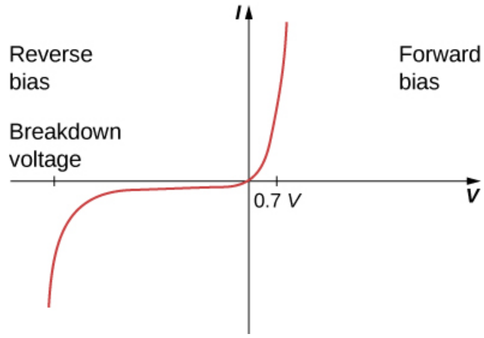


Figure 5.5.4: When the voltage across the diode is negative and small, there is very little current flow through the diode. As the voltage reaches the breakdown voltage, the diode conducts. When the voltage across the diode is positive and greater than 0.7 V (the actual voltage value depends on the diode), the diode conducts. As the voltage applied increases, the current through the diode increases, but the voltage across the diode remains approximately 0.7 V.

Ohm's law is commonly stated as $V = IR$, but originally it was stated as a microscopic view, in terms of the current density, the conductivity, and the electrical field. This microscopic view suggests the proportionality $V \propto I$ comes from the drift velocity of the free electrons in the metal that results from an applied electrical field. As stated earlier, the current density is proportional to the applied electrical field. The reformulation of Ohm's law is credited to Gustav Kirchhoff, whose name we will see again in the next chapter.

5.6: Electrical Energy and Power

LEARNING OBJECTIVES

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

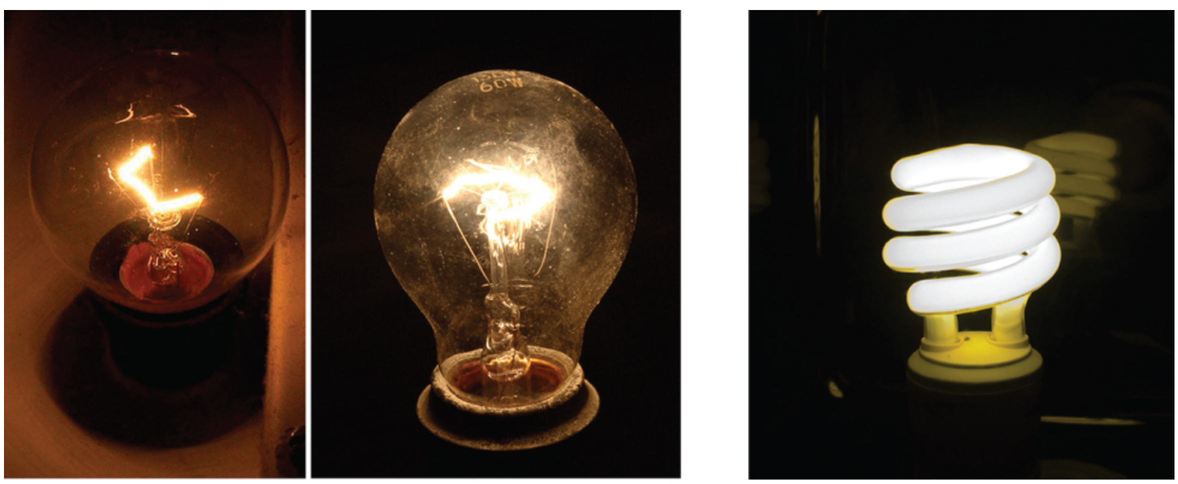
- Express electrical power in terms of the voltage and the current
- Describe the power dissipated by a resistor in an electric circuit
- Calculate the energy efficiency and cost effectiveness of appliances and equipment

In an electric circuit, electrical energy is continuously converted into other forms of energy. For example, when a current flows in a conductor, electrical energy is converted into thermal energy within the conductor. The electrical field, supplied by the voltage source, accelerates the free electrons, increasing their kinetic energy for a short time. This increased kinetic energy is converted into thermal energy through collisions with the ions of the lattice structure of the conductor. Previously, we defined power as the rate at which work is done by a force measured in watts. Power can also be defined as the rate at which energy is transferred. In this section, we discuss the time rate of energy transfer, or power, in an electric circuit.

Power in Electric Circuits

Power is associated by many people with electricity. Power transmission lines might come to mind. We also think of light bulbs in terms of their power ratings in watts. What is the expression for **electric power**?

Let us compare a 25-W bulb with a 60-W bulb (Figure $5.6.1a$). The 60-W bulb glows brighter than the 25-W bulb. Although it is not shown, a 60-W light bulb is also warmer than the 25-W bulb. The heat and light is produced by from the conversion of electrical energy. The kinetic energy lost by the electrons in collisions is converted into the internal energy of the conductor and radiation. How are voltage, current, and resistance related to electric power?



(a) (b)

Figure $5.6.1$: (a) Pictured above are two incandescent bulbs: a 25-W bulb (left) and a 60-W bulb (right). The 60-W bulb provides a higher intensity light than the 25-W bulb. The electrical energy supplied to the light bulbs is converted into heat and light. (b) This compact fluorescent light (CFL) bulb puts out the same intensity of light as the 60-W bulb, but at 1/4 to 1/10 the input power. (credit a: modification of works by "Dickbauch"/Wikimedia Commons and Greg Westfall; credit b: modification of work by "dbgg1979"/Flickr)

To calculate electric power, consider a voltage difference existing across a material (Figure $5.6.2$). The electric potential V_1 is higher than the electric potential at V_2, and the voltage difference is negative $V = V_2 - V_1$. As discussed in Electric Potential, an electrical field exists between the two potentials, which points from the higher potential to the lower potential. Recall that the electrical potential is defined as the potential energy per charge, $V = \Delta U/q$, and the charge ΔQ loses potential energy moving through the potential difference.

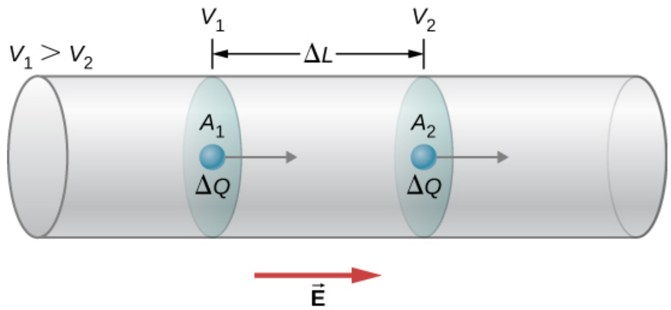


Figure 5.6.2: When there is a potential difference across a conductor, an electrical field is present that points in the direction from the higher potential to the lower potential.

If the charge is positive, the charge experiences a force due to the electrical field $\vec{F} = m\vec{a} = \Delta Q\vec{E}$. This force is necessary to keep the charge moving. This force does not act to accelerate the charge through the entire distance ΔL because of the interactions of the charge with atoms and free electrons in the material. The speed, and therefore the kinetic energy, of the charge do not increase during the entire trip across ΔL, and charge passing through area A_2 has the same drift velocity v_d as the charge that passes through area A_1. However, work is done on the charge, by the electrical field, which changes the potential energy. Since the change in the electrical potential difference is negative, the electrical field is found to be

$$E = -\frac{(V_2 - V_1)}{\Delta L} = \frac{V}{\Delta L}.$$

The work done on the charge is equal to the electric force times the length at which the force is applied,

$$W = F\Delta L = (\Delta QE)\Delta L = \left(\Delta Q\frac{V}{\Delta L}\right)\Delta L = \Delta QV = \Delta U.$$

The charge moves at a drift velocity v_d so the work done on the charge results in a loss of potential energy, but the average kinetic energy remains constant. The lost electrical potential energy appears as thermal energy in the material. On a microscopic scale, the energy transfer is due to collisions between the charge and the molecules of the material, which leads to an increase in temperature in the material. The loss of potential energy results in an increase in the temperature of the material, which is dissipated as radiation. In a resistor, it is dissipated as heat, and in a light bulb, it is dissipated as heat and light.

The power dissipated by the material as heat and light is equal to the time rate of change of the work:

$$P = IV = I(IR) = I^2R$$

or

$$P = IV = \left(\frac{V}{R}\right)V = \frac{V^2}{R}.$$

If a resistor is connected to a battery, the power dissipated as radiant energy by the wires and the resistor is equal to

$$P = IV = I^2R = \frac{V^2}{R}.$$

The power supplied from the battery is equal to current times the voltage, $P = IV$.

Definition: Electric Power

The electric power gained or lost by any device has the form

$$P = IV.$$

The power dissipated by a resistor has the form

$$P = I^2R = \frac{V^2}{R}.$$

Different insights can be gained from the three different expressions for electric power. For example, $P = V^2/R$ implies that the lower the resistance connected to a given voltage source, the greater the power delivered. Furthermore, since voltage is squared in $P = V^2/R$, the effect of applying a higher voltage is perhaps greater than expected. Thus, when the voltage is doubled to a 25-W bulb, its power nearly quadruples to about 100 W, burning it out. If the bulb's resistance remained constant, its power would be exactly 100 W, but at the higher temperature, its resistance is higher, too.

Example 5.6.1: Calculating Power in Electric Devices

A DC winch motor is rated at 20.00 A with a voltage of 115 V. When the motor is running at its maximum power, it can lift an object with a weight of 4900.00 N a distance of 10.00 m, in 30.00 s, at a constant speed.

a. What is the power consumed by the motor?
b. What is the power used in lifting the object? Ignore air resistance. (c) Assuming that the difference in the power consumed by the motor and the power used lifting the object are dissipated as heat by the resistance of the motor, estimate the resistance of the motor?

Strategy

a. The power consumed by the motor can be found using $P = IV$.
b. The power used in lifting the object at a constant speed can be found using $P = Fv$, where the speed is the distance divided by the time. The upward force supplied by the motor is equal to the weight of the object because the acceleration is zero. (c) The resistance of the motor can be found using $P = I^2R$.

Solution

1. The power consumed by the motor is equal to $P = IV$ and the current is given as 20.00 A and the voltage is 115.00 V:
$$P = IV = (20.00\ A)115.00\ V = 2300.00\ W.$$
2. The power used lifting the object is equal to $P = Fv$ where the force is equal to the weight of the object (1960 N) and the magnitude of the velocity is
$$v = \frac{10.00\ m}{30.00\ s} = 0.33\frac{m}{s}$$
$$P = Fv = (4900\ N)0.33\ m/s = 1633.33\ W.$$
3. The difference in the power equals $2300.00\ W - 1633.33\ W = 666.67\ W$and the resistance can be found using $P = I^2R$:
$$R = \frac{P}{I^2} = \frac{666.67\ W}{(20.00\ A)^2} = 1.67\ \Omega.$$

Significance The resistance of the motor is quite small. The resistance of the motor is due to many windings of copper wire. The power dissipated by the motor can be significant since the thermal power dissipated by the motor is proportional to the square of the current $(P = I^2R)$.

Exercise 5.6.1

Electric motors have a reasonably high efficiency. A 100-hp motor can have an efficiency of 90% and a 1-hp motor can have an efficiency of 80%. Why is it important to use high-performance motors?

Answer

Even though electric motors are highly efficient 10–20% of the power consumed is wasted, not being used for doing useful work. Most of the 10–20% of the power lost is transferred into heat dissipated by the copper wires used to make the coils of the motor. This heat adds to the heat of the environment and adds to the demand on power plants providing the power. The demand on the power plant can lead to increased greenhouse gases, particularly if the power plant uses coal or gas as fuel.

A fuse (Figure 5.6.3) is a device that protects a circuit from currents that are too high. A fuse is basically a short piece of wire between two contacts. As we have seen, when a current is running through a conductor, the kinetic energy of the charge carriers is converted into thermal energy in the conductor. The piece of wire in the fuse is under tension and has a low melting point. The wire is designed to heat up and break at the rated current. The fuse is destroyed and must be replaced, but it protects the rest of the circuit. Fuses act quickly, but there is a small time delay while the wire heats up and breaks.

Figure $5.6.3$: A fuse consists of a piece of wire between two contacts. When a current passes through the wire that is greater than the rated current, the wire melts, breaking the connection. Pictured is a "blown" fuse where the wire broke protecting a circuit (credit: modification of work by "Shardayyy"/Flickr).

Circuit breakers are also rated for a maximum current, and open to protect the circuit, but can be reset. Circuit breakers react much faster. The operation of circuit breakers is not within the scope of this chapter and will be discussed in later chapters. Another method of protecting equipment and people is the ground fault circuit interrupter (GFCI), which is common in bathrooms and kitchens. The GFCI outlets respond very quickly to changes in current. These outlets open when there is a change in magnetic field produced by current-carrying conductors, which is also beyond the scope of this chapter and is covered in a later chapter.

The Cost of Electricity

The more electric appliances you use and the longer they are left on, the higher your electric bill. This familiar fact is based on the relationship between energy and power. You pay for the energy used. Since $P = \frac{dE}{dt}$, we see that

$$E = \int Pdt$$

is the energy used by a device using power **P** for a time interval **t**. If power is delivered at a constant rate, then then the energy can be found by $E = Pt$. For example, the more light bulbs burning, the greater **P** used; the longer they are on, the greater **t** is.

The energy unit on electric bills is the kilowatt-hour $(kW \cdot h)$, consistent with the relationship $E = Pt$. It is easy to estimate the cost of operating electrical appliances if you have some idea of their power consumption rate in watts or kilowatts, the time they are on in hours, and the cost per kilowatt-hour for your electric utility. Kilowatt-hours, like all other specialized energy units such as food calories, can be converted into joules. You can prove to yourself that $1\, kW \cdot h = 3.6 \times 10^6\, J$.

The electrical energy (**E**) used can be reduced either by reducing the time of use or by reducing the power consumption of that appliance or fixture. This not only reduces the cost but also results in a reduced impact on the environment. Improvements to lighting are some of the fastest ways to reduce the electrical energy used in a home or business. About 20% of a home's use of energy goes to lighting, and the number for commercial establishments is closer to 40%. Fluorescent lights are about four times more efficient than incandescent lights—this is true for both the long tubes and the compact fluorescent lights (CFLs), e.g., Figure $5.6.1b$. Thus, a 60-W incandescent bulb can be replaced by a 15-W CFL, which has the same brightness and color. CFLs have a bent tube inside a globe or a spiral-shaped tube, all connected to a standard screw-in base that fits standard incandescent light sockets. (Original problems with color, flicker, shape, and high initial investment for CFLs have been addressed in recent years.)

The heat transfer from these CFLs is less, and they last up to 10 times longer than incandescent bulbs. The significance of an investment in such bulbs is addressed in the next example. New white LED lights (which are clusters of small LED bulbs) are even more efficient (twice that of CFLs) and last five times longer than CFLs.

Example \(\PageIndex{2}\): Calculating the Cost Effectiveness of LED Bulb

The typical replacement for a 100-W incandescent bulb is a 20-W LED bulb. The 20-W LED bulb can provide the same amount of light output as the 100-W incandescent light bulb. What is the cost savings for using the LED bulb in place of the incandescent bulb for one year, assuming $0.10 per kilowatt-hour is the average energy rate charged by the power company? Assume that the bulb is turned on for three hours a day.

Strategy

a. Calculate the energy used during the year for each bulb, using $E = Pt$.
b. Multiply the energy by the cost.

Solution

a. Calculate the power for each bulb.

$$E_{Incandescent} = Pt = 100\,W\left(\frac{1\,kW}{1000\,W}\right)\left(\frac{3\,h}{day}\right)(365\,days) = 109.5\,kW \cdot h$$

$$E_{LED} = Pt = 20\,W\left(\frac{1\,kW}{1000\,W}\right)\left(\frac{3\,h}{day}\right)(365\,days) = 21.9\,kW \cdot h$$

b. Calculate the cost for each.

$$cost_{Incandescent} = 109.5\,kW \cdot h\left(\frac{\$0.10}{kW \cdot h}\right) = \$10.95$$

$$cost_{LED} = 21.90\,kW \cdot h\left(\frac{\$0.10}{kW \cdot h}\right) = \$2.19$$

Significance

A LED bulb uses 80% less energy than the incandescent bulb, saving $8.76 over the incandescent bulb for one year. The LED bulb can cost $20.00 and the 100-W incandescent bulb can cost $0.75, which should be calculated into the computation. A typical lifespan of an incandescent bulb is 1200 hours and is 50,000 hours for the LED bulb. The incandescent bulb would last 1.08 years at 3 hours a day and the LED bulb would last 45.66 years. The initial cost of the LED bulb is high, but the cost to the home owner will be $0.69 for the incandescent bulbs versus $0.44 for the LED bulbs per year. (Note that the LED bulbs are coming down in price.) The cost savings per year is approximately $8.50, and that is just for one bulb.

Exercise 5.6.2

Is the efficiency of the various light bulbs the only consideration when comparing the various light bulbs?

Answer

No, the efficiency is a very important consideration of the light bulbs, but there are many other considerations. As mentioned above, the cost of the bulbs and the life span of the bulbs are important considerations. For example, CFL bulbs contain mercury, a neurotoxin, and must be disposed of as hazardous waste. When replacing incandescent bulbs that are being controlled by a dimmer switch with LED, the dimmer switch may need to be replaced. The dimmer switches for LED lights are comparably priced to the incandescent light switches, but this is an initial cost which should be considered. The spectrum of light should also be considered, but there is a broad range of color temperatures available, so you should be able to find one that fits your needs. None of these considerations mentioned are meant to discourage the use of LED or CFL light bulbs, but they are considerations.

Changing light bulbs from incandescent bulbs to CFL or LED bulbs is a simple way to reduce energy consumption in homes and commercial sites. CFL bulbs operate with a much different mechanism than do incandescent lights. The mechanism is complex and beyond the scope of this chapter, but here is a very general description of the mechanism. CFL bulbs contain argon and mercury vapor housed within a spiral-shaped tube. The CFL bulbs use a “ballast” that increases the voltage used by the CFL bulb. The ballast produce an electrical current, which passes through the gas mixture and excites the gas molecules. The excited gas molecules produce ultraviolet (UV) light, which in turn stimulates the fluorescent coating on the inside of the tube. This coating fluoresces in the visible spectrum, emitting visible light. Traditional fluorescent tubes and CFL bulbs had a short time delay of up to a few seconds while the mixture was being “warmed up” and the molecules reached an excited state. It should be noted that these bulbs do contain mercury, which is poisonous, but if the bulb is broken, the mercury is never released. Even if the bulb is broken, the mercury tends to remain in the fluorescent coating. The amount is also quite small and the advantage of the energy saving may outweigh the disadvantage of using mercury.

The CFL light bulbs are being replaced with LED light bulbs, where LED stands for "light-emitting diode." The diode was briefly discussed as a nonohmic device, made of semiconducting material, which essentially permits current flow in one direction. LEDs are a special type of diode made of semiconducting materials infused with impurities in combinations and concentrations that enable the extra energy from the movement of the electrons during electrical excitation to be converted into visible light. Semiconducting devices will be explained in greater detail in Condensed Matter Physics.

Commercial LEDs are quickly becoming the standard for commercial and residential lighting, replacing incandescent and CFL bulbs. They are designed for the visible spectrum and are constructed from gallium doped with arsenic and phosphorous atoms. The color emitted from an LED depends on the materials used in the semiconductor and the current. In the early years of LED development, small LEDs found on circuit boards were red, green, and yellow, but LED light bulbs can now be programmed to produce millions of colors of light as well as many different hues of white light.

Comparison of Incandescent, CFL, and LED Light Bulbs

The energy savings can be significant when replacing an incandescent light bulb or a CFL light bulb with an LED light. Light bulbs are rated by the amount of power that the bulb consumes, and the amount of light output is measured in lumens. The lumen (lm) is the SI -derived unit of luminous flux and is a measure of the total quantity of visible light emitted by a source. A 60-W incandescent light bulb can be replaced with a 13- to 15-W CFL bulb or a 6- to 8-W LED bulb, all three of which have a light output of approximately 800 lm. A table of light output for some commonly used light bulbs appears in Table $5.6.1$.

The life spans of the three types of bulbs are significantly different. An LED bulb has a life span of 50,000 hours, whereas the CFL has a lifespan of 8000 hours and the incandescent lasts a mere 1200 hours. The LED bulb is the most durable, easily withstanding rough treatment such as jarring and bumping. The incandescent light bulb has little tolerance to the same treatment since the filament and glass can easily break. The CFL bulb is also less durable than the LED bulb because of its glass construction. The amount of heat emitted is 3.4 btu/h for the 8-W LED bulb, 85 btu/h for the 60-W incandescent bulb, and 30 btu/h for the CFL bulb. As mentioned earlier, a major drawback of the CFL bulb is that it contains mercury, a neurotoxin, and must be disposed of as hazardous waste. From these data, it is easy to understand why the LED light bulb is quickly becoming the standard in lighting.

Table $5.6.1$: Light Output of LED, Incandescent, and CFL Light Bulbs

Light Output (lumens)	LED Light Bulbs (watts)	Incandescent Light Bulbs (watts)	CFL Light Bulbs (watts)
450	4−5	40	9−13
800	6−8	60	13−15
1100	9−13	75	18−25
1600	16−20	100	23−30
2600	25−28	150	30−55

Summary of Relationships

In this chapter, we have discussed relationships between voltages, current, resistance, and power. Figure $5.6.4$ shows a summary of the relationships between these measurable quantities for ohmic devices. (Recall that ohmic devices follow Ohm's law $V = IR$.) For example, if you need to calculate the power, use the pink section, which shows that $P = VI$, $P = \frac{V^2}{R}$, and $P = I^2R$.

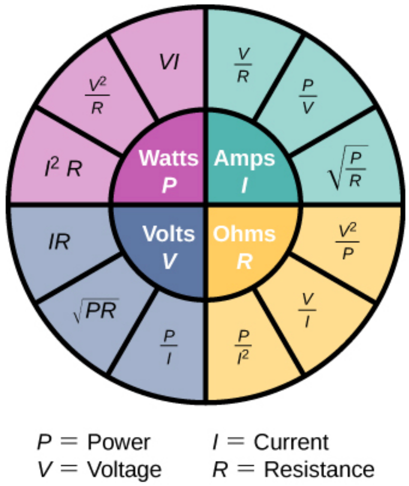


Figure 5.6.4: This circle shows a summary of the equations for the relationships between power, current, voltage, and resistance.

Which equation you use depends on what values you are given, or you measure. For example if you are given the current and the resistance, use $P = I^2R$. Although all the possible combinations may seem overwhelming, don't forget that they all are combinations of just two equations, Ohm's law $(V = IR)$ and power $(P = IV)$.

5.7: Alternating Current versus Direct Current

Learning Objectives

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

- Explain the differences and similarities between AC and DC current.
- Calculate rms voltage, current, and average power.
- Explain why AC current is used for power transmission.

Alternating Current

Most of the examples dealt with so far, and particularly those utilizing batteries, have constant voltage sources. Once the current is established, it is thus also a 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. If the source varies periodically, particularly sinusoidally, the circuit is known as an alternating current circuit. Examples include the commercial and residential power that serves so many of our needs. Figure $5.7.1$ shows graphs of voltage and current versus time for typical DC and AC power. The AC voltages and frequencies commonly used in homes and businesses vary around the world.

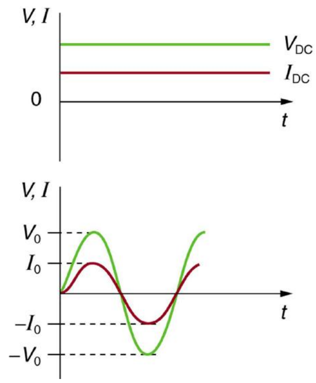


Figure $5.7.1$: (a) DC voltage and current are constant in time, once the current is established. (b) A graph of voltage and current versus time for 60-Hz AC power. The voltage and current are sinusoidal and are in phase for a simple resistance circuit. The frequencies and peak voltages of AC sources differ greatly.

Figure $5.7.2$ shows a schematic of a simple circuit with an AC voltage source. The voltage between the terminals fluctuates as shown, with the **AC voltage** given by

$$V = V_0 sin 2\pi ft, \tag{5.7.1}$$

where V is the voltage at time t, V_0, V_0 is the peak voltage, and f is the frequency in hertz. For this simple resistance circuit, $I = V/R$, and so the **AC current** is

$$I = I_0 sin 2\pi ft, \tag{5.7.2}$$

where I is the current at time t, and $I_0 = V_0/R$ is the peak current. For this example, the voltage and current are said to be in phase, as seen in Figure $5.7.1b$.

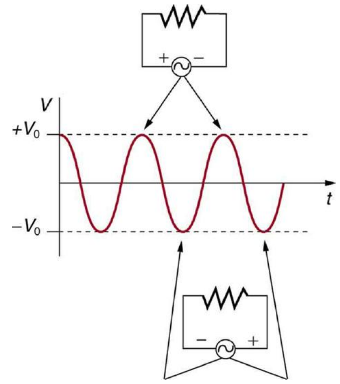


Figure 5.7.2: The potential difference V between the terminals of an AC voltage source fluctuates as shown. The mathematical expression for V is given by $V = V_0 sin 2\pi ft$.

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 a stroboscopic effect evidencing AC. The fact that the light output fluctuates means that the power is fluctuating. The power supplied is $P = IV$. Using the expressions for I and V above, we see that the time dependence of power is $P = I_0 V_0 sin^2 2\pi ft$, as shown in Figure 5.7.3.

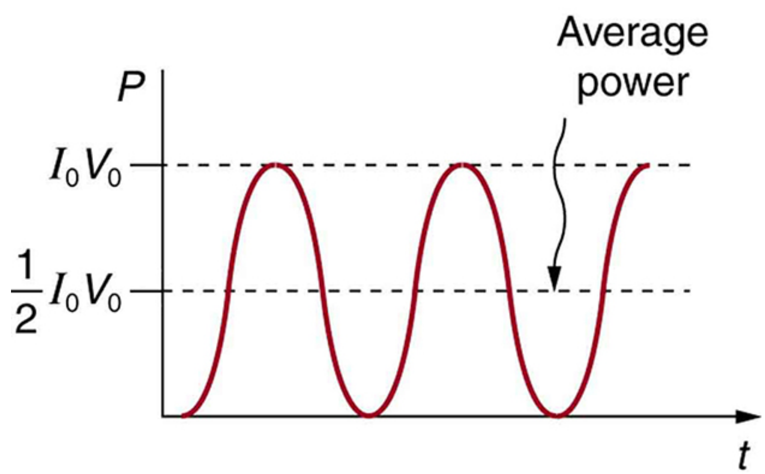


Figure 5.7.3: AC power as a function of time. Since the voltage and current are in phase here, their product is non-negative and fluctuates between zero and $I_0 V_0$. Average power is $(1/2) I_0 V_0$.

Making Connections: Take-Home Experiment—AC/DC Lights

Wave your hand back and forth between your face and a fluorescent light bulb. Do you observe the same thing with the headlights on your car? Explain what you observe. *Warning: Do not look directly at very bright light.*

We are most often concerned with average power rather than its fluctuations—that 60-W light bulb in your desk lamp has an average power consumption of 60 W, for example. As illustrated in Figure 3, the average power P_{ave} is

$$P_{ave} = \frac{1}{2} I_0 V_0. \tag{5.7.3}$$

This is evident from the graph, since the areas above and below the $(1/2) I_0 V_0$ line are equal, but it can also be proven using trigonometric identities. Similarly, we define an average or **rms current** I_{rms} and average or **rms voltage** V_{rms} to be, respectively,

$$I_{rms} = \frac{I_0}{\sqrt{2}} \tag{5.7.4}$$

and

$$V_{rms} = \frac{V_0}{\sqrt{2}}. \tag{5.7.5}$$

where rms stands for root mean square, a particular kind of average. In general, to obtain a root mean square, the particular quantity is squared, its mean (or average) is found, and the square root is taken. This is useful for AC, since the average value is zero. Now,

$$P_{ave} = I_{rms} V_{rms}, \tag{5.7.6}$$

which gives

$$P_{ave} = \frac{I_0}{\sqrt{2}} \cdot \frac{V_0}{\sqrt{2}} = \frac{1}{2} I_0 V_0, \tag{5.7.7}$$

as stated above. It is standard practice to quote I_{rms}, V_{rms}, and P_{ave} rather than the peak values. For example, most household electricity is 120 V AC, which means that V_{rms} is 120 V. The common 10-A circuit breaker will interrupt a sustained I_{rms} greater than 10 A. Your 1.0-kW microwave oven consumes $P_{ave} = 1.0 kW$, and so on. You can think of these rms and average values as the equivalent DC values for a simple resistive circuit.

To summarize, when dealing with AC, Ohm's law and the equations for power are completely analogous to those for DC, but rms and average values are used for AC. Thus, for AC, Ohm's law is written

$$I_{rms} = \frac{V_{rms}}{R}. \tag{5.7.8}$$

The various expressions for AC power P_{ave} are

$$P_{ave} = I_{rms} V_{rms}, \tag{5.7.9}$$

$$P_{ave} = \frac{V_{rms}^2}{R}, \tag{5.7.10}$$

and

$$P_{ave} = I_{rms}^2 R. \tag{5.7.11}$$

Example 5.7.1: Peak Voltage and Power for AC

(a) What is the value of the peak voltage for 120-V AC power?

Strategy

We are told that V_{rms} is 120V and P_{ave} is 60.0 W. We can use $V_{rms} = \frac{V_0}{\sqrt{2}}$ to find the peak voltage, and we can manipulate the definition of power to find the peak power from the given average power.

Solution

Solving the equation $V_{rms} = \frac{V_0}{\sqrt{2}}$ for the peak voltage V_0 and substituting the known value for V_{rms} gives

$$V_0 = \sqrt{2} V_{rms} = 1.414\,(120 V) = 170 V. \tag{5.7.12}$$

Discussion

This means that the AC voltage swings from 170 V to $-170 V$ and back 60 times every second. An equivalent DC voltage is a constant 120 V.

(b) What is the peak power consumption rate of a 60.0-W AC light bulb?

Solution

Peak power is peak current times peak voltage. Thus,

$$P_0 = I_0 V_0 = 2\left(\frac{1}{2} I_0 V_0\right) = 2 P_{ave}. \tag{5.7.13}$$

We know the average power is 60.0 W, and so

$$P_0 = 2\,(60.0 W) = 120 W. \tag{5.7.14}$$

Discussion

So the power swings from zero to 120 W one hundred twenty times per second (twice each cycle), and the power averages 60 W.

Why Use AC for Power Distribution?

Most large power-distribution systems are AC. Moreover, the power is transmitted at much higher voltages than the 120-V AC (240 V in most parts of the world) we use in homes and on the job. Economies of scale make it cheaper to build a few very large electric power-generation plants than to build numerous small ones. This necessitates sending power long distances, and it is obviously important that energy losses en route be minimized. High voltages can be transmitted with much smaller power losses than low voltages, as we shall see. (See Figure 4.) For safety reasons, the voltage at the user is reduced to familiar values. The crucial factor is that it is much easier to increase and decrease AC voltages than DC, so AC is used in most large power distribution systems.

Figure 5.7.4: Power is distributed over large distances at high voltage to reduce power loss in the transmission lines. The voltages generated at the power plant are stepped up by passive devices called transformers (see Transformers) to 330,000 volts (or more in some places worldwide). At the point of use, the transformers reduce the voltage

Example 5.7.2: Power losses are less for high-voltage transmission

(a) What current is needed to transmit 100 MW of power at 200 kV?

Strategy

We are given $P_{ave} = 100 MW$, $V_{rms} = 200 kV$, and the resistance of the lines is $R = 1.00 \Omega$. Using these givens, we can find the current flowing (from $P = IV$) and then the power dissipated in the lines ($P = I^2 R$), and we take the ratio to the total power transmitted.

Solution

To find the current, we rearrange the relationship $P_{ave} = I_{rms} V_{rms}$ and substitute known values. This gives

$$I_{rms} = \frac{P_{ave}}{V_{rms}} = \frac{100 \times 10^6 W}{200 \times 0^3 V} = 500 A. \tag{5.7.15}$$

(b) What is the power dissipated by the transmission lines if they have a resistance of 1.00Ω?

Solution

Knowing the current and given the resistance of the lines, the power dissipated in them is found from $P_{ave} = I^2_{rms} R.$. Substituting the known value gives

$$P_{ave} = I^2_{rms} R = (500 A)^2 (1.00 \Omega) = 250 kW. \tag{5.7.16}$$

(c) What percentage of the power is lost in the transmission lines?

Solution

The percent loss is the ratio of this lost power to the total or input power, multiplied by 100:

$$\tag{5.7.17}$$

Discussion

One-fourth of a percent is an acceptable loss. Note that if 100 MW of power had been transmitted at 25 kV, then a current of 4000 A would have been needed. This would result in a power loss in the lines of 16.0 MW, or 16.0% rather than 0.250%. The lower the voltage, the more current is needed, and the greater the power loss in the fixed-resistance transmission lines. Of course, lower-resistance lines can be built, but this requires larger and more expensive wires. If superconducting lines could be economically produced, there would be no loss in the transmission lines at all. But, as we shall see in a later chapter, there is a limit to current in superconductors, too. In short, high voltages are more economical for transmitting power, and AC voltage is much easier to raise and lower, so that AC is used in most large-scale power distribution systems.

It is widely recognized that high voltages pose greater hazards than low voltages. But, in fact, some high voltages, such as those associated with common static electricity, can be harmless. So it is not voltage alone that determines a hazard. It is not so widely recognized that AC shocks are often more harmful than similar DC shocks. Thomas Edison thought that AC shocks were more harmful and set up a DC power-distribution system in New York City in the late 1800s. There were bitter fights, in particular between Edison and George Westinghouse and Nikola Tesla, who were advocating the use of AC in early power-distribution systems. AC has prevailed largely due to transformers and lower power losses with high-voltage transmission.

PHET EXPLORATIONS: GENERATOR

Generate electricity with a bar magnet! Discover the physics behind the phenomena by exploring magnets and how you can use them to make a bulb light.

Figure $\mathbf{5.7.5}$***:*** Generator

Summary

- Direct current (DC) is the flow of electric current in only one direction. It refers to systems where the source voltage is constant.
- The voltage source of an alternating current (AC) system puts out $V = V_0 \sin 2\pi ft$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz.
- In a simple circuit, $I = V/R$ and AC current is $I = I_0 \sin 2\pi ft$, where I is the current at time t, and $I_0 = V_0/R$ is the peak current.
- The average AC power is $P_{ave} = \frac{1}{2} I_0 V_0$.
- Average (rms) current I_{rms} and average (rms) voltage V_{rms} and $I_{rms} = \frac{I_0}{\sqrt{2}}$ and $V_{rms} = \frac{V_0}{\sqrt{2}}$, where rms stands for root mean square.
- Thus, $P_{ave} = I_{rms} V_{rms}$.
- Ohm's law for AC is $I_{rms} = \frac{V_{rms}}{R}$.
- Expressions for the average power of an AC circuit are $P_{ave} = I_{rms} V_{rms}$, $P_{ave} = \frac{V_{rms}^2}{R}$, and $P_{ave} = I_{rms}^2 R$, analogous to the expressions for DC circuits.

Glossary

direct current

(DC) the flow of electric charge in only one direction

alternating current

(AC) the flow of electric charge that periodically reverses direction

AC voltage

voltage that fluctuates sinusoidally with time, expressed as $V = V_0 \sin 2\pi ft$, where V is the voltage at time t, V_0 is the peak voltage, and f is the frequency in hertz

AC current

current that fluctuates sinusoidally with time, expressed as $I = I_0 \sin 2\pi ft$, where I is the current at time t, I_0 is the peak current, and f is the frequency in hertz

rms current

the root mean square of the current, $I_{rms} = I_0/\sqrt{2}$, where I_0 is the peak current, in an AC system

rms voltage

the root mean square of the voltage, $V_{rms} = V_0/\sqrt{2}$, where V_0 is the peak voltage, in an AC system

---

5.8: Current and Resistance (Summary)

Key Terms

ampere (amp)	SI unit for current; $1 A=1 C/s$
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
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
nonohmic	type of a material for which Ohm's law is not valid
ohm	(Ω) unit of electrical resistance, $1 \Omega=1 V/A$
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

Key Equations

Average electrical current	$I_{ave}=\frac{\Delta Q}{\Delta t}$
Definition of an ampere	$1 A=1 C/s$
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

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 $1A=1C/s$.
- Current consists of the flow of free charges, such as electrons, protons, and ions.

Basic 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.

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.

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 $J = \sigma E$.

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 ⋅ h).
- The total energy used over a time interval can be found by $E = \int P dt$.

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).

5.9: Current and Resistance (Exercises)

Conceptual Questions

Electrical Current

1. Can a wire carry a current and still be neutral—that is, have a total charge of zero? Explain.

2. Car batteries are rated in ampere-hours (**A·h**). To what physical quantity do ampere-hours correspond (voltage, current, charge, energy, power,…)?

3. When working with high-power electric circuits, it is advised that whenever possible, you work "one-handed" or "keep one hand in your pocket." Why is this a sensible suggestion?

Basic Model of Conduction in Metals

4. Incandescent light bulbs are being replaced with more efficient LED and CFL light bulbs. Is there any obvious evidence that incandescent light bulbs might not be that energy efficient? Is energy converted into anything but visible light?

5. It was stated that the motion of an electron appears nearly random when an electrical field is applied to the conductor. What makes the motion nearly random and differentiates it from the random motion of molecules in a gas?

6. Electric circuits are sometimes explained using a conceptual model of water flowing through a pipe. In this conceptual model, the voltage source is represented as a pump that pumps water through pipes and the pipes connect components in the circuit. Is a conceptual model of water flowing through a pipe an adequate representation of the circuit? How are electrons and wires similar to water molecules and pipes? How are they different?

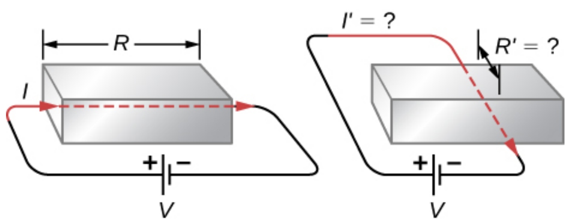
7. An incandescent light bulb is partially evacuated. Why do you suppose that is?

Resistivity and Resistance

8. The IR drop across a resistor means that there is a change in potential or voltage across the resistor. Is there any change in current as it passes through a resistor? Explain.

9. Do impurities in semiconducting materials listed in Table 9.1 supply free charges? (Hint: Examine the range of resistivity for each and determine whether the pure semiconductor has the higher or lower conductivity.)

10. Does the resistance of an object depend on the path current takes through it? Consider, for example, a rectangular bar—is its resistance the same along its length as across its width?

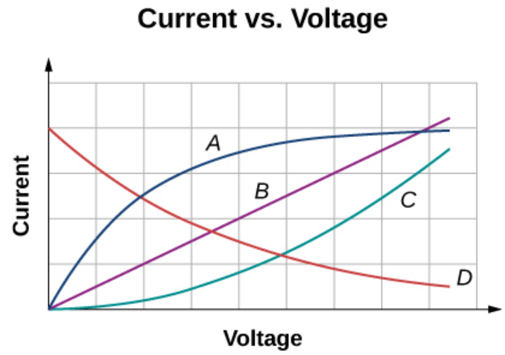


11. If aluminum and copper wires of the same length have the same resistance, which has the larger diameter? Why?

Ohm's Law

12. In Determining Field from Potential, resistance was defined as $R \equiv \frac{V}{I}$. In this section, we presented Ohm's law, which is commonly expressed as **V=IR.** The equations look exactly alike. What is the difference between Ohm's law and the definition of resistance?

13. Shown below are the results of an experiment where four devices were connected across a variable voltage source. The voltage is increased and the current is measured. Which device, if any, is an ohmic device?



14. The current I is measured through a sample of an ohmic material as a voltage V is applied. (a) What is the current when the voltage is doubled to 2V (assume the change in temperature of the material is negligible)? (b) What is the voltage applied is the current measured is 0.2I (assume the change in temperature of the material is negligible)? What will happen to the current if the material if the voltage remains constant, but the temperature of the material increases significantly?

Electrical Energy and Power

15. Common household appliances are rated at 110 V, but power companies deliver voltage in the kilovolt range and then step the voltage down using transformers to 110 V to be used in homes. You will learn in later chapters that transformers consist of many turns of wire, which warm up as current flows through them, wasting some of the energy that is given off as heat. This sounds inefficient. Why do the power companies transport electric power using this method?

16. Your electric bill gives your consumption in units of kilowatt-hour (kW ⋅⋅ h). Does this unit represent the amount of charge, current, voltage, power, or energy you buy?

17. Resistors are commonly rated at $\frac{1}{8}W$, $\frac{1}{4}W$, $\frac{1}{2}W$, 1 W and 2 W for use in electrical circuits. If a current of **I=2.00A** is accidentally passed through a **R=1.00Ω** resistor rated at 1 W, what would be the most probable outcome? Is there anything that can be done to prevent such an accident?

18. An immersion heater is a small appliance used to heat a cup of water for tea by passing current through a resistor. If the voltage applied to the appliance is doubled, will the time required to heat the water change? By how much? Is this a good idea?

Problems

Electrical Current

21. A Van de Graaff generator is one of the original particle accelerators and can be used to accelerate charged particles like protons or electrons. You may have seen it used to make human hair stand on end or produce large sparks. One application of the Van de Graaff generator is to create X-rays by bombarding a hard metal target with the beam. Consider a beam of protons at 1.00 keV and a current of 5.00 mA produced by the generator.

(a) What is the speed of the protons?

(b) How many protons are produced each second?

22. A cathode ray tube (CRT) is a device that produces a focused beam of electrons in a vacuum. The electrons strike a phosphor-coated glass screen at the end of the tube, which produces a bright spot of light. The position of the bright spot of light on the screen can be adjusted by deflecting the electrons with electrical fields, magnetic fields, or both. Although the CRT tube was once commonly found in televisions, computer displays, and oscilloscopes, newer appliances use a liquid crystal display (LCD) or plasma screen. You still may come across a CRT in your study of science. Consider a CRT with an electron beam average current of 25.00μA25.00μA. How many electrons strike the screen every minute?

23. How many electrons flow through a point in a wire in 3.00 s if there is a constant current of **I=4.00A**?

24. A conductor carries a current that is decreasing exponentially with time. The current is modeled as $I = I_0 e^{-t/\tau}$, where $I_0 = 3.00A$ is the current at time **t=0.00s** and **τ=0.50s** is the time constant. How much charge flows through the conductor between **t=0.00s** and **t=3τ**?

25. The quantity of charge through a conductor is modeled as $Q = 4.00\frac{C}{s^4}t^4 - 1.00\frac{C}{s}t + 6.00mC$. What is the current at time **t=3.00s**?

26. The current through a conductor is modeled as $I(t) = I_m sin(2\pi[60Hz]t)$. Write an equation for the charge as a function of time.

27. The charge on a capacitor in a circuit is modeled as $Q(t) = Q_{max} cos(\omega t + \phi)$. What is the current through the circuit as a function of time?

Basic Model of Conduction in Metals

28. An aluminum wire 1.628 mm in diameter (14-gauge) carries a current of 3.00 amps.

(a) What is the absolute value of the charge density in the wire?

(b) What is the drift velocity of the electrons?

(c) What would be the drift velocity if the same gauge copper were used instead of aluminum? The density of copper is $8.96g/cm^3$ and the density of aluminum is $2.70g/cm^3$. The molar mass of aluminum is 26.98 g/mol and the molar mass of copper is 63.5 g/mol. Assume each atom of metal contributes one free electron.

29. The current of an electron beam has a measured current of **I=50.00μA** with a radius of 1.00 mm. What is the magnitude of the current density of the beam?

30. A high-energy proton accelerator produces a proton beam with a radius of **r=0.90mm**. The beam current is **I=9.00μA** and is constant. The charge density of the beam is $n = 6.00 \times 10^{11}$ protons per cubic meter.

(a) What is the current density of the beam?

(b) What is the drift velocity of the beam?

(c) How much time does it take for 1.00×10^{10} protons to be emitted by the accelerator?

31. Consider a wire of a circular cross-section with a radius of **R=3.00mm**. The magnitude of the current density is modeled as $J = cr^2 = 5.00 \times 10^6 \frac{A}{m^4} r^2$. What is the current through the inner section of the wire from the center to **r=0.5R**?

32. A cylindrical wire has a current density from the center of the wire's cross section as $J(r) = Cr^2$ where r is in meters, J is in amps per square meter, and $C = 10^3$ A/m^4. This current density continues to the end of the wire at a radius of 1.0 mm. Calculate the current just outside of this wire.

33. The current supplied to an air conditioner unit is 4.00 amps. The air conditioner is wired using a 10-gauge (diameter 2.588 mm) wire. The charge density is $n = 8.48 \times 10^{28} \frac{electrons}{m^3}$. Find the magnitude of

(a) current density and

(b) the drift velocity.

Resistivity and Resistance

34. What current flows through the bulb of a 3.00-V flashlight when its hot resistance is **3.60Ω**?

35. Calculate the effective resistance of a pocket calculator that has a 1.35-V battery and through which 0.200 mA flows.

36. How many volts are supplied to operate an indicator light on a DVD player that has a resistance of **140Ω**, given that 25.0 mA passes through it?

37. What is the resistance of a 20.0-m-long piece of 12-gauge copper wire having a 2.053-mm diameter?

38. The diameter of 0-gauge copper wire is 8.252 mm. Find the resistance of a 1.00-km length of such wire used for power transmission.

39. If the 0.100-mm-diameter tungsten filament in a light bulb is to have a resistance of **0.200Ω** at **20.0°C**, how long should it be?

40. A lead rod has a length of 30.00 cm and a resistance of **5.00μΩ**. What is the radius of the rod?

41. Find the ratio of the diameter of aluminum to copper wire, if they have the same resistance per unit length (as they might in household wiring).

42. What current flows through a 2.54-cm-diameter rod of pure silicon that is 20.0 cm long, when 1.00×10^3 is applied to it? (Such a rod may be used to make nuclear-particle detectors, for example.)

43. (a) To what temperature must you raise a copper wire, originally at $20.0°C$, to double its resistance, neglecting any changes in dimensions? (b) Does this happen in household wiring under ordinary circumstances?

44. A resistor made of nichrome wire is used in an application where its resistance cannot change more than 1.00% from its value at $20.0°C$. Over what temperature range can it be used?

45. Of what material is a resistor made if its resistance is 40.0% greater at $100.0°C$ than at $20.0°C$?

46. An electronic device designed to operate at any temperature in the range from $-10.0°C$ to $55.0°C$ contains pure carbon resistors. By what factor does their resistance increase over this range?**47.**

(a) Of what material is a wire made, if it is 25.0 m long with a diameter of 0.100 mm and has a resistance of 77.7Ω at $20.0°C$? (b) What is its resistance at $150.0°C$?

48. Assuming a constant temperature coefficient of resistivity, what is the maximum percent decrease in the resistance of a constantan wire starting at $20.0°C$?

49. A copper wire has a resistance of 0.500Ω at $20.0°C$, and an iron wire has a resistance of 0.525Ω at the same temperature. At what temperature are their resistances equal?

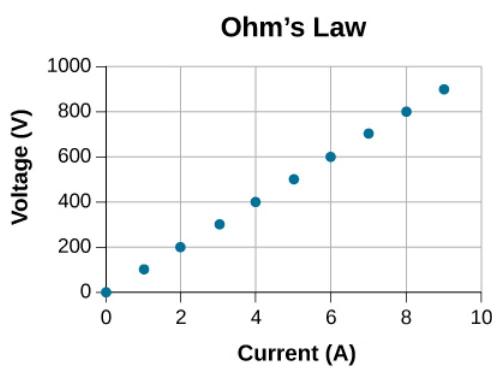
Ohm's Law

50. A $2.2\text{-k}\Omega$ resistor is connected across a D cell battery (1.5 V). What is the current through the resistor?

51. A resistor rated at $250k\Omega$ is connected across two D cell batteries (each 1.50 V) in series, with a total voltage of 3.00 V. The manufacturer advertises that their resistors are within 5% of the rated value. What are the possible minimum current and maximum current through the resistor?

52. A resistor is connected in series with a power supply of 20.00 V. The current measure is 0.50 A. What is the resistance of the resistor?

53. A resistor is placed in a circuit with an adjustable voltage source. The voltage across and the current through the resistor and the measurements are shown below. Estimate the resistance of the resistor.



54. The following table show the measurements of a current through and the voltage across a sample of material. Plot the data, and assuming the object is an ohmic device, estimate the resistance.

Table: Measurements of current through and the voltage across a sample of material

I(A)	V(V)
0	3
2	23
4	39
6	58
8	77

10	100
12	119
14	142
16	162

Electrical Energy and Power

55. A **20.00-V** battery is used to supply current to a **10-kΩ** resistor. Assume the voltage drop across any wires used for connections is negligible.

(a) What is the current through the resistor?

(b) What is the power dissipated by the resistor?

(c) What is the power input from the battery, assuming all the electrical power is dissipated by the resistor?

(d) What happens to the energy dissipated by the resistor?

56. What is the maximum voltage that can be applied to a **20-kΩ** resistor rated at $\frac{1}{4}W$?

57. A heater is being designed that uses a coil of 14-gauge nichrome wire to generate 300 W using a voltage of **V=110V.** How long should the engineer make the wire?

58. An alternative to CFL bulbs and incandescent bulbs are light-emitting diode (LED) bulbs. A 100-W incandescent bulb can be replaced by a 16-W LED bulb. Both produce 1600 lumens of light. Assuming the cost of electricity is $0.10 per kilowatt-hour, how much does it cost to run the bulb for one year if it runs for four hours a day?

59. The power dissipated by a resistor with a resistance of **R=100Ω** is **P=2.0W**. What are the current through and the voltage drop across the resistor?

60. Running late to catch a plane, a driver accidentally leaves the headlights on after parking the car in the airport parking lot. During takeoff, the driver realizes the mistake. Having just replaced the battery, the driver knows that the battery is a 12-V automobile battery, rated at 100 **A·h**. The driver, knowing there is nothing that can be done, estimates how long the lights will shine, assuming there are two 12-V headlights, each rated at 40 W. What did the driver conclude?

61. A physics student has a single-occupancy dorm room. The student has a small refrigerator that runs with a current of 3.00 A and a voltage of 110 V, a lamp that contains a 100-W bulb, an overhead light with a 60-W bulb, and various other small devices adding up to 3.00 W.

(a) Assuming the power plant that supplies 110 V electricity to the dorm is 10 km away and the two aluminum transmission cables use 0-gauge wire with a diameter of 8.252 mm, estimate the percentage of the total power supplied by the power company that is lost in the transmission.

(b) What would be the result is the power company delivered the electric power at 110 kV?

62. A 0.50-W, **220-Ω** resistor carries the maximum current possible without damaging the resistor. If the current were reduced to half the value, what would be the power consumed?

Additional Problems

69. A coaxial cable consists of an inner conductor with radius $r_i = 0.25cm$ and an outer radius of $r_o = 0.5cm$ and has a length of 10 meters. Plastic, with a resistivity of $\rho = 2.00 \times 10^{13} \Omega \cdot m$, separates the two conductors. What is the resistance of the cable?

70. A 10.00-meter long wire cable that is made of copper has a resistance of 0.051 ohms.

(a) What is the weight if the wire was made of copper?

(b) What is the weight of a 10.00-meter-long wire of the same gauge made of aluminum?

(c)What is the resistance of the aluminum wire? The density of copper is $8960 kg/m^3$ and the density of aluminum is $2760 kg/m^3$.

71. A nichrome rod that is 3.00 mm long with a cross-sectional area of $1.00 mm^2$ is used for a digital thermometer.

(a) What is the resistance at room temperature?

(b) What is the resistance at body temperature?

72. The temperature in Philadelphia, PA can vary between **68.00°F** and **100.00°F** in one summer day. By what percentage will an aluminum wire's resistance change during the day?

73. When 100.0 V is applied across a 5-gauge (diameter 4.621 mm) wire that is 10 m long, the magnitude of the current density is $2.0 \times 10^8 A/m^2$. What is the resistivity of the wire?

74. A wire with a resistance of **5.0Ω** is drawn out through a die so that its new length is twice times its original length. Find the resistance of the longer wire. You may assume that the resistivity and density of the material are unchanged.

75. What is the resistivity of a wire of 5-gauge wire ($A = 16.8 \times 10^{-6} m^2$), 5.00 m length, and **5.10mΩ** resistance?

76. Coils are often used in electrical and electronic circuits. Consider a coil which is formed by winding 1000 turns of insulated 20-gauge copper wire (area $0.52 mm^2$) in a single layer on a cylindrical non-conducting core of radius 2.0 mm. What is the resistance of the coil? Neglect the thickness of the insulation.

77. Currents of approximately 0.06 A can be potentially fatal. Currents in that range can make the heart fibrillate (beat in an uncontrolled manner). The resistance of a dry human body can be approximately **100kΩ**.

(a) What voltage can cause 0.2 A through a dry human body?

(b) When a human body is wet, the resistance can fall to **100Ω**. What voltage can cause harm to a wet body?

78. A 20.00-ohm, 5.00-watt resistor is placed in series with a power supply.

(a) What is the maximum voltage that can be applied to the resistor without harming the resistor?

(b) What would be the current through the resistor?

79. A battery with an emf of 24.00 V delivers a constant current of 2.00 mA to an appliance. How much work does the battery do in three minutes?

80. A 12.00-V battery has an internal resistance of a tenth of an ohm.

(a) What is the current if the battery terminals are momentarily shorted together?

(b) What is the terminal voltage if the battery delivers 0.25 amps to a circuit?

Challenge Problems

81. A 10-gauge copper wire has a cross-sectional area $A = 5.26 mm^2$ and carries a current of **I=5.00A**. The density of copper is $\rho = 89.50 g/cm^3$. One mole of copper atoms ($6.02 \times 10^{23} atoms$) has a mass of approximately 63.50 g. What is the magnitude of the drift velocity of the electrons, assuming that each copper atom contributes one free electron to the current?

82. The current through a 12-gauge wire is given as **I(t)=(5.00A)sin(2π60Hzt)**. What is the current density at time 15.00 ms?

83. A particle accelerator produces a beam with a radius of 1.25 mm with a current of 2.00 mA. Each proton has a kinetic energy of 10.00 MeV.

(a) What is the velocity of the protons?

(b) What is the number (n) of protons per unit volume?

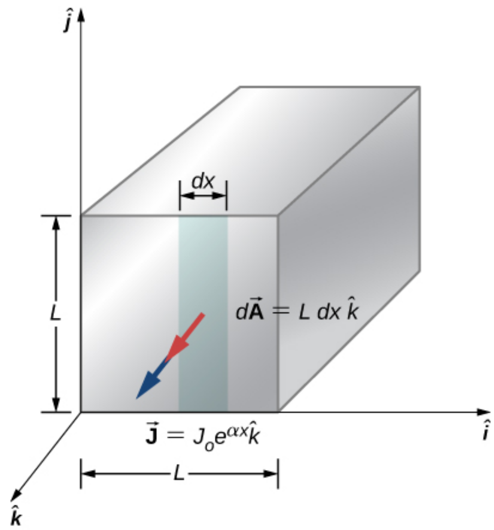
(c) How many electrons pass a cross sectional area each second?

84. In this chapter, most examples and problems involved direct current (DC). DC circuits have the current flowing in one direction, from positive to negative. When the current was changing, it was changed linearly from $I = -I_{max}$ to $I = +I_{max}$ and the voltage changed linearly from $V = -V_{max}$ to $V = +V_{max}$, where $V_{max} = I_{max} R$. Suppose a voltage source is placed in series with a resistor of **R=10Ω** that supplied a current that alternated as a sine wave, for example, $I(t) = (3.00 A) sin(\frac{2\pi}{4.00 s} t)$. (a) What would a graph of the voltage drop across the resistor **V(t)** versus time look like? (b)

What would a plot of **V(t)** versus **I(t)** for one period look like? (Hint: If you are not sure, try plotting **V(t)** versus **I(t)** using a spreadsheet.)

85. A current of **I=25A** is drawn from a 100-V battery for 30 seconds. By how much is the chemical energy reduced?

86. Consider a square rod of material with sides of length **L=3.00cm** with a current density of $J = J_0 e^{\alpha x} \hat{k} = (0.35 \frac{A}{m^2}) e^{(2.1 \times 10^{-3} m^{-1}) x} \hat{k}$ as shown below. Find the current that passes through the face of



87. A resistor of an unknown resistance is placed in an insulated container filled with 0.75 kg of water. A voltage source is connected in series with the resistor and a current of 1.2 amps flows through the resistor for 10 minutes. During this time, the temperature of the water is measured and the temperature change during this time is **ΔT=10.00°C**.

(a) What is the resistance of the resistor?

(b) What is the voltage supplied by the power supply?

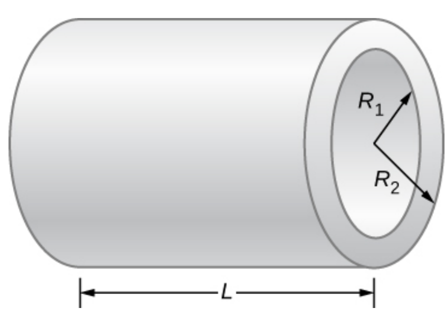
88. The charge that flows through a point in a wire as a function of time is modeled as $q(t) = q_0 e^{-t/T} = 10.0 C e^{-t/5s}$.

(a) What is the initial current through the wire at time **t=0.00s**?

(b) Find the current at time $t = \frac{1}{2} T$.

(c) At what time t will the current be reduced by one-half $I = \frac{1}{2} I_0$?

89. Consider a resistor made from a hollow cylinder of carbon as shown below. The inner radius of the cylinder is Ri=0.20mmRi=0.20mm and the outer radius is $R_0 = 0.30 mm$. The length of the resistor is **L=0.90mm**. The resistivity of the carbon is $\rho = 3.5 \times 10^{-5} \Omega \cdot m$. (a) Prove that the resistance perpendicular from the axis is $R = \frac{\rho}{2\pi L} ln(\frac{R_0}{R_i})$. (b) What is the resistance?



90. What is the current through a cylindrical wire of radius **R=0.1mm** if the current density is $J = \frac{J_0}{R} r$, where $J_0 = 32000 \frac{A}{m^2}$?

91. A student uses a 100.00-W, 115.00-V radiant heater to heat the student's dorm room, during the hours between sunset and sunrise, 6:00 p.m. to 7:00 a.m.

(a) What current does the heater operate at?

(b) How many electrons move through the heater?

(c) What is the resistance of the heater?

(d) How much heat was added to the dorm room?

92. A 12-V car battery is used to power a 20.00-W, 12.00-V lamp during the physics club camping trip/star party. The cable to the lamp is 2.00 meters long, 14-gauge copper wire with a charge density of $n = 9.50 \times 10^{28} m^{-3}$.

(a) What is the current draw by the lamp?

(b) How long would it take an electron to get from the battery to the lamp?

93. A physics student uses a 115.00-V immersion heater to heat 400.00 grams (almost two cups) of water for herbal tea. During the two minutes it takes the water to heat, the physics student becomes bored and decides to figure out the resistance of the heater. The student starts with the assumption that the water is initially at the temperature of the room $T_i = 25.00^\circ C$ and reaches $T_f = 100.00^\circ C$. The specific heat of the water is $c = 4180 \frac{J}{kg}$. What is the resistance of the heater?

Contributors and Attributions

5.10: Current and Resistance (Answers)

Conceptual Questions

1. If a wire is carrying a current, charges enter the wire from the voltage source's positive terminal and leave at the negative terminal, so the total charge remains zero while the current flows through it.

3. Using one hand will reduce the possibility of "completing the circuit" and having current run through your body, especially current running through your heart.

5. Even though the electrons collide with atoms and other electrons in the wire, they travel from the negative terminal to the positive terminal, so they drift in one direction. Gas molecules travel in completely random directions.

7. In the early years of light bulbs, the bulbs are partially evacuated to reduce the amount of heat conducted through the air to the glass envelope. Dissipating the heat would cool the filament, increasing the amount of energy needed to produce light from the filament. It also protects the glass from the heat produced from the hot filament. If the glass heats, it expands, and as it cools, it contacts. This expansion and contraction could cause the glass to become brittle and crack, reducing the life of the bulbs. Many bulbs are now partially filled with an inert gas. It is also useful to remove the oxygen to reduce the possibility of the filament actually burning. When the original filaments were replaced with more efficient tungsten filaments, atoms from the tungsten would evaporate off the filament at such high temperatures. The atoms collide with the atoms of the inert gas and land back on the filament.

9. In carbon, resistivity increases with the amount of impurities, meaning fewer free charges. In silicon and germanium, impurities decrease resistivity, meaning more free electrons.

11. Copper has a lower resistivity than aluminum, so if length is the same, copper must have the smaller diameter.

13. Device B shows a linear relationship and the device is ohmic.

15. Although the conductors have a low resistance, the lines from the power company can be kilometers long. Using a high voltage reduces the current that is required to supply the power demand and that reduces line losses.

17. The resistor would overheat, possibly to the point of causing the resistor to burn. Fuses are commonly added to circuits to prevent such accidents.

Problems

21. a. $v = 4.38 \times 10^5 \frac{m}{s}$;

b. $\Delta q = 5.00 \times 10^{-3} C$,no. of protons=$3.13 \times 10^{16}$

23. $I = \frac{\Delta Q}{\Delta t}$, $\Delta Q = 12.00C$, no. of electrons=7.46×10^{19}

25. $I(t) = 0.016 \frac{C}{s^4} t^3 - 0.001 \frac{C}{s}$ $I(3.00s) = 0.431A$

27. $I(t) = -I_{max} sin(\omega t + \phi)$

29. $|J| = 15.92A/m^2$

31. $I = 3.98 \times 10^{-5} A$

33. a. $|J| = 7.60 \times 10^5 \frac{A}{m^2}$;

b. $v_d = 5.60 \times 10^{-5} \frac{m}{s}$

35. $R = 6.750k\Omega$

37. $R = 0.10\Omega$

39. $R = \rho \frac{L}{A}$; $L = 3cm$

41. $\frac{R_{Al}/L_{Al}}{R_{Cu}/L_{Cu}} = \frac{\rho_{Al}\frac{1}{\pi(\frac{D_{Al}}{2})^2}}{\rho_{Cu}\frac{1}{\pi(\frac{D_{Cu}}{2})^2}} = \frac{\rho_{Al}}{\rho_{Cu}}(\frac{D_{Cu}}{D_{Al}})^2 = 1, \frac{D_{Al}}{D_{Cu}} = \sqrt{\frac{\rho_{AI}}{\rho_{Cu}}}$

43. a. $R = R_0(1+\alpha\Delta T), 2 = 1+\alpha\Delta T, \Delta T = 256.4^\circ C, T = 276.4^\circ C$;

b. Under normal conditions, no it should not occur.

45. $R = R_0(1+\alpha\Delta T)\ \ \alpha = 0.006^\circ C^{-1}$, iron

47. a. $R = \rho\frac{L}{A}, \rho = 2.44\times 10^-8\Omega\cdot m$, gold; $R = \rho\frac{L}{A}(1+\alpha\Delta T)$

b. $R = 2.44\times 10^{-8}\Omega\cdot m(\frac{25m}{\pi(\frac{0.100\times 10^{-3}m}{2})^2})(1+0.0034^\circ C^{-1}(150^\circ C - 20^\circ C))R = 112\Omega$

49. $R_{Fe} = 0.525\Omega, R_{Cu} = 0.500\Omega, \alpha_{Fe} = 0.0065^\circ C^{-1}, \alpha_{Cu} = 0.0039^\circ C^{-1}, R_{Fe} = R_{Cu}, R_{0Fe}(1+\alpha_{Fe}(T-T_0))$
$= R_{0Cu}(1+\alpha_{Cu}(T-T_0)), \frac{R_{0Fe}}{R_{0Cu}}(1+\alpha_{Fe}(T-T_0)) = 1+\alpha_{Cu}(T-T_0), T = 2.91^\circ C$

51. $R_{min} = 2.375\times 10^5\Omega, I_{min} = 12.63\mu A$

$R_{max} = 2.625\times 10^5\Omega, I_{max} = 11.43\mu A$

53. $R = 100\Omega$

55. a. $I = 2mA$;

b. $P = 0.04mW$;

c. $P = 0.04mW$;

d. It is converted into heat.

57. $P = \frac{V^2}{R}, R = 40\Omega, A = 2.08mm^2, \rho = 100\times 10^{-8}\Omega\cdot m, R = \rho\frac{L}{A}, L = 83m$

59. $I = 0.14A, V = 14V$

61. a. $I \approx 3.00A + \frac{100W}{110V} + \frac{60W}{110V} + \frac{3.00W}{110V} = 4.48A$

$P = 493W$

$R = 9.91\Omega$,

$P_{loss} = 200.W$

b. $P = 493W$

$I = 0.0045A$

$R = 9.91\Omega$

$P_{loss} = 201\mu W$

Additional Problems

69. $dR = \frac{\rho}{2\pi rL}dr$

$R = \frac{\rho}{2\pi L}ln\frac{r_o}{r_i}$

$R = 2.21\times 10^{11}\Omega$

71. a. $R_0 = 0.003\Omega$;

b. $T_c = 37.0^\circ C\ R = 0.00302\Omega$

73. $\rho = 5.00 \times 10^{-8}\Omega \cdot m$

75. $\rho = 1.71 \times 10^{-8}\Omega \cdot m$

77. $a.\ V = 6000V;$

b. $V = 60V$

79. $P = \frac{W}{t}, W = 8.64J$

Challenge Problems

81. $V = 7.09cm^3\ n = 8.49 \times 10^{28}\frac{electrons}{m^2}\ \ v_d = 7.00 \times 10^{-5}\frac{m}{s}$

83. a. $v = 4.38 \times 10^7 m/s;$

b. $v = 5.81 \times 10^{13}\ \frac{protons}{m^3};$

c. 1.25 $\frac{electrons}{m^3}$

85. $E = 75kJ$

87. a. $P = 52W\ R = 36\Omega;$

b. $V = 43.54V$

89. a. $R = \frac{\rho}{2\pi L} ln(\frac{R_0}{R_i});$

b. $R = 2.5m\Omega$

91. a. $I = 0.870A;$

b. $\text{#electrons} = 2.54 \times 10^{23}$

c. $R = 132\Omega;$

d. $q = 4.68 \times 10^6 J$

93. $P = 1045W, P = \frac{V^2}{R}, R = 12.27\Omega$

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).

CHAPTER OVERVIEW

6: Direct-Current (DC) Resistor Circuits

6.1: Introduction

In the preceding few chapters, we discussed the electrical concepts of voltage, current, and resistance. In this chapter, we use these concepts to begin to learn how to analyze direct-current (DC) circuits. A circuit is a collection of electrical components connected to accomplish a specific task. Figure 6.1.1 shows an amplifier circuit, which takes a small-amplitude signal and amplifies it to power the speakers in earbuds. Although the circuit looks complex, it actually consists of a set of series, parallel, and series-parallel circuits. Direct-current circuits contain current that only flows in one direction in a given section of the circuit.

An early section of this chapter covers the analysis of series and parallel circuits that consist of resistors. Later in this chapter, we introduce the basic equations and techniques to analyze any circuit, including those that are not reducible through simplifying parallel and series elements. But first, we need to understand how to power a circuit.

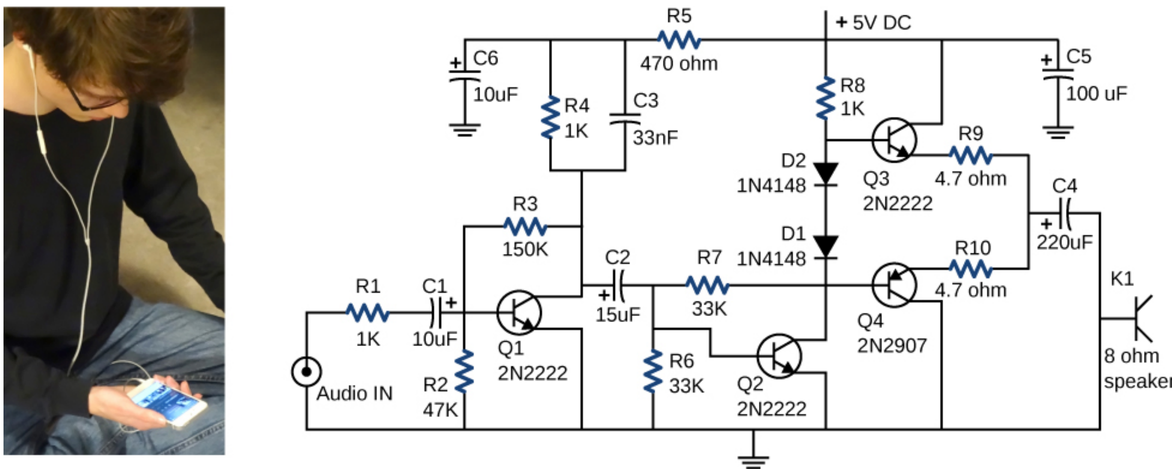


Figure 6.1.1: This circuit shown is used to amplify small signals and power the earbud speakers attached to a cellular phone. This circuit's components include resistors, capacitors, and diodes, and transistors. Circuits using similar components are found in all types of equipment and appliances you encounter in everyday life, such as alarm clocks, televisions, computers, and refrigerators. This chapter will only discuss resistor circuits. (credit: Jane Whitney)

6.2: Source Voltage

Learning Objectives

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

- Describe the source voltage and the internal resistance of a battery.
- Explain the basic operation of a battery.

If you forget to turn off your car lights, they slowly dim as the battery runs down. Why don't they suddenly blink off when the battery's energy is gone? Their gradual dimming implies that the battery output voltage decreases as the battery is depleted. The reason for the decrease in output voltage for depleted batteries is that all voltage sources have two fundamental parts—a source of electrical energy and an internal resistance. In this section, we examine the energy source and the internal resistance.

Introduction to Source Voltage

Voltage has many sources, a few of which are shown in Figure $6.2.2$. All such devices create a **potential difference** and can supply current if connected to a circuit. A special type of potential difference is known as **source voltage.** (Note: Source voltage is called electromotive force (emf) in some text, but this usage is deprecated because emf is not a force and does not have dimensions of newtons in the SI system.)

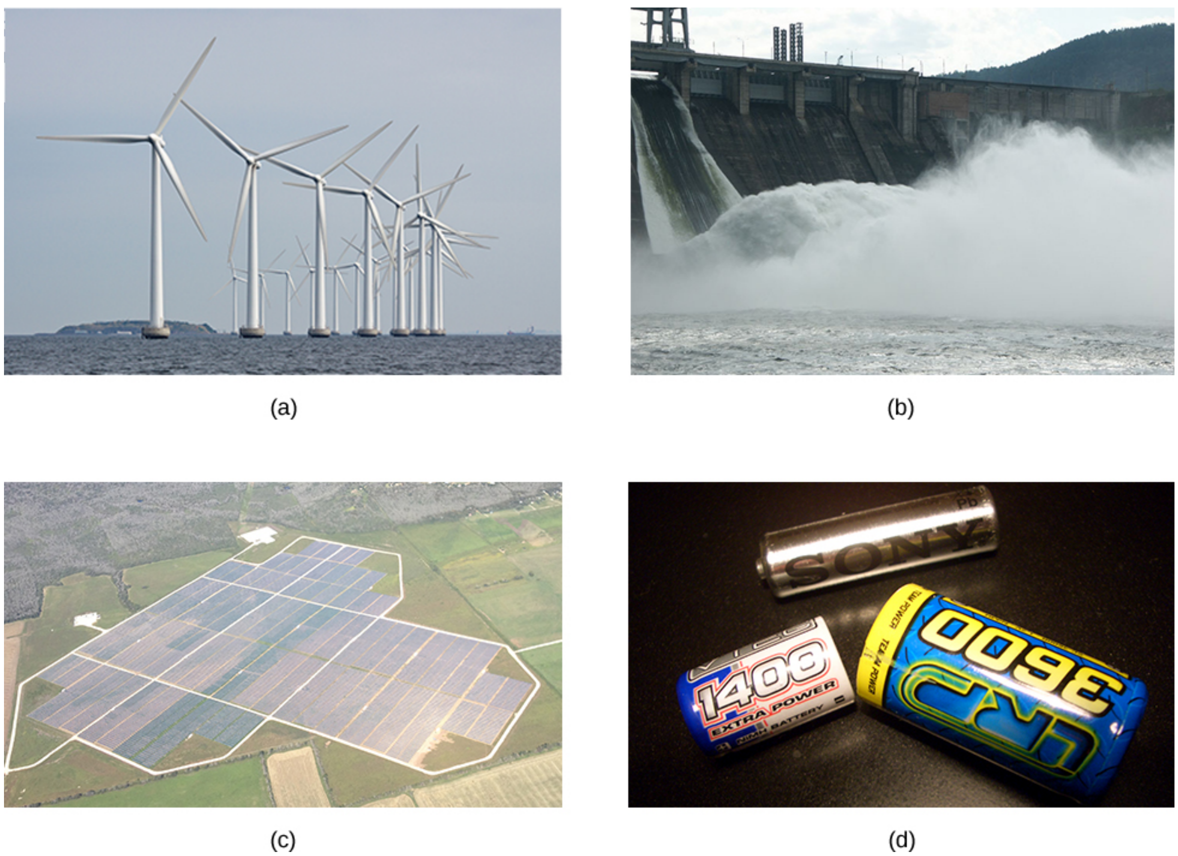


Figure $6.2.1$: A variety of voltage sources. (a) The Brazos Wind Farm in Fluvanna, Texas; (b) the Krasnoyarsk Dam in Russia; (c) a solar farm; (d) a group of nickel metal hydride batteries. The voltage output of each device depends on its construction and load. The voltage output equals source voltage only if there is no load. (credit a: modification of work by "Leaflet"/Wikimedia Commons; credit b: modification of work by Alex Polezhaev; credit c: modification of work by US Department of Energy; credit d: modification of work by Tiaa Monto)

Consider a simple circuit of a 12-V lamp attached to a 12-V battery, as shown in Figure $6.2.2$. The **battery** can be modeled as a two-terminal device that keeps one terminal at a higher electric potential than the second terminal. The higher electric potential is sometimes called the positive terminal and is labeled with a plus sign. The lower-potential terminal is sometimes called the negative terminal and labeled with a minus sign.

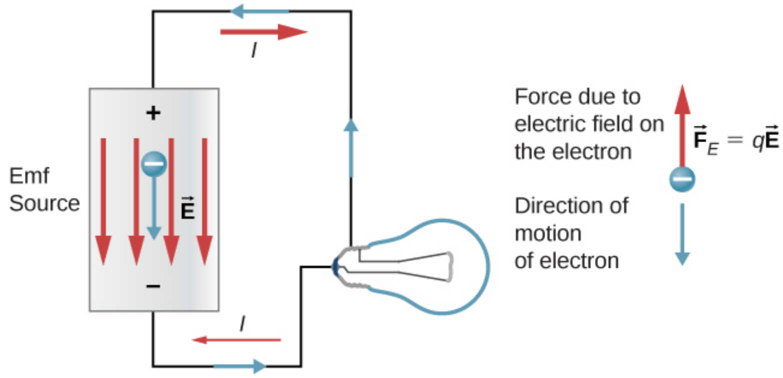


Figure 6.2.2: A battery maintains one terminal at a higher electric potential than the other terminal, acting as a source of current in a circuit.

When the battery is not connected to the lamp, there is no net flow of charge within the battery. Once the battery is connected to the lamp, charges flow from one terminal of the battery, through the lamp (causing the lamp to light), and back to the other terminal of the battery. If we consider positive (conventional) current flow, positive charges leave the positive terminal, travel through the lamp, and enter the negative terminal.

Positive current flow is useful for most of the circuit analysis in this chapter, but in metallic wires and resistors, electrons contribute the most to current, flowing in the opposite direction of positive current flow. Therefore, it is more realistic to consider the movement of electrons for the analysis of the circuit in Figure 6.2.2. The electrons leave the negative terminal, travel through the lamp, and return to the positive terminal. For the battery to maintain the potential difference between the two terminals, negative charges (electrons) must be moved from the positive terminal to the negative terminal. The battery acts as a charge pump, moving negative charges from the positive terminal to the negative terminal to maintain the potential difference. This increases the potential energy of the charges and, therefore, the electric potential of the charges.

The force on the negative charge from the electric field is in the opposite direction of the electric field, as shown in Figure 6.2.2. For the negative charges to be moved to the negative terminal, work must be done on the negative charges. This requires energy, which comes from chemical reactions in the battery. The potential is kept high on the positive terminal and low on the negative terminal to maintain the potential difference between the two terminals. The source voltage is equal to the work done on the charge per unit charge $\left(\varepsilon = \frac{dW}{dq}\right)$ when there is no current flowing. Since the unit for work is the joule and the unit for charge is the coulomb, the unit for source voltage is the volt $(1\,V = 1\,J/C)$.

The **terminal voltage** $V_{terminal}$ of a battery is the voltage measured across the terminals of the battery when there is no load connected to the terminal. An ideal battery maintains a constant terminal voltage, independent of the current between the two terminals. An ideal battery has no internal resistance, and the terminal voltage is equal to the source voltage of the battery. In the next section, we will show that a real battery does have internal resistance and the terminal voltage is always less than the source voltage of the battery.

The Origin of Battery Potential

The first battery was invented in 1799 by Alessandro Volta and was also known as the **voltaic pile**. Modern batteries have improved considerably since those early times, but often still retain the concept of a series of cells, each of which creates a potential difference. The combination of chemicals and the makeup of the terminals in a battery determine its source voltage. The **lead acid battery** used in cars and other vehicles is one of the most common combinations of chemicals. Figure 6.2.3 shows a single cell (one of six) of this battery. The cathode (positive) terminal of the cell is connected to a lead oxide plate, whereas the anode (negative) terminal is connected to a lead plate. Both plates are immersed in sulfuric acid, the electrolyte for the system.

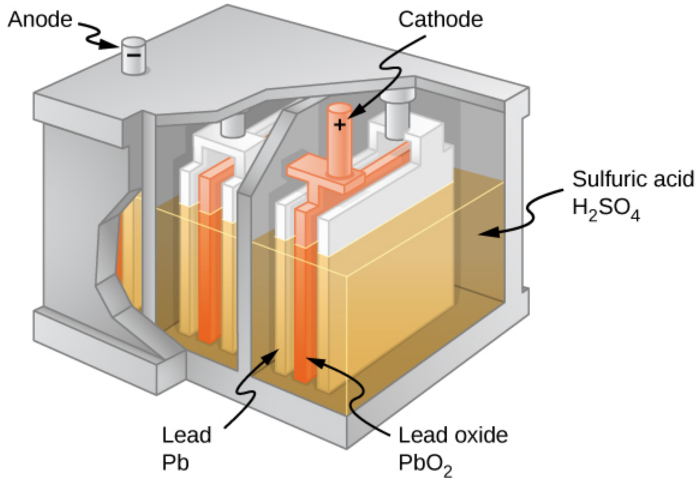


Figure 6.2.3: Chemical reactions in a lead-acid cell separate charge, sending negative charge to the anode, which is connected to the lead plates. The lead oxide plates are connected to the positive or cathode terminal of the cell. Sulfuric acid conducts the charge, as well as participates in the chemical reaction.

Knowing a little about how the chemicals in a lead-acid battery interact helps in understanding the potential created by the battery. Figure 6.2.4 shows the result of a single chemical reaction. Two electrons are placed on the **anode**, making it negative, provided that the **cathode** supplies two electrons. This leaves the cathode positively charged, because it has lost two electrons. In short, a separation of charge has been driven by a chemical reaction.

Note that the reaction does not take place unless there is a complete circuit to allow two electrons to be supplied to the cathode. Under many circumstances, these electrons come from the anode, flow through a resistance, and return to the cathode. Note also that since the chemical reactions involve substances with resistance, it is not possible to create the source voltage without an internal resistance.

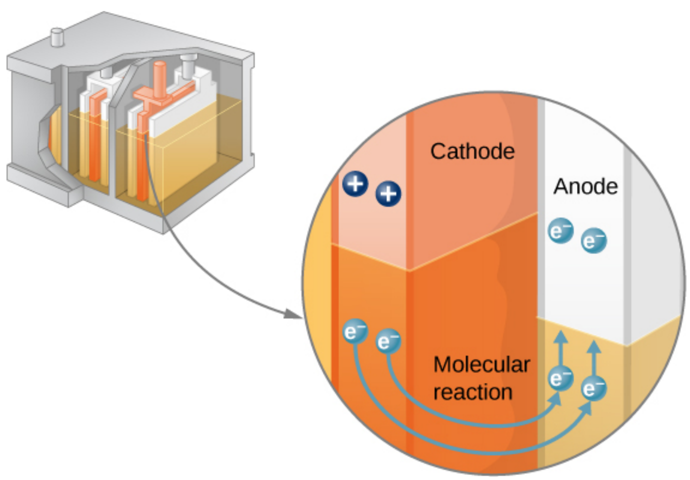


Figure 6.2.4: In a lead-acid battery, two electrons are forced onto the anode of a cell, and two electrons are removed from the cathode of the cell. The chemical reaction in a lead-acid battery places two electrons on the anode and removes two from the cathode. It requires a closed circuit to proceed, since the two electrons must be supplied to the cathode.

Internal Resistance and Terminal Voltage

The amount of resistance to the flow of current within the voltage source is called the **internal resistance**. The internal resistance r of a battery can behave in complex ways. It generally increases as a battery is depleted, due to the oxidation of the plates or the reduction of the acidity of the electrolyte. However, internal resistance may also depend on the magnitude and direction of the current through a voltage source, its temperature, and even its history. The internal resistance of rechargeable nickel-cadmium cells, for example, depends on how many times and how deeply they have been depleted. A simple model for a battery consists of an idealized source voltage ε and an internal resistance r (Figure 6.2.5).

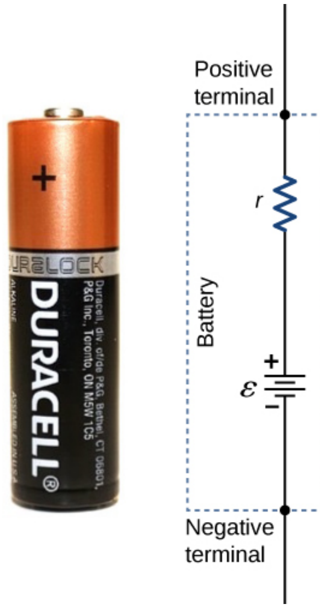


Figure 6.2.5: A battery can be modeled as an idealized source voltage (ε) with an internal resistance (**r**). The terminal voltage of the battery is $V_{terminal} = \varepsilon - Ir$.

Suppose an external resistor, known as the load resistance R, is connected to a voltage source such as a battery, as in Figure 6.2.6. The figure shows a model of a battery with a source voltage ε, an internal resistance r, and a load resistor R connected across its terminals. Using conventional current flow, positive charges leave the positive terminal of the battery, travel through the resistor, and return to the negative terminal of the battery. The terminal voltage of the battery depends on the source voltage, the internal resistance, and the current, and is equal to

Note

$$V_{terminal} = \varepsilon - Ir \tag{6.2.1}$$

For a given source voltage and internal resistance, the terminal voltage decreases as the current increases due to the potential drop Ir of the internal resistance.

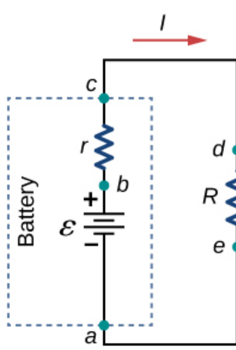


Figure 6.2.6: Schematic of a voltage source and its load resistor R. Since the internal resistance r is in series with the load, it can significantly affect the terminal voltage and the current delivered to the load.

A graph of the potential difference across each element the circuit is shown in Figure 6.2.7. A current I runs through the circuit, and the potential drop across the internal resistor is equal to Ir. The terminal voltage is equal to $\varepsilon - Ir$, which is equal to the **potential drop** across the load resistor $IR = \varepsilon - Ir$. As with potential energy, it is the change in voltage that is important. When the term "voltage" is used, we assume that it is actually the change in the potential, or ΔV. However, Δ is often omitted for convenience.

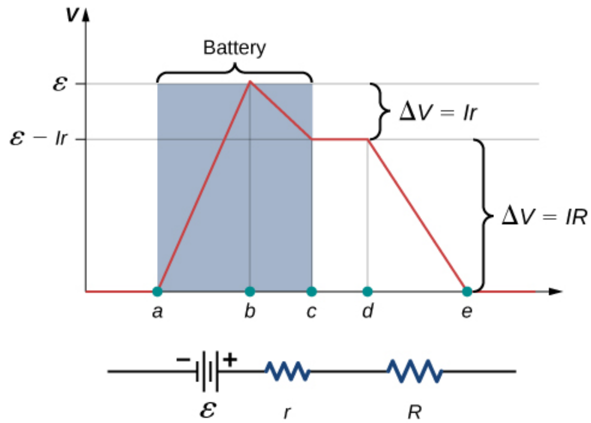


Figure 6.2.7: A graph of the voltage through the circuit of a battery and a load resistance. The electric potential increases due to the source voltage of the battery from the chemical reactions doing work on the charges. There is a decrease in the electric potential in the battery due to the internal resistance. The potential decreases due to the internal resistance $-Ir$, making the terminal voltage of the battery equal to $(\varepsilon - Ir)$. The voltage then decreases across the resistor by IR. The current is equal to $I = \frac{\varepsilon}{r+R}$.

The current through the load resistor is $I = \frac{\varepsilon}{r+R}$. We see from this expression that the smaller the internal resistance r, the greater the current the voltage source supplies to its load R. As batteries are depleted, r increases. If r becomes a significant fraction of the load resistance, then the current is significantly reduced, as the following example illustrates.

Example 6.2.1: Analyzing a Circuit with a Battery and a Load

A given battery has a 12.00-V source voltage and an internal resistance of $0.100\,\Omega$. (a) Calculate its terminal voltage when connected to a $10.00\,\Omega$ load. (b) What is the terminal voltage when connected to a $0.500\,\Omega$ load? (c) What power does the $0.500\,\Omega$ load dissipate? (d) If the internal resistance grows to $0.500\,\Omega$, find the current, terminal voltage, and power dissipated by a $0.500\,\Omega$ load.

Strategy

The analysis above gave an expression for current when internal resistance is taken into account. Once the current is found, the terminal voltage can be calculated by using the equation $V_{terminal} = \varepsilon - Ir$. Once current is found, we can also find the power dissipated by the resistor.

Solution

1. Entering the given values for the source voltage, load resistance, and internal resistance into the expression above yields

$$I = \frac{\varepsilon}{R+r} = \frac{12.00\,V}{10.10\,\Omega} = 1.188\,A. \tag{6.2.2}$$

   Enter the known values into the equation $V_{terminal} = \varepsilon - Ir$ to get the terminal voltage:

$$V_{terminal} = \varepsilon - Ir = 12.00\,V - (1.188\,A)(0.100\,\Omega) = 11.90\,V. \tag{6.2.3}$$

   The terminal voltage here is only slightly lower than the source voltage, implying that the current drawn by this light load is not significant.

2. Similarly, with $R_{load} = 0.500\,\Omega$, the current is

$$I = \frac{\varepsilon}{R+r} = \frac{12.00\,V}{0.600\,\Omega} = 20.00\,A. \tag{6.2.4}$$

   The terminal voltage is now

$$V_{terminal} = \varepsilon - Ir = 12.00\,V - (20.00\,A)(0.100\,\Omega) = 10.00\,V. \tag{6.2.5}$$

   The terminal voltage exhibits a more significant reduction compared with source voltage, implying $0.500\,\Omega$ is a heavy load for this battery. A “heavy load” signifies a larger draw of current from the source but not a larger resistance.

3. The power dissipated by the $0.500\,\Omega$ load can be found using the formula $P = I^2R$. Entering the known values gives

$$P = I^2R = (20.0\,A)^2(0.500\,\Omega) = 2.00 \times 10^2\,W. \tag{6.2.6}$$

Note that this power can also be obtained using the expression $\frac{V^2}{R}$ or IV, where V is the terminal voltage (10.0 V in this case).

4. Here, the internal resistance has increased, perhaps due to the depletion of the battery, to the point where it is as great as the load resistance. As before, we first find the current by entering the known values into the expression, yielding

$$I = \frac{\varepsilon}{R+r} = \frac{12.00\,V}{1.00\,\Omega} = 12.00\,A. \tag{6.2.7}$$

Now the terminal voltage is

$$V_{terminal} = \varepsilon - Ir = 12.00\,V - (12.00\,A)(0.500\,\Omega) = 6.00\,V, \tag{6.2.8}$$

and the power dissipated by the load is

$$P = I^2R = (12.00\,A)^2(0.500\,\Omega) = 72.00\,W. \tag{6.2.9}$$

We see that the increased internal resistance has significantly decreased the terminal voltage, current, and power delivered to a load.

Significance

The internal resistance of a battery can increase for many reasons. For example, the internal resistance of a rechargeable battery increases as the number of times the battery is recharged increases. The increased internal resistance may have two effects on the battery. First, the terminal voltage will decrease. Second, the battery may overheat due to the increased power dissipated by the internal resistance.

Exercise 6.2.1

If you place a wire directly across the two terminal of a battery, effectively shorting out the terminals, the battery will begin to get hot. Why do you suppose this happens?

Solution

If a wire is connected across the terminals, the load resistance is close to zero, or at least considerably less than the internal resistance of the battery. Since the internal resistance is small, the current through the circuit will be large, $I = \frac{\varepsilon}{R+r} = \frac{\varepsilon}{0+r} = \frac{\varepsilon}{r}$. The large current causes a high power to be dissipated by the internal resistance $(P = I^2r)$. The power is dissipated as heat.

Battery Testers

Battery testers, such as those in Figure $6.2.8$, use small load resistors to intentionally draw current to determine whether the terminal potential drops below an acceptable level. Although it is difficult to measure the internal resistance of a battery, battery testers can provide a measurement of the internal resistance of the battery. If internal resistance is high, the battery is weak, as evidenced by its low terminal voltage.

(a)

(b)

Figure 6.2.8: Battery testers measure terminal voltage under a load to determine the condition of a battery. (a) A US Navy electronics technician uses a battery tester to test large batteries aboard the aircraft carrier USS Nimitz. The battery tester she uses has a small resistance that can dissipate large amounts of power. (b) The small device shown is used on small batteries and has a digital display to indicate the acceptability of the terminal voltage. (credit a: modification of work by Jason A. Johnston; credit b: modification of work by Keith Williamson)

Some batteries can be recharged by passing a current through them in the direction opposite to the current they supply to an appliance. This is done routinely in cars and in batteries for small electrical appliances and electronic devices (Figure 6.2.9). The voltage output of the battery charger must be greater than the source voltage of the battery to reverse the current through it. This causes the terminal voltage of the battery to be greater than the source voltage, since $V = \varepsilon - Ir$ and I is now negative.

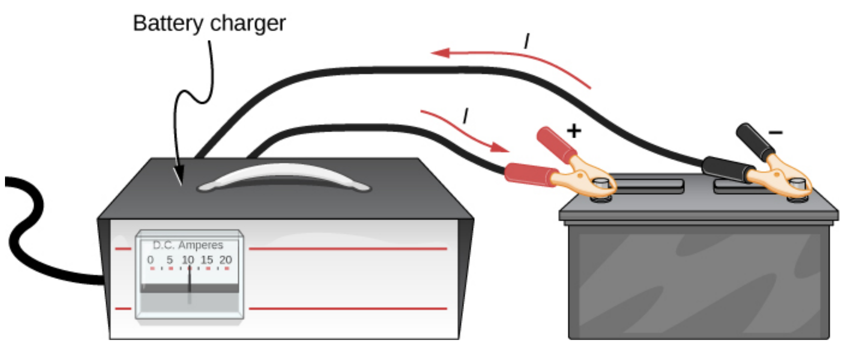


Figure 6.2.9: A car battery charger reverses the normal direction of current through a battery, reversing its chemical reaction and replenishing its chemical potential.

It is important to understand the consequences of the internal resistance of devices which provide source voltage, such as batteries and solar cells, but often, the analysis of circuits is done with the terminal voltage of the battery, as we have done in the previous sections. The terminal voltage is referred to as simply as V, dropping the subscript "terminal." This is because the internal resistance of the battery is difficult to measure directly and can change over time.

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).

6.3: Resistors in Series and Parallel

Learning Objectives

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

- Define the term equivalent resistance
- Calculate the equivalent resistance of resistors connected in series
- Calculate the equivalent resistance of resistors connected in parallel

In Current and Resistance, we described the term 'resistance' and explained the basic design of a resistor. Basically, a resistor limits the flow of charge in a circuit and is an ohmic device where $V = IR$. Most circuits have more than one resistor. If several resistors are connected together and connected to a battery, the current supplied by the battery depends on the **equivalent resistance** of the circuit.

The equivalent resistance of a combination of resistors depends on both their individual values and how they are connected. The simplest combinations of resistors are series and parallel connections (Figure $6.3.1$). In a **series circuit**, the output current of the first resistor flows into the input of the second resistor; therefore, the current is the same in each resistor. In a **parallel circuit**, all of the resistor leads on one side of the resistors are connected together and all the leads on the other side are connected together. In the case of a parallel configuration, each resistor has the same potential drop across it, and the currents through each resistor may be different, depending on the resistor. The sum of the individual currents equals the current that flows into the parallel connections.

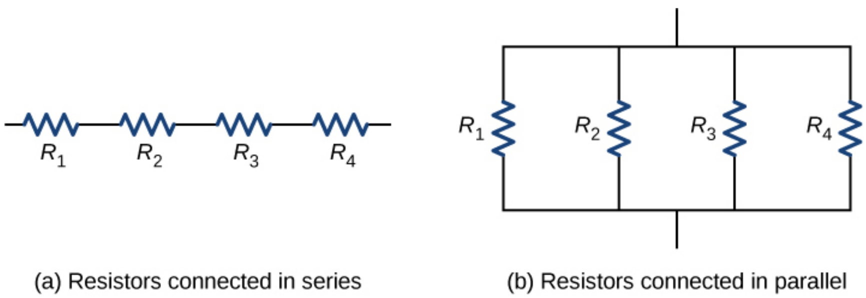


Figure $6.3.1$: (a) For a series connection of resistors, the current is the same in each resistor. (b) For a parallel connection of resistors, the voltage is the same across each resistor.

Resistors in Series

Resistors are said to be in series whenever the current flows through the resistors sequentially. Consider Figure $6.3.2$, which shows three resistors in series with an applied voltage equal to V_{ab}. Since there is only one path for the charges to flow through, the current is the same through each resistor. The equivalent resistance of a set of resistors in a series connection is equal to the algebraic sum of the individual resistances.

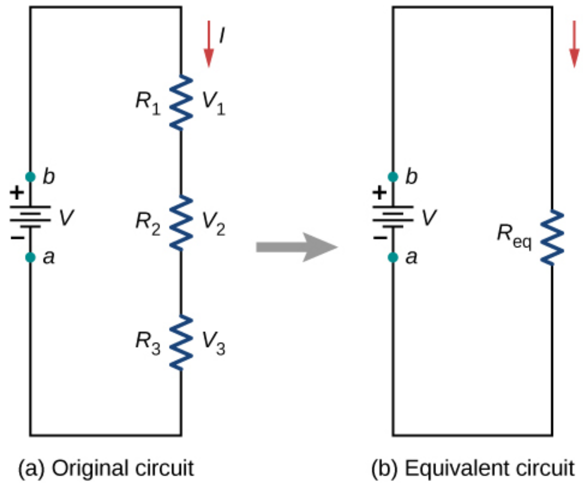


Figure 6.3.2: (a) Three resistors connected in series to a voltage source. (b) The original circuit is reduced to an equivalent resistance and a voltage source.

In Figure $6.3.2$, the current coming from the voltage source flows through each resistor, so the current through each resistor is the same. The current through the circuit depends on the voltage supplied by the voltage source and the resistance of the resistors. For each resistor, a potential drop occurs that is equal to the loss of electric potential energy as a current travels through each resistor. According to Ohm's law, the potential drop V across a resistor when a current flows through it is calculated using the equation $V = IR$, where I is the current in amps (A) and R is the resistance in ohms (Ω). Since energy is conserved, and the voltage is equal to the potential energy per charge, the sum of the voltage applied to the circuit by the source and the potential drops across the individual resistors around a loop should be equal to zero:

$$\sum_{i=1}^{N} V_i = 0. \tag{6.3.1}$$

This equation is often referred to as Kirchhoff's loop law, which we will look at in more detail later in this chapter. For Figure $6.3.2$, the sum of the potential drop of each resistor and the voltage supplied by the voltage source should equal zero:

$$\begin{aligned} V - V_1 - V_2 - V_3 &= 0, \\ V &= V_1 + V_2 + V_3, \\ &= IR_1 + IR_2 + IR_3, \end{aligned}$$

Solving for I

$$\begin{aligned} I &= \frac{V}{R_1 + R_2 + R_3} \\ &= \frac{V}{R_S}. \end{aligned}$$

Since the current through each component is the same, the equality can be simplified to an equivalent resistance (R_S), which is just the sum of the resistances of the individual resistors.

✔ Equivalent Resistance in Series Circuits

Any number of resistors can be connected in series. If N resistors are connected in series, the **equivalent resistance** is

$$R_S = R_1 + R_2 + R_3 + \ldots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i. \tag{6.3.2}$$

One result of components connected in a series circuit is that if something happens to one component, it affects all the other components. For example, if several lamps are connected in series and one bulb burns out, all the other lamps go dark.

Example 6.3.1: Equivalent Resistance, Current, and Power in a Series Circuit

A battery with a terminal voltage of 9 V is connected to a circuit consisting of four $20\,\Omega$ and one $10\,\Omega$ resistors all in series (Figure 6.3.3). Assume the battery has negligible internal resistance.

a. Calculate the equivalent resistance of the circuit.
b. Calculate the current through each resistor.
c. Calculate the potential drop across each resistor.
d. Determine the total power dissipated by the resistors and the power supplied by the battery.

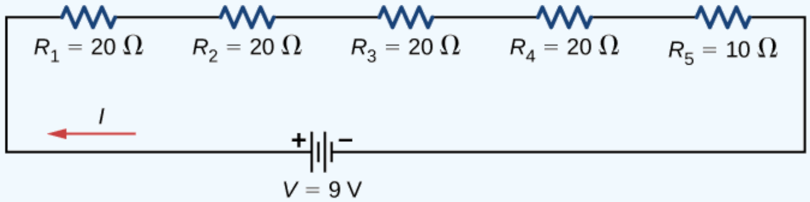


Figure 6.3.3: A simple series circuit with five resistors.

Strategy

In a series circuit, the equivalent resistance is the algebraic sum of the resistances. The current through the circuit can be found from Ohm's law and is equal to the voltage divided by the equivalent resistance. The potential drop across each resistor can be found using Ohm's law. The power dissipated by each resistor can be found using $P = I^2R$, and the total power dissipated by the resistors is equal to the sum of the power dissipated by each resistor. The power supplied by the battery can be found using $P = I\epsilon$.

Solution

1. The equivalent resistance is the algebraic sum of the resistances (Equation 6.3.2):

$$\begin{aligned} R_S &= R_1 + R_2 + R_3 + R_4 + R_5 \\ &= 20\,\Omega + 20\,\Omega + 20\,\Omega + 20\,\Omega + 10\,\Omega = 90\,\Omega. \end{aligned}$$

2. The current through the circuit is the same for each resistor in a series circuit and is equal to the applied voltage divided by the equivalent resistance:

$$I = \frac{V}{R_S} = \frac{9\,V}{90\,\Omega} = 0.1\,A.$$

Note that the sum of the potential drops across each resistor is equal to the voltage supplied by the battery.

3. The power dissipated by a resistor is equal to $P = I^2R$, and the power supplied by the battery is equal to $P = I\epsilon$.

$$P_1 = P_2 = P_3 = P_4 = (0.1\,A)^2(20\,\Omega) = 0.2\,W,$$

$$P_5 = (0.1\,A)^2(10\,\Omega) = 0.1\,W,$$

$$P_{dissipated} = 0.2\,W + 0.2\,W + 0.2\,W + 0.2\,W + 0.1\,W = 0.9\,W,$$

$$P_{source} = I\epsilon = (0.1\,A)(9\,V) = 0.9\,W.$$

Significance

There are several reasons why we would use multiple resistors instead of just one resistor with a resistance equal to the equivalent resistance of the circuit. Perhaps a resistor of the required size is not available, or we need to dissipate the heat generated, or we want to minimize the cost of resistors. Each resistor may cost a few cents to a few dollars, but when multiplied by thousands of units, the cost saving may be appreciable.

Exercise 6.3.1

Some strings of miniature holiday lights are made to short out when a bulb burns out. The device that causes the short is called a shunt, which allows current to flow around the open circuit. A "short" is like putting a piece of wire across the component. The bulbs are usually grouped in series of nine bulbs. If too many bulbs burn out, the shunts eventually open. What causes this?

Answer

The equivalent resistance of nine bulbs connected in series is 9**R**. The current is $I = V/9\,R$. If one bulb burns out, the equivalent resistance is 8**R**, and the voltage does not change, but the current increases $(I = V/8\,R$. As more bulbs burn out, the current becomes even higher. Eventually, the current becomes too high, burning out the shunt.

Let's briefly summarize the major features of resistors in series:

1. Series resistances add together to get the equivalent resistance (Equation 6.3.2):

$$R_S = R_1 + R_2 + R_3 + \ldots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i. \tag{6.3.3}$$

2. The same current flows through each resistor in series.
3. Individual resistors in series do not get the total source voltage, but divide it. The total potential drop across a series configuration of resistors is equal to the sum of the potential drops across each resistor.

Resistors in Parallel

Figure 6.3.4 shows resistors in parallel, wired to a voltage source. Resistors are in parallel when one end of all the resistors are connected by a continuous wire of negligible resistance and the other end of all the resistors are also connected to one another through a continuous wire of negligible resistance. The potential drop across each resistor is the same. Current through each resistor can be found using Ohm's law $I = V/R$, where the voltage is constant across each resistor. For example, an automobile's headlights, radio, and other systems are wired in parallel, so that each subsystem utilizes the full voltage of the source and can operate completely independently. The same is true of the wiring in your house or any building.

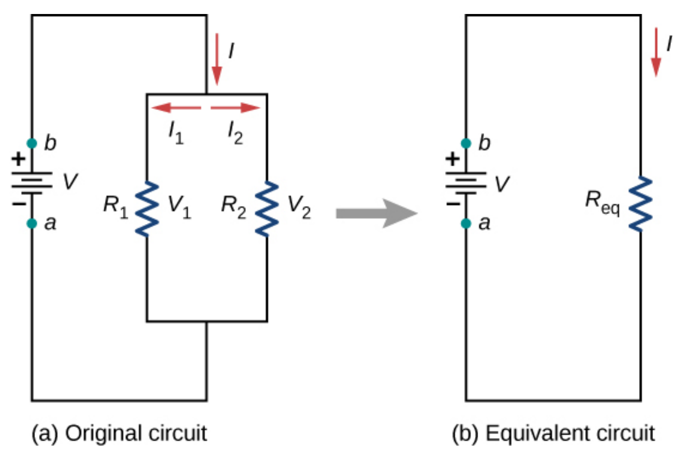


Figure 6.3.4: Two resistors connected in parallel to a voltage source. (b) The original circuit is reduced to an equivalent resistance and a voltage source.

The current flowing from the voltage source in Figure 6.3.4 depends on the voltage supplied by the voltage source and the equivalent resistance of the circuit. In this case, the current flows from the voltage source and enters a junction, or node, where the circuit splits flowing through resistors R_1 and R_2. As the charges flow from the battery, some go through resistor R_1 and some flow through resistor R_2. The sum of the currents flowing into a junction must be equal to the sum of the currents flowing out of the junction:

$$\sum I_{in} = \sum I_{out}.$$

This equation is referred to as **Kirchhoff's junction rule** and will be discussed in detail in the next section. In Figure $6.3.4$, the junction rule gives $I = I_1 + I_2$. There are two loops in this circuit, which leads to the equations $V = I_1 R_1$ and $I_1 R_1 = I_2 R_2$. Note the voltage across the resistors in parallel are the same ($V = V_1 = V_2$) and the current is additive:

$$\begin{aligned} I &= I_1 + I_2 \\ &= \frac{V_1}{R_1} + \frac{V_2}{R_2} \\ &= \frac{V}{R_1} + \frac{V}{R_2} \\ &= V\left(\frac{1}{R_1} + \frac{1}{R_2}\right) = \frac{V}{R_P} \end{aligned}$$

Solving for the R_P

$$R_P = \left(\frac{1}{R_1} + \frac{1}{R_2}\right)^{-1}. \tag{6.3.4}$$

Equivalent Resistance in Parallel Circuits

Generalizing to any number of N resistors, the equivalent resistance R_P of a parallel connection is related to the individual resistances by

$$R_P = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots + \frac{1}{R_{N-1}} + \frac{1}{R_N}\right)^{-1} = \left(\sum_{i=1}^{N} \frac{1}{R_i}\right)^{-1}. \tag{6.3.5}$$

This relationship results in an equivalent resistance R_P that is less than the smallest of the individual resistances. When resistors are connected in parallel, more current flows from the source than would flow for any of them individually, so the total resistance is lower.

Example 6.3.2: Analysis of a parallel circuit

Three resistors $R_1 = 1.00\,\Omega$, $R_2 = 2.00\,\Omega$, and $R_3 = 2.00\,\Omega$, are connected in parallel. The parallel connection is attached to a $V = 3.00\,V$ voltage source.

a. What is the equivalent resistance?
b. Find the current supplied by the source to the parallel circuit.
c. Calculate the currents in each resistor and show that these add together to equal the current output of the source.
d. Calculate the power dissipated by each resistor.
e. Find the power output of the source and show that it equals the total power dissipated by the resistors.

Strategy

(a) The total resistance for a parallel combination of resistors is found using Equation 6.3.5. (Note that in these calculations, each intermediate answer is shown with an extra digit.)

(b) The current supplied by the source can be found from Ohm's law, substituting R_P for the total resistance $I = \frac{V}{R_P}$.

(c) The individual currents are easily calculated from Ohm's law $\left(I_i = \frac{V_i}{R_i}\right)$, since each resistor gets the full voltage. The total current is the sum of the individual currents:

$$I = \sum_i I_i.$$

(d) The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P_i = V^2/R_i$, since each resistor gets full voltage.

(e) The total power can also be calculated in several ways, use $P = IV$.

Solution

1. The total resistance for a parallel combination of resistors is found using Equation 6.3.5. Entering known values gives

$$R_P = \left(\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}\right)^{-1} = \left(\frac{1}{1.00\,\Omega} + \frac{1}{2.00\,\Omega} + \frac{1}{2.00\,\Omega}\right)^{-1} = 0.50\,\Omega.$$

   The total resistance with the correct number of significant digits is $R_{eq} = 0.50\,\Omega$. As predicted, R_P is less than the smallest individual resistance.
2. The total current can be found from Ohm's law, substituting R_P for the total resistance. This gives

$$I = \frac{V}{R_P} = \frac{3.00\,V}{0.50\,\Omega} = 6.00\,A.$$

   Current **I** for each device is much larger than for the same devices connected in series (see the previous example). A circuit with parallel connections has a smaller total resistance than the resistors connected in series.
3. The individual currents are easily calculated from Ohm's law, since each resistor gets the full voltage. Thus,

$$I_1 = \frac{V}{R_1} = \frac{3.00\,V}{1.00\,\Omega} = 3.00\,A.$$

   Similarly,

$$I_2 = \frac{V}{R_2} = \frac{3.00\,V}{2.00\,\Omega} = 1.50\,A$$

   and

$$I_3 = \frac{V}{R_3} = \frac{3.00\,V}{2.00\,\Omega} = 1.50\,A.$$

   The total current is the sum of the individual currents:

$$I_1 + I_2 + I_3 = 6.00\,A.$$

4. The power dissipated by each resistor can be found using any of the equations relating power to current, voltage, and resistance, since all three are known. Let us use $P = V^2/R$, since each resistor gets full voltage. Thus,

$$P_1 = \frac{V^2}{R_1} = \frac{(3.00\,V)^2}{1.00\,\Omega} = 9.00\,W.$$

   Similarly,

$$P_2 = \frac{V^2}{R_2} = \frac{(3.00\,V)^2}{2.00\,\Omega} = 4.50\,W.$$

   and

$$P_3 = \frac{V^2}{R_3} = \frac{(3.00\,V)^2}{2.00\,\Omega} = 4.50\,W.$$

5. The total power can also be calculated in several ways. Choosing $P = IV$ and entering the total current yields

$$P = IV = (6.00\,A)(3.00\,V) = 18.00\,W.$$

Significance

Total power dissipated by the resistors is also 18.00 W:

$$P_1 + P_2 + P_3 = 9.00\,W + 4.50\,W + 4.50\,W = 18.00\,W.$$

Notice that the total power dissipated by the resistors equals the power supplied by the source.

? Exercise $6.3.2A$

Consider the same potential difference $(V = 3.00\,V)$ applied to the same three resistors connected in series. Would the equivalent resistance of the series circuit be higher, lower, or equal to the three resistor in parallel? Would the current through the series circuit be higher, lower, or equal to the current provided by the same voltage applied to the parallel circuit? How would the power dissipated by the resistor in series compare to the power dissipated by the resistors in parallel?

Solution

The equivalent of the series circuit would be $R_{eq} = 1.00\,\Omega + 2.00\,\Omega + 2.00\,\Omega = 5.00\,\Omega$, which is higher than the equivalent resistance of the parallel circuit $R_{eq} = 0.50\,\Omega$. The equivalent resistor of any number of resistors is always higher than the equivalent resistance of the same resistors connected in parallel. The current through for the series circuit would be $I = \frac{3.00\,V}{5.00\,\Omega} = 0.60\,A$, which is lower than the sum of the currents through each resistor in the parallel circuit, $I = 6.00\,A$. This is not surprising since the equivalent resistance of the series circuit is higher. The current through a series connection of any number of resistors will always be lower than the current into a parallel connection of the same resistors, since the equivalent resistance of the series circuit will be higher than the parallel circuit. The power dissipated by the resistors in series would be $P = 1.800\,W$, which is lower than the power dissipated in the parallel circuit $P = 18.00\,W$.

? Exercise $6.3.2B$

How would you use a river and two waterfalls to model a parallel configuration of two resistors? How does this analogy break down?

Solution

A river, flowing horizontally at a constant rate, splits in two and flows over two waterfalls. The water molecules are analogous to the electrons in the parallel circuits. The number of water molecules that flow in the river and falls must be equal to the number of molecules that flow over each waterfall, just like sum of the current through each resistor must be equal to the current flowing into the parallel circuit. The water molecules in the river have energy due to their motion and height. The potential energy of the water molecules in the river is constant due to their equal heights. This is analogous to the constant change in voltage across a parallel circuit. Voltage is the potential energy across each resistor.

The analogy quickly breaks down when considering the energy. In the waterfall, the potential energy is converted into kinetic energy of the water molecules. In the case of electrons flowing through a resistor, the potential drop is converted into heat and light, not into the kinetic energy of the electrons.

Let us summarize the major features of resistors in parallel:

1. Equivalent resistance is found from Equation $6.3.5$ and is smaller than any individual resistance in the combination.
2. The potential drop across each resistor in parallel is the same.
3. Parallel resistors do not each get the total current; they divide it. The current entering a parallel combination of resistors is equal to the sum of the current through each resistor in parallel.

Combinations of Series and Parallel

More complex connections of resistors are often just combinations of series and parallel connections. Such combinations are common, especially when wire resistance is considered. In that case, wire resistance is in series with other resistances that are in parallel.

Combinations of series and parallel can be reduced to a single equivalent resistance using the technique illustrated in Figure $6.3.5$. Various parts can be identified as either series or parallel connections, reduced to their equivalent resistances, and then further reduced until a single equivalent resistance is left. The process is more time consuming than difficult. Here, we note the equivalent resistance as R_{eq}.

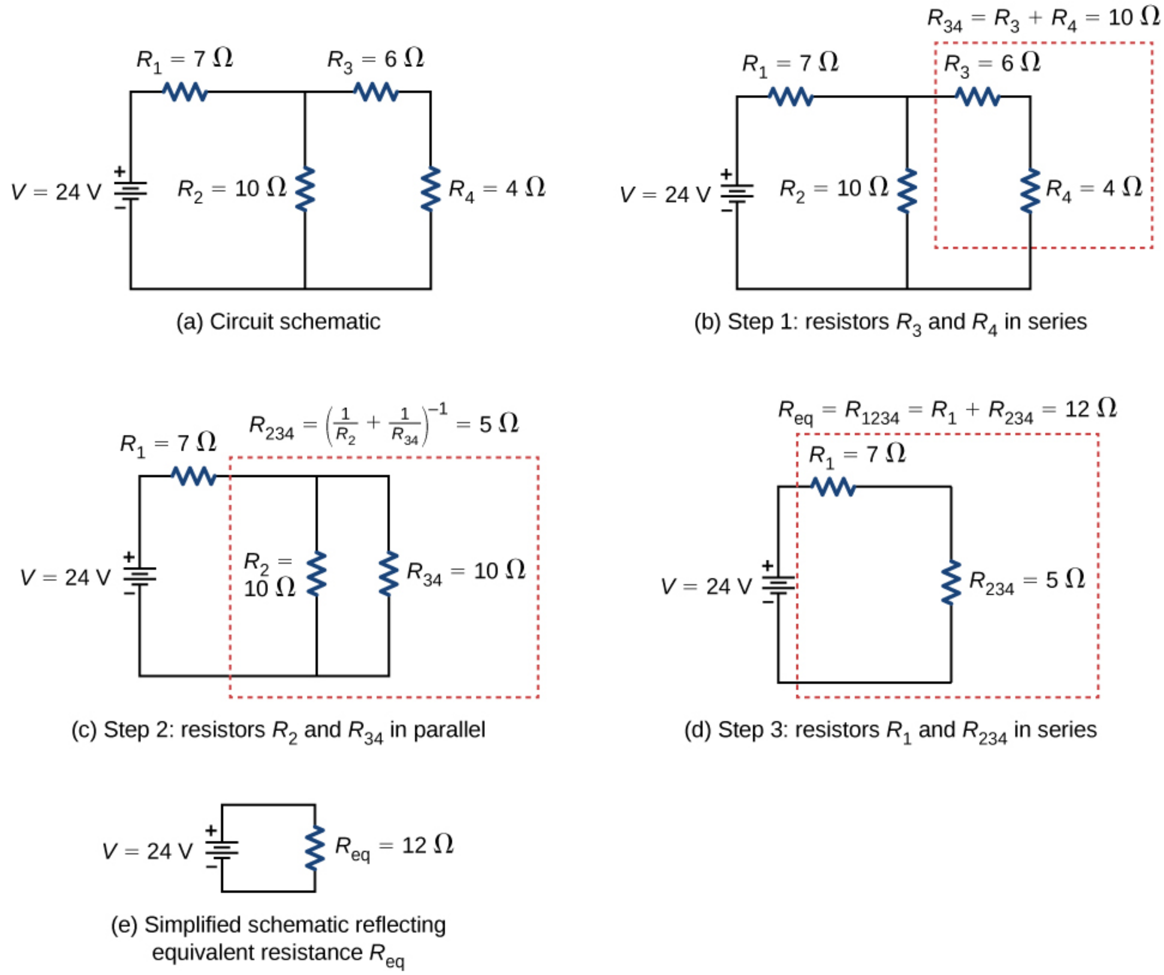


Figure 6.3.5: (a) The original circuit of four resistors. (b) Step 1: The resistors R_3 and R_4 are in series and the equivalent resistance is $R_{34} = 10\,\Omega$ (c) Step 2: The reduced circuit shows resistors R_2 and R_{34} are in parallel, with an equivalent resistance of $R_{234} = 5\,\Omega$. (d) Step 3: The reduced circuit shows that R_1 and R_{234} are in series with an equivalent resistance of $R_{1234} = 12\,\Omega$ which is the equivalent resistance R_{eq}. (e) The reduced circuit with a voltage source of $V = 24\,V$ with an equivalent resistance of $R_{eq} = 12\,\Omega$. This results in a current of $I = 2\,A$ from the voltage source.

Notice that resistors R_3 and R_4 are in series. They can be combined into a single equivalent resistance. One method of keeping track of the process is to include the resistors as subscripts. Here the equivalent resistance of R_3 and R_4 is

$$R_{34} = R_3 + R_4 = 6\,\Omega + 4\,\Omega = 10\,\Omega.$$

The circuit now reduces to three resistors, shown in Figure $6.3.5c$. Redrawing, we now see that resistors R_2 and R_{34} constitute a parallel circuit. Those two resistors can be reduced to an equivalent resistance:

$$R_{234} = \left(\frac{1}{R_2} + \frac{1}{R_{34}}\right)^{-1} = \left(\frac{1}{10\,\Omega} + \frac{1}{10\,\Omega}\right)^{-1} = 5\,\Omega.$$

This step of the process reduces the circuit to two resistors, shown in in Figure $6.3.5d$. Here, the circuit reduces to two resistors, which in this case are in series. These two resistors can be reduced to an equivalent resistance, which is the equivalent resistance of the circuit:

$$R_{eq} = R_{1234} = R_1 + R_{234} = 7\,\Omega + 5\Omega = 12\,\Omega.$$

The main goal of this circuit analysis is reached, and the circuit is now reduced to a single resistor and single voltage source.

Now we can analyze the circuit. The current provided by the voltage source is $I = \frac{V}{R_{eq}} = \frac{24\,V}{12\,\Omega} = 2\,A$. This current runs through resistor R_1 and is designated as I_1. The potential drop across R_1 can be found using Ohm's law:

$$V_1 = I_1 R_1 = (2\,A)(7\,\Omega) = 14\,V.$$

Looking at Figure $6.3.5c$, this leaves $24\,V - 14\,V = 10\,V$ to be dropped across the parallel combination of R_2 and R_{34}. The current through R_2 can be found using Ohm's law:

$$I_2 = \frac{V_2}{R_2} = \frac{10\,V}{10\,\Omega} = 1\,A.$$

The resistors R_3 and R_4 are in series so the currents I_3 and I_4 are equal to

$$I_3 = I_4 = I - I_2 = 2\,A - 1\,A = 1\,A.$$

Using Ohm's law, we can find the potential drop across the last two resistors. The potential drops are $V_3 = I_3 R_3 = 6\,V$ and $V_4 = I_4 R_4 = 4\,V$. The final analysis is to look at the power supplied by the voltage source and the power dissipated by the resistors. The power dissipated by the resistors is

$$\begin{aligned} P_1 &= I_1^2 R_1 = (2\,A)^2 (7\,\Omega) = 28\,W, \\ P_2 &= I_2^2 R_2 = (1\,A)^2 (10\,\Omega) = 10\,W, \\ P_3 &= I_3^2 R_3 = (1\,A)^2 (6\,\Omega) = 6\,W, \\ P_4 &= I_4^2 R_4 = (1\,A)^2 (4\,\Omega) = 4\,W, \\ P_{dissipated} &= P_1 + P_2 + P_3 + P_4 = 48\,W. \end{aligned}$$

The total energy is constant in any process. Therefore, the power supplied by the voltage source is

$$\begin{aligned} P_s &= IV \\ &= (2\,A)(24\,V) = 48\,W \end{aligned}$$

Analyzing the power supplied to the circuit and the power dissipated by the resistors is a good check for the validity of the analysis; they should be equal.

✓ Example 6.3.3: Combining Series and parallel circuits

Figure $6.3.6$ shows resistors wired in a combination of series and parallel. We can consider R_1 to be the resistance of wires leading to R_2 and R_3.

a. Find the equivalent resistance of the circuit.
b. What is the potential drop V_1 across resistor R_1?
c. Find the current I_2 through resistor R_2.
d. What power is dissipated by R_2?

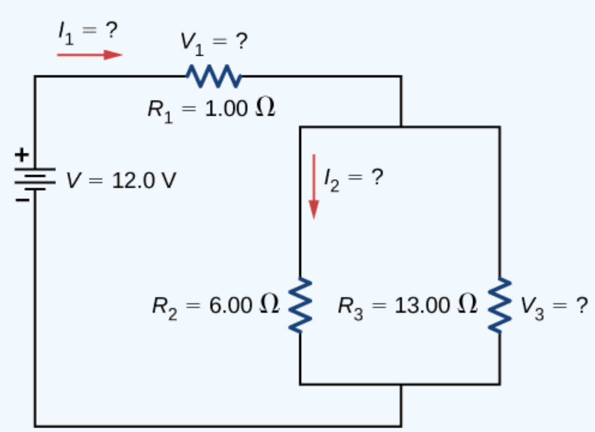


Figure $6.3.6$: These three resistors are connected to a voltage source so that R_2 and R_3 are in parallel with one another and that combination is in series with R_1.

Strategy

(a) To find the equivalent resistance, first find the equivalent resistance of the parallel connection of R_2 and R_3. Then use this result to find the equivalent resistance of the series connection with R_1.

(b) The current through R_1 can be found using Ohm's law and the voltage applied. The current through R_1 is equal to the current from the battery. The potential drop V_1 across the resistor R_1 (which represents the resistance in the connecting wires) can be found using Ohm's law.

(c) The current through R_2 can be found using Ohm's law $I_2 = \frac{V_2}{R_2}$. The voltage across R_2 can be found using $V_2 = V - V_1$.

(d) Using Ohm's law $(V_2 = I_2R_2)$, the power dissipated by the resistor can also be found using $P_2 = I_2^2R_2 = \frac{V_2^2}{R_2}$.

Solution

1. To find the equivalent resistance of the circuit, notice that the parallel connection of R_2 and R_3 is in series with R_1, so the equivalent resistance is

$$R_{eq} = R_1 + \left(\frac{1}{R_2} + \frac{1}{R_3}\right)^{-1} = 1.00\,\Omega + \left(\frac{1}{6.00\,\Omega} + \frac{1}{13.00\,\Omega}\right)^{-1} = 5.10\,\Omega.$$

   The total resistance of this combination is intermediate between the pure series and pure parallel values ($20.0\,\Omega$ and $0.804\,\Omega$, respectively).

2. The current through R_1 is equal to the current supplied by the battery:

$$I_1 = I = \frac{V}{R_{eq}} = \frac{12.0\,V}{5.10\,\Omega} = 2.35\,A.$$

   The voltage across R_1 is

$$V_1 = I_1R_1 = (2.35\,A)(1\,\Omega) = 2.35\,V.$$

   The voltage applied to R_2 and R_3 is less than the voltage supplied by the battery by an amount V_1. When wire resistance is large, it can significantly affect the operation of the devices represented by R_2 and R_3.

3. To find the current through R_2, we must first find the voltage applied to it. The voltage across the two resistors in parallel is the same:

$$V_2 = V_3 = V - V_1 = 12.0\,V - 2.35\,V = 9.65\,V.$$

   Now we can find the current I_2 through resistance R_2 using Ohm's law:

$$I_2 = \frac{V_2}{R_2} = \frac{9.65\,V}{6.00\,\Omega} = 1.61\,A.$$

   The current is less than the 2.00 A that flowed through R_2 when it was connected in parallel to the battery in the previous parallel circuit example.

4. The power dissipated by R_2 is given by

$$P_2 = I_2^2R_2 = (1.61\,A)^2(6.00\,\Omega) = 15.5\,W.$$

Significance

The analysis of complex circuits can often be simplified by reducing the circuit to a voltage source and an equivalent resistance. Even if the entire circuit cannot be reduced to a single voltage source and a single equivalent resistance, portions of the circuit may be reduced, greatly simplifying the analysis.

Exercise 6.3.3

Consider the electrical circuits in your home. Give at least two examples of circuits that must use a combination of series and parallel circuits to operate efficiently.

Solution

All the overhead lighting circuits are in parallel and connected to the main supply line, so when one bulb burns out, all the overhead lighting does not go dark. Each overhead light will have at least one switch in series with the light, so you can turn it on and off.

A refrigerator has a compressor and a light that goes on when the door opens. There is usually only one cord for the refrigerator to plug into the wall. The circuit containing the compressor and the circuit containing the lighting circuit are in parallel, but there is a switch in series with the light. A thermostat controls a switch that is in series with the compressor to control the temperature of the refrigerator.

Practical Implications

One implication of this last example is that resistance in wires reduces the current and power delivered to a resistor. If wire resistance is relatively large, as in a worn (or a very long) extension cord, then this loss can be significant. If a large current is drawn, the **IR** drop in the wires can also be significant and may become apparent from the heat generated in the cord.

For example, when you are rummaging in the **refrigerator** and the motor comes on, the refrigerator light dims momentarily. Similarly, you can see the passenger compartment light dim when you start the engine of your car (although this may be due to resistance inside the battery itself).

What is happening in these high-current situations is illustrated in Figure $6.3.7$. The device represented by R_3 has a very low resistance, so when it is switched on, a large current flows. This increased current causes a larger **IR** drop in the wires represented by R_1, reducing the voltage across the light bulb (which is R_2), which then dims noticeably.

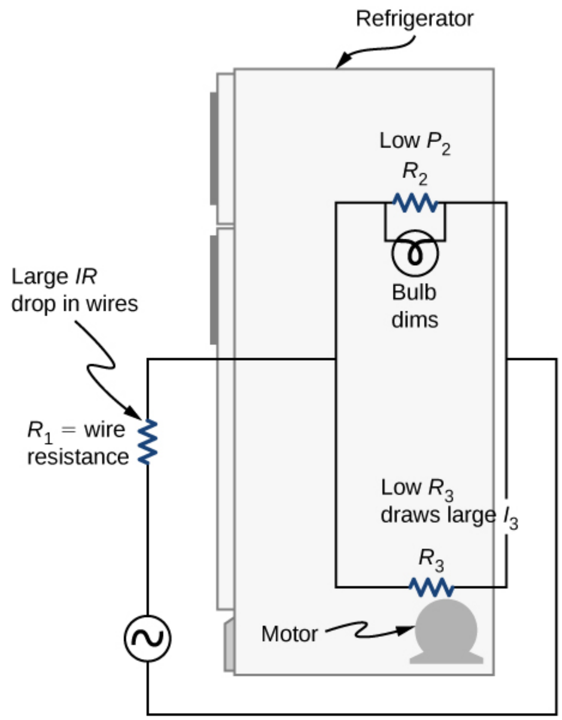


Figure $6.3.7$: Why do lights dim when a large appliance is switched on? The answer is that the large current the appliance motor draws causes a significant **IR** drop in the wires and reduces the voltage across the light.

✓ Problem-Solving Strategy: Series and Parallel Resistors

1. Draw a clear circuit diagram, labeling all resistors and voltage sources. This step includes a list of the known values for the problem, since they are labeled in your circuit diagram.
2. Identify exactly what needs to be determined in the problem (identify the unknowns). A written list is useful.
3. Determine whether resistors are in series, parallel, or a combination of both series and parallel. Examine the circuit diagram to make this assessment. Resistors are in series if the same current must pass sequentially through them.
4. Use the appropriate list of major features for series or parallel connections to solve for the unknowns. There is one list for series and another for parallel.

5. Check to see whether the answers are reasonable and consistent.

Example 6.3.4: Combining Series and Parallel circuits

Two resistors connected in series (R_1, R_2) are connected to two resistors that are connected in parallel (R_3, R_4). The series-parallel combination is connected to a battery. Each resistor has a resistance of 10.00 Ohms. The wires connecting the resistors and battery have negligible resistance. A current of 2.00 Amps runs through resistor R_1. What is the voltage supplied by the voltage source?

Strategy

Use the steps in the preceding problem-solving strategy to find the solution for this example.

Solution

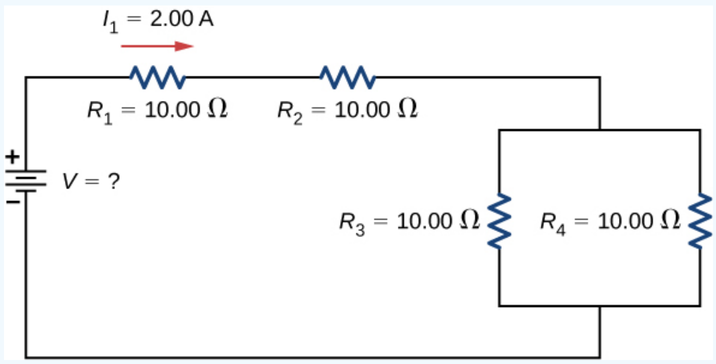


Figure 6.3.8: To find the unknown voltage, we must first find the equivalent resistance of the circuit.

1. Draw a clear circuit diagram (Figure 6.3.8).
2. The unknown is the voltage of the battery. In order to find the voltage supplied by the battery, the equivalent resistance must be found.
3. In this circuit, we already know that the resistors R_1 and R_2 are in series and the resistors R_3 and R_4 are in parallel. The equivalent resistance of the parallel configuration of the resistors R_3 and R_4 is in series with the series configuration of resistors R_1 and R_2.
4. The voltage supplied by the battery can be found by multiplying the current from the battery and the equivalent resistance of the circuit. The current from the battery is equal to the current through R_1 and is equal to 2.00 A. We need to find the equivalent resistance by reducing the circuit. To reduce the circuit, first consider the two resistors in parallel. The equivalent resistance is

$$R_{34} = \left(\frac{1}{10.00\,\Omega} + \frac{1}{10.00\,\Omega}\right)^{-1} = 5.00\,\Omega.$$

   This parallel combination is in series with the other two resistors, so the equivalent resistance of the circuit is $R_{eq} = R_1 + R_2 + R_{34} = (25.00\,\Omega$. The voltage supplied by the battery is therefore $V = IR_{eq} = 2.00\,A(25.00\,\Omega) = 50.00\,V$.
5. One way to check the consistency of your results is to calculate the power supplied by the battery and the power dissipated by the resistors. The power supplied by the battery is $P_{batt} = IV = 100.00\,W$.

Since they are in series, the current through R_2 equals the current through R_1. Since $R_3 = R_4$, the current through each will be 1.00 Amps. The power dissipated by the resistors is equal to the sum of the power dissipated by each resistor:

$$\begin{aligned} P &= I_1^2 R_1 + I_2^2 R_2 + I_3^2 R_3 + I_4^2 R_4 \\ &= 40.00\,W + 40.00\,W + 10.00\,W + 10.00\,W = 100.\,W. \end{aligned}$$

Since the power dissipated by the resistors equals the power supplied by the battery, our solution seems consistent.

Significance

If a problem has a combination of series and parallel, as in this example, it can be reduced in steps by using the preceding problem-solving strategy and by considering individual groups of series or parallel connections. When finding R_{eq} for a parallel connection, the reciprocal must be taken with care. In addition, units and numerical results must be reasonable. Equivalent series resistance should be greater, whereas equivalent parallel resistance should be smaller, for example. Power should be greater for the same devices in parallel compared with series, and so on.

6.4: Kirchhoff's Rules and Resistor Circuits

Learning Objectives

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

- State Kirchhoff's junction rule
- State Kirchhoff's loop rule
- Analyze complex circuits using Kirchhoff's rules

We have just seen that some circuits may be analyzed by reducing a circuit to a single voltage source and an equivalent resistance. Many complex circuits cannot be analyzed with the series-parallel techniques developed in the preceding sections. In this section, we elaborate on the use of Kirchhoff's rules to analyze more complex circuits. For example, the circuit in Figure $6.4.1$ is known as a **multi-loop circuit**, which consists of junctions. A junction, also known as a node, is a connection of three or more wires. In this circuit, the previous methods cannot be used, because not all the resistors are in clear series or parallel configurations that can be reduced. Give it a try. The resistors R_1 and R_2 are in series and can be reduced to an equivalent resistance. The same is true of resistors R_4 and R_5. But what do you do then?

Even though this circuit cannot be analyzed using the methods already learned, two circuit analysis rules can be used to analyze any circuit, simple or complex. The rules are known as **Kirchhoff's rules**, after their inventor Gustav **Kirchhoff** (1824–1887).

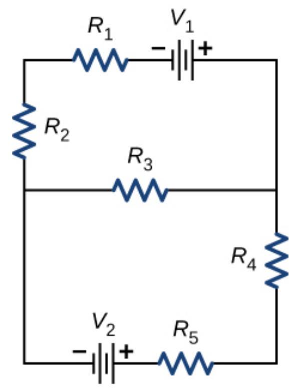


Figure $6.4.1$: This circuit cannot be reduced to a combination of series and parallel connections. However, we can use Kirchhoff's rules to analyze it.

Kirchhoff's Rules

- Kirchhoff's first rule—the junction rule. The sum of all currents entering a junction must equal the sum of all currents leaving the junction:

$$\sum I_{in} = \sum I_{out}. \tag{6.4.1}$$

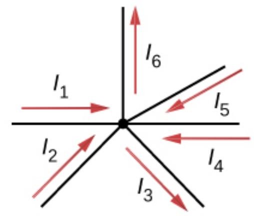
- Kirchhoff's second rule—the loop rule. The algebraic sum of changes in potential around any closed circuit path (loop) must be zero:

$$\sum V = 0. \tag{6.4.2}$$

We now provide explanations of these two rules, followed by problem-solving hints for applying them and a worked example that uses them.

Kirchhoff's Junction Rule

Kirchhoff's junction rule applies to the charge entering and leaving a junction (Figure $6.4.2$). As stated earlier, a junction, or node, is a connection of three or more wires. Current is the flow of charge, and charge is conserved; thus, whatever charge flows into the junction must flow out.



$$\sum I_{in} = \sum I_{out}$$
$$I_1 + I_2 + I_4 + I_5 = I_3 + I_6$$

Figure $6.4.2$: Charge must be conserved, so the sum of currents into a junction must be equal to the sum of currents out of the junction.

Although it is an over-simplification, an analogy can be made with water pipes connected in a plumbing junction. If the wires in Figure $6.4.2$ were replaced by water pipes, and the water was assumed to be incompressible, the volume of water flowing into the junction must equal the volume of water flowing out of the junction.

Kirchhoff's Loop Rule

Kirchhoff's loop rule applies to potential differences. The loop rule is stated in terms of potential V rather than potential energy, but the two are related since $U = qV$. In a closed loop, whatever energy is supplied by a voltage source, the energy must be transferred into other forms by the devices in the loop, since there are no other ways in which energy can be transferred into or out of the circuit. Kirchhoff's loop rule states that the algebraic sum of potential differences, including voltage supplied by the voltage sources and resistive elements, in any loop must be equal to zero. For example, consider a simple loop with no junctions, as in Figure $6.4.3$.

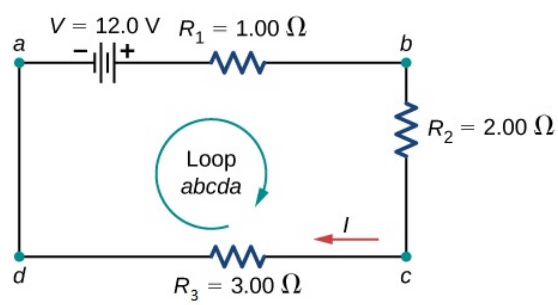


Figure $6.4.3$: A simple loop with no junctions. Kirchhoff's loop rule states that the algebraic sum of the voltage differences is equal to zero.

The circuit consists of a voltage source and three external load resistors. The labels **a**, **b**, **c**, and **d** serve as references, and have no other significance. The usefulness of these labels will become apparent soon. The loop is designated as Loop **abcda**, and the labels help keep track of the voltage differences as we travel around the circuit. Start at point **a** and travel to point **b**. The voltage of the voltage source is added to the equation and the potential drop of the resistor R_1 is subtracted. From point **b** to **c**, the potential drop across R_2 is subtracted. From **c** to **d**, the potential drop across R_3 is subtracted. From points **d** to **a**, nothing is done because there are no components.

Figure $6.4.4$ shows a graph of the voltage as we travel around the loop. Voltage increases as we cross the battery, whereas voltage decreases as we travel across a resistor. The **potential drop**, or change in the electric potential, is equal to the current through the resistor times the resistance of the resistor. Since the wires have negligible resistance, the voltage remains constant as we cross the wires connecting the components.

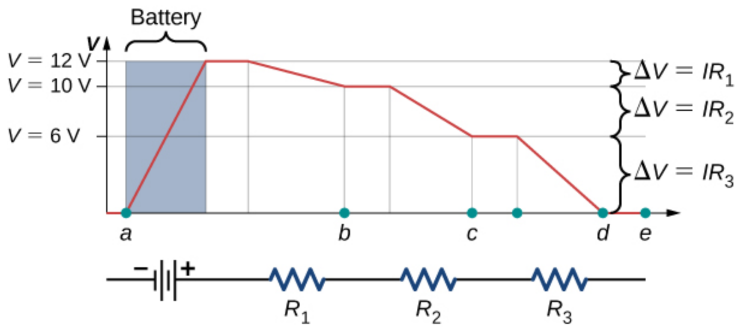


Figure 6.4.4: A voltage graph as we travel around the circuit. The voltage increases as we cross the battery and decreases as we cross each resistor. Since the resistance of the wire is quite small, we assume that the voltage remains constant as we cross the wires connecting the components.

Then Kirchhoff's loop rule states

$$V - IR_1 - IR_2 - IR_3 = 0. \tag{6.4.3}$$

The loop equation can be used to find the current through the loop:

$$I = \frac{V}{R_1 + R_2 + R_3} = \frac{12.00\,V}{1.00\,\Omega + 2.00\,\Omega + 3.00\,\Omega} = 2.00\,A. \tag{6.4.4}$$

This loop could have been analyzed using the previous methods, but we will demonstrate the power of Kirchhoff's method in the next section.

Applying Kirchhoff's Rules

By applying Kirchhoff's rules, we generate a set of linear equations that allow us to find the unknown values in circuits. These may be currents, voltages, or resistances. Each time a rule is applied, it produces an equation. If there are as many independent equations as unknowns, then the problem can be solved.

Using Kirchhoff's method of analysis requires several steps, as listed in the following procedure.

✔ Problem-Solving Strategy: Kirchhoff's Rules

1. Label points in the circuit diagram using lowercase letters **a**, **b**, **c**, …. These labels simply help with orientation.
2. Locate the junctions in the circuit. The junctions are points where three or more wires connect. Label each junction with the currents and directions into and out of it. Make sure at least one current points into the junction and at least one current points out of the junction.
3. Choose the loops in the circuit. Every component must be contained in at least one loop, but a component may be contained in more than one loop.
4. Apply the junction rule. Again, some junctions should not be included in the analysis. You need only use enough nodes to include every current.
5. Apply the loop rule. Use the map in Figure 6.4.5.

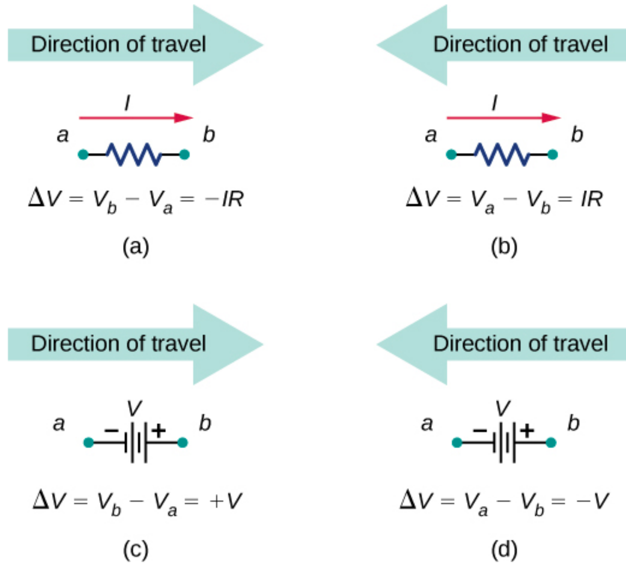


Figure $6.4.5$: Each of these resistors and voltage sources is traversed from **a** to **b**. (a) When moving across a resistor in the same direction as the current flow, subtract the potential drop. (b) When moving across a resistor in the opposite direction as the current flow, add the potential drop. (c) When moving across a voltage source from the negative terminal to the positive terminal, add the potential drop. (d) When moving across a voltage source from the positive terminal to the negative terminal, subtract the potential drop.

Let's examine some steps in this procedure more closely. When locating the junctions in the circuit, do not be concerned about the direction of the currents. If the direction of current flow is not obvious, choosing any direction is sufficient as long as at least one current points into the junction and at least one current points out of the junction. If the arrow is in the opposite direction of the conventional current flow, the result for the current in question will be negative but the answer will still be correct.

The number of nodes depends on the circuit. Each current should be included in a node and thus included in at least one junction equation. Do not include nodes that are not linearly independent, meaning nodes that contain the same information.

Consider Figure $6.4.6$. There are two junctions in this circuit: Junction **b** and Junction **e**. Points **a**, **c**, **d**, and **f** are not junctions, because a junction must have three or more connections. The equation for Junction **b** is $I_1 = I_2 + I_3$, and the equation for Junction **e** is $I_2 + I_3 = I_1$. These are equivalent equations, so it is necessary to keep only one of them.

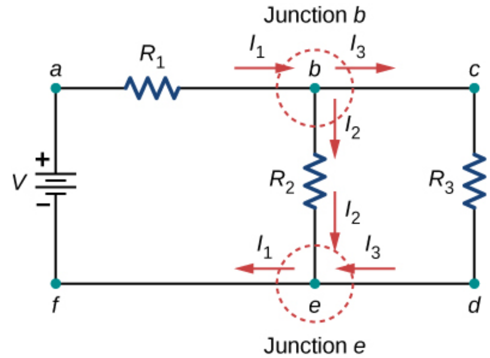


Figure $6.4.6$: At first glance, this circuit contains two junctions, Junction **b** and Junction **e**, but only one should be considered because their junction equations are equivalent.

When choosing the loops in the circuit, you need enough loops so that each component is covered once, without repeating loops. Figure $6.4.7$ shows four choices for loops to solve a sample circuit; choices (a), (b), and (c) have a sufficient amount of loops to solve the circuit completely. Option (d) reflects more loops than necessary to solve the circuit.

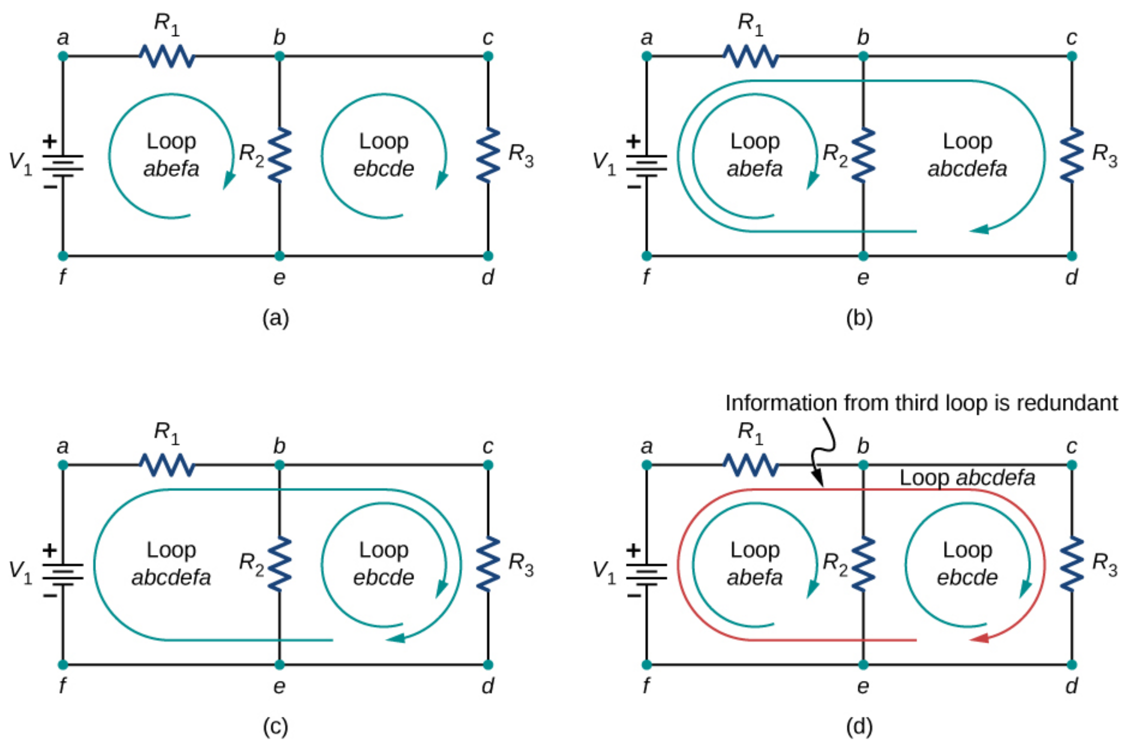


Figure 6.4.7: Panels (a)–(c) are sufficient for the analysis of the circuit. In each case, the two loops shown contain all the circuit elements necessary to solve the circuit completely. Panel (d) shows three loops used, which is more than necessary. Any two loops in the system will contain all information needed to solve the circuit. Adding the third loop provides redundant information.

Consider the circuit in Figure 6.4.8a. Let us analyze this circuit to find the current through each resistor. First, label the circuit as shown in part (b).

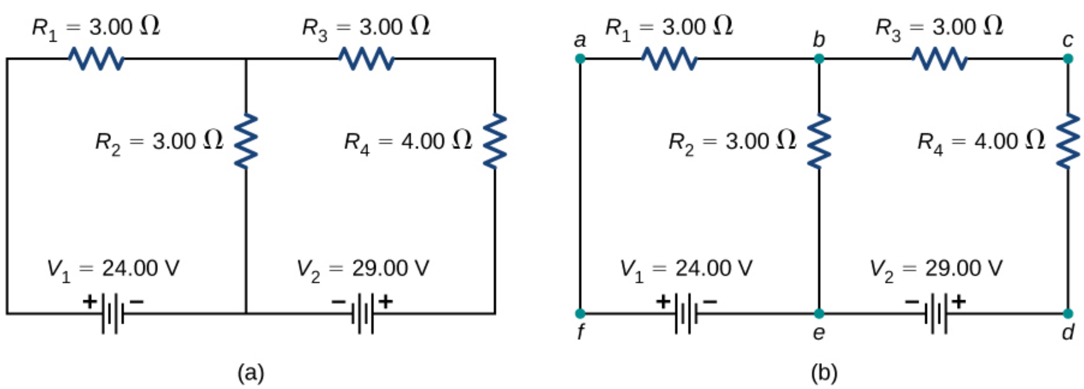


Figure 6.4.8: (a) A multi-loop circuit. (b) Label the circuit to help with orientation.

Next, determine the junctions. In this circuit, points **b** and **e** each have three wires connected, making them junctions. Start to apply Kirchhoff's junction rule $\left(\sum I_{in} = \sum I_{out}\right)$ by drawing arrows representing the currents and labeling each arrow, as shown in Figure 6.4.9. Junction **b** shows that $I_1 = I_2 + I_3$ and Junction **e** shows that $I_2 + I_3 = I_1$. Since Junction **e** gives the same information of Junction **b**, it can be disregarded. This circuit has three unknowns, so we need three linearly independent equations to analyze it.

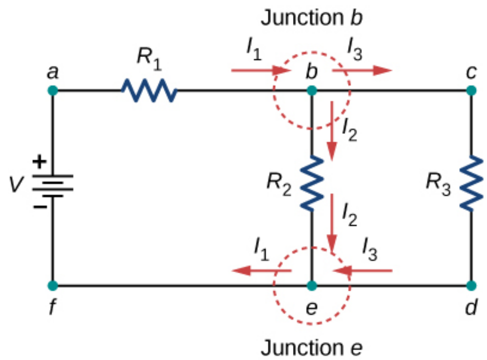


Figure 6.4.9: (a) This circuit has two junctions, labeled b and e, but only node b is used in the analysis. (b) Labeled arrows represent the currents into and out of the junctions.

Next we need to choose the loops. In Figure 6.4.10, Loop **abefa** includes the voltage source V_1 and resistors R_1 and R_2. The loop starts at point **a**, then travels through points **b**, **e**, and **f**, and then back to point **a**. The second loop, Loop **ebcde**, starts at point **e** and includes resistors R_2 and R_3, and the voltage source V_2.

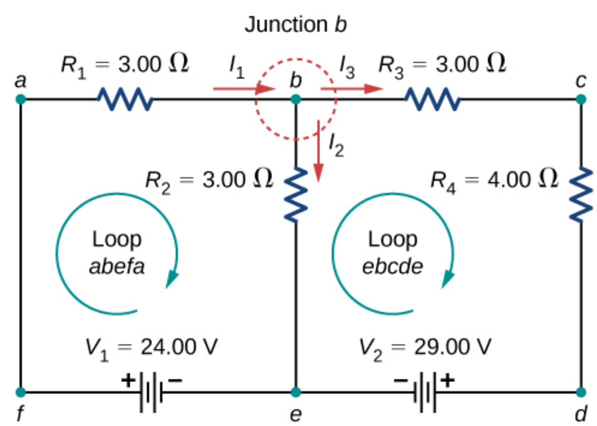


Figure 6.4.10: Choose the loops in the circuit.

Now we can apply Kirchhoff's loop rule, using the map in Figure 6.4.5. Starting at point **a** and moving to point **b**, the resistor R_1 is crossed in the same direction as the current flow I_1, so the potential drop I_1R_1 is subtracted. Moving from point **b** to point **e**, the resistor R_2 is crossed in the same direction as the current flow I_2 so the potential drop I_2R_2 is subtracted. Moving from point **e** to point **f**, the voltage source V_1 is crossed from the negative terminal to the positive terminal, so V_1 is added. There are no components between points **f** and **a.** The sum of the voltage differences must equal zero:

$$Loop\ abefa:\ -I_1R_1 - I_2R_2 + V_1 = 0\ or\ V_1 = I_1R_1 + I_2R_2. \tag{6.4.5}$$

Finally, we check loop **ebcde**. We start at point **e** and move to point **b**, crossing R_2 in the opposite direction as the current flow I_2. The potential drop I_2R_2 is added. Next, we cross R_3 and R_4 in the same direction as the current flow I_3 and subtract the potential drops I_3R_3 and I_3R_4. Note that the current is the same through resistors R_3 and R_4,because they are connected in series. Finally, the voltage source is crossed from the positive terminal to the negative terminal, and the voltage source V_2 is subtracted. The sum of these voltage differences equals zero and yields the loop equation

$$Loop\ ebcde:\ I_2R_2 - I_3(R_3 + R_4) - V_2 = 0. \tag{6.4.6}$$

We now have three equations, which we can solve for the three unknowns.

$$\text{Junction b: } I_1 - I_2 - I_3 = 0. \tag{6.4.7}$$

$$\text{Loop abefa: } I_1R_1 + I_2R_2 = V_1. \tag{6.4.8}$$

$$\text{Loop ebcde: } I_2R_2 - I_3(R_3 + R_4) = V_2. \tag{6.4.9}$$

To solve the three equations for the three unknown currents, start by eliminating current I_2. First add Equation 6.4.7 times R_2 to Equation 6.4.8. The result is Equation 6.4.11:

$$(R_1 + R_2)I_1 - R_2I_3 = V_1. \tag{6.4.10}$$

$$6\,\Omega I_1 - 3\Omega I_3 = 24\,V. \tag{6.4.11}$$

Next, subtract Equation 6.4.9 from Equation 6.4.8. The result is Equation 6.4.13:

$$I_1 R_1 + I_3(R_3 + R_4) = V_1 - V_2. \tag{6.4.12}$$

$$3\Omega I_1 + 7\Omega I_3 = -5\,V. \tag{6.4.13}$$

We can solve Equations 6.4.11 and 6.4.13 for current I_1. Adding seven times Equation 6.4.11 and three times Equation 6.4.13 results in $51\,\Omega I_1 = 153\,V$, or $I_1 = 3.00\,A$. Using Equation 6.4.11 results in $I_3 = -2.00\,A$. Finally, Equation 6.4.7 yields $I_2 = I_1 - I_3 = 5.00\,A$. One way to check that the solutions are consistent is to check the power supplied by the voltage sources and the power dissipated by the resistors:

$$P_{in} = I_1 V_1 + I_3 V_2 = 130\,W,$$

$$P_{out} = I_1^2 R_1 + I_2^2 R_2 + I_3^2 R_3 + I_3^2 R_4 = 130\,W.$$

Note that the solution for the current I_3 is negative. This is the correct answer, but suggests that the arrow originally drawn in the junction analysis is the direction opposite of conventional current flow. The power supplied by the second voltage source is 58 W and not −58 W.

Example 6.4.1: Calculating Current by Using Kirchhoff's Rules

Find the currents flowing in the circuit in Figure 6.4.11.

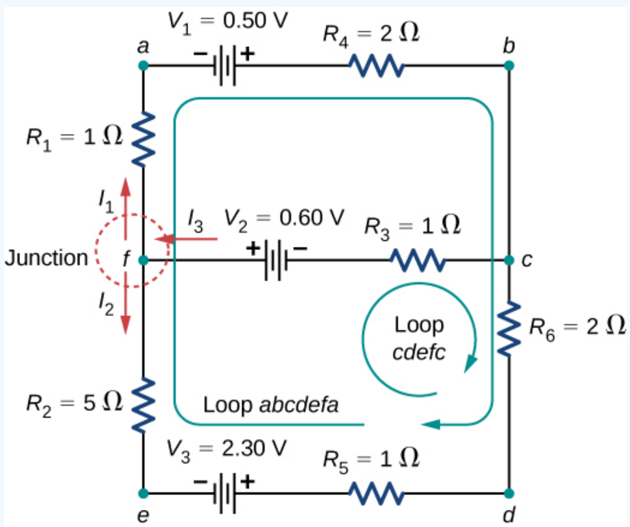


Figure 6.4.11: This circuit is a combination of series and parallel configurations of resistors and voltage sources. This circuit cannot be analyzed using the techniques discussed in the section on Source Voltage but can be analyzed using Kirchhoff's rules.

Strategy

This circuit is sufficiently complex that the currents cannot be found using Ohm's law and the series-parallel techniques—it is necessary to use Kirchhoff's rules. Currents have been labeled I_1, I_2, and I_3 in the figure, and assumptions have been made about their directions. Locations on the diagram have been labeled with letters **a** through **h**. In the solution, we apply the junction and loop rules, seeking three independent equations to allow us to solve for the three unknown currents.

Solution

Applying the junction and loop rules yields the following three equations. We have three unknowns, so three equations are required.

$$Junction\, c:\; I_1 + I_2 = I_3. \tag{6.4.14}$$

$$Loop\, abcdefa:\; I_1(R_1 + R_4) - I_2(R_2 + R_5 + R_6) = V_1 - V_3. \tag{6.4.15}$$

$$Loop\, cdefc:\; I_2(R_2 + R_5 + R_6) + I_3 R_3 = V_2 + V_3. \tag{6.4.16}$$

Simplify the equations by placing the unknowns on one side of the equations.

$$Junction\, c:\; I_1 + I_2 - I_3 = 0. \tag{6.4.17}$$

$$Loop\, abcdefa:\; I_1(3\Omega) - I_2(8\Omega) = 0.5\,V - 2.30\,V. \tag{6.4.18}$$

$$Loop\, cdefc:\; I_2(8\Omega) + I_3(1\Omega) = 0.6\,V + 2.30\,V. \tag{6.4.19}$$

Simplify the equations. The first loop equation can be simplified by dividing both sides by 3.00. The second loop equation can be simplified by dividing both sides by 6.00.

$$Junction\, c:\; I_1 + I_2 - I_3 = 0. \tag{6.4.20}$$

$$Loop\, abcdefa:\; I_1(3\Omega) - I_2(8\Omega) = -1.8\,V. \tag{6.4.21}$$

$$Loop\, cdefc:\; I_2(8\Omega) + I_3(1\Omega) = 2.90\,V. \tag{6.4.22}$$

The results are

$$I_1 = 0.20\,A,\; I_2 = 0.30\,A,\; I_3 = 0.50\,A. \tag{6.4.23}$$

Significance

A method to check the calculations is to compute the power dissipated by the resistors and the power supplied by the voltage sources:

$$P_{R_1} = I_1^2 R_1 = 0.04\,W. \tag{6.4.24}$$

$$P_{R_2} = I_2^2 R_2 = 0.45\,W. \tag{6.4.25}$$

$$P_{R_3} = I_3^2 R_3 = 0.25\,W. \tag{6.4.26}$$

$$P_{R_4} = I_1^2 R_4 = 0.08\,W. \tag{6.4.27}$$

$$P_{R_5} = I_2^2 R_5 = 0.09\,W. \tag{6.4.28}$$

$$P_{R_6} = I_2^2 R_1 = 0.18\,W. \tag{6.4.29}$$

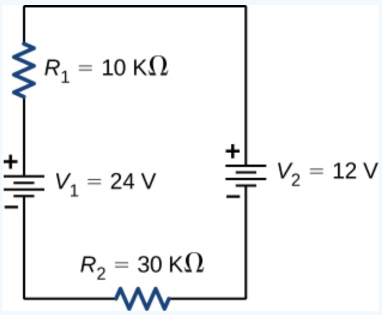
$$P_{dissipated} = 1.09\,W. \tag{6.4.30}$$

$$P_{source} = I_1 V_1 + I_2 V_3 + I_3 V_2 = 0.10\; + 0.69\,W + 0.30\,W = 1.09\,W. \tag{6.4.31}$$

The power supplied equals the power dissipated by the resistors.

? Exercise 6.4.1

In considering the following schematic and the power supplied and consumed by a circuit, will a voltage source always provide power to the circuit, or can a voltage source consume power?



Answer

The circuit can be analyzed using Kirchhoff's loop rule. The first voltage source supplies power: $P_{in} = IV_1 = 7.20\,mW$. The second voltage source consumes power: $P_{out} = IV_2 + I^2R_1 + I^2R_2 = 7.2\,mW$.

✓ Example 6.4.2: Calculating Current by Using Kirchhoff's Rules

Find the current flowing in the circuit in Figure 6.4.12.

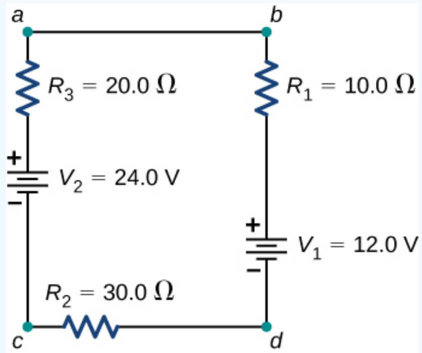


Figure 6.4.12: This circuit consists of three resistors and two batteries connected in series. Note that the batteries are connected with opposite polarities.

Strategy

This circuit can be analyzed using Kirchhoff's rules. There is only one loop and no nodes. Choose the direction of current flow. For this example, we will use the clockwise direction from point **a** to point **b**. Consider Loop **abcda** and use Figure 6.4.5 to write the loop equation. Note that according to Figure 6.4.5, battery V_1 will be added and battery V_2 will be subtracted.

Solution

Applying the junction rule yields the following three equations. We have one unknown, so one equation is required:

$$Loop\ abcda: \ -IR_1 - V_1 - IR_2 + V_2 - IR_3 = 0. \tag{6.4.32}$$

Simplify the equations by placing the unknowns on one side of the equations. Use the values given in the figure.

$$I(R_1 + R_2 + R_3) = V_2 - V_1. \tag{6.4.33}$$

$$I = \frac{V_2 - V_1}{R_1 + R_2 + R_3} = \frac{24\ V - 12\ V}{10.0\ \Omega + 30.0\ \Omega + 10.0\ \Omega} = 0.20\ A. \tag{6.4.34}$$

Significance

The power dissipated or consumed by the circuit equals the power supplied to the circuit, but notice that the current in the battery V_1 is flowing through the battery from the positive terminal to the negative terminal and consumes power.

$$P_{R_1} = I^2 R_1 = 0.40\ W \tag{6.4.35}$$

$$P_{R_2} = I^2 R_2 = 1.20\ W \tag{6.4.36}$$

$$P_{R_3} = I^2 R_3 = 0.80\ W \tag{6.4.37}$$

$$P_{V_1} = IV_1 = 2.40\ W \tag{6.4.38}$$

$$P_{dissipated} = 4.80\ W \tag{6.4.39}$$

$$P_{source} = IV_2 = 4.80\ W \tag{6.4.40}$$

The power supplied equals the power dissipated by the resistors and consumed by the battery V_1.

? Exercise 6.4.2

When using Kirchhoff's laws, you need to decide which loops to use and the direction of current flow through each loop. In analyzing the circuit in Example 6.4.2, the direction of current flow was chosen to be clockwise, from point **a** to point **b**. How would the results change if the direction of the current was chosen to be counterclockwise, from point **b** to point **a**?

Answer

The current calculated would be equal to $I = -0.20\,A$ instead of $I = 0.20\,A$. The sum of the power dissipated and the power consumed would still equal the power supplied.

Multiple Voltage Sources

Many devices require more than one battery. Multiple voltage sources, such as batteries, can be connected in series configurations, parallel configurations, or a combination of the two.

In series, the positive terminal of one battery is connected to the negative terminal of another battery. Any number of voltage sources, including batteries, can be connected in series. Two batteries connected in series are shown in Figure 6.4.13. Using Kirchhoff's loop rule for the circuit in part (b) gives the result

$$\varepsilon_1 - Ir_1 + \varepsilon_2 - Ir_2 - IR = 0, \tag{6.4.41}$$

$$[(\varepsilon_1 + \varepsilon_2) - I(r_1 + r_2)] - IR = 0. \tag{6.4.42}$$

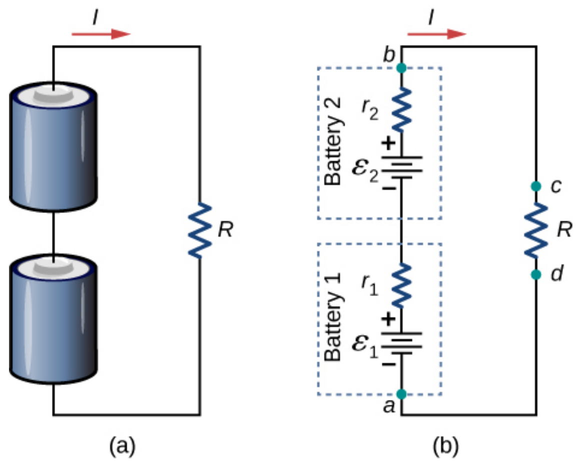


Figure 6.4.13: (a) Two batteries connected in series with a load resistor. (b) The circuit diagram of the two batteries and the load resistor, with each battery modeled as an idealized source voltage and an internal resistance.

When voltage sources are in series, their internal resistances can be added together and their source voltages can be added together to get the total values. Series connections of voltage sources are common—for example, in flashlights, toys, and other appliances. Usually, the cells are in series in order to produce a larger total source voltage. In Figure 6.4.13, the terminal voltage is

$$V_{terminal} = (\varepsilon_1 - Ir_1) + (\varepsilon_2 - Ir_2) = [(\varepsilon_1 + \varepsilon_2) - I(r_1 + r_2) - I(r_1 + r_2)] = (\varepsilon_1 + \varepsilon_2) + Ir_{eq}. \tag{6.4.43}$$

Note that the same current **I** is found in each battery because they are connected in series. The disadvantage of series connections of cells is that their internal resistances are additive.

Batteries are connected in series to increase the voltage supplied to the circuit. For instance, an LED flashlight may have two AAA cell batteries, each with a terminal voltage of 1.5 V, to provide 3.0 V to the flashlight.

Any number of batteries can be connected in series. For **N** batteries in series, the terminal voltage is equal to

Note

$$V_{terminal} = (\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{N-1} + \varepsilon_N) - I(r_1 + r_2 + \ldots + r_{N-1} + r_N) = \sum_{i=1}^{N} \varepsilon_i - Ir_{eq} \tag{6.4.44}$$

where the equivalent resistance is

$$r_{eq} = \sum_{i=1}^{N} r_i \tag{6.4.45}$$

When a load is placed across voltage sources in series, as in Figure 6.4.14, we can find the current:

$$(\varepsilon_1 - Ir_1) + (\varepsilon_2 - Ir_2) = IR, \tag{6.4.46}$$

$$Ir_1 + Ir_2 + IR = \varepsilon_1 + \varepsilon_2, \tag{6.4.47}$$

$$I = \frac{\varepsilon_1 + \varepsilon_2}{r_1 + r_2 + R}. \tag{6.4.48}$$

As expected, the internal resistances increase the equivalent resistance.

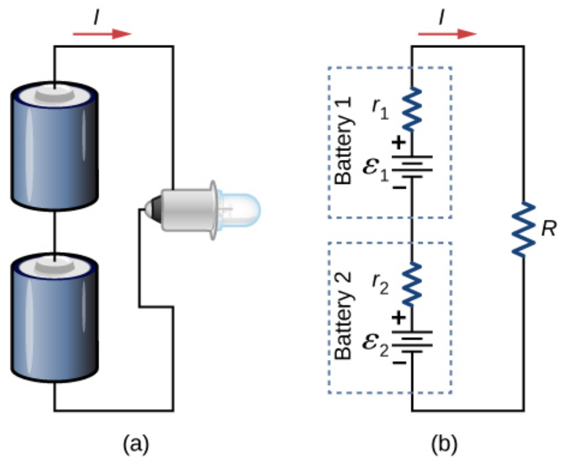


Figure 6.4.14: Two batteries connect in series to an LED bulb, as found in a flashlight.

Voltage sources, such as batteries, can also be connected in parallel. Figure 6.4.15 shows two batteries with identical source voltages in parallel and connected to a load resistance. When the batteries are connected in parallel, the positive terminals are connected together and the negative terminals are connected together, and the load resistance is connected to the positive and negative terminals. Normally, voltage sources in parallel have identical source voltages. In this simple case, since the voltage sources are in parallel, the total source voltage is the same as the individual source voltages of each battery.

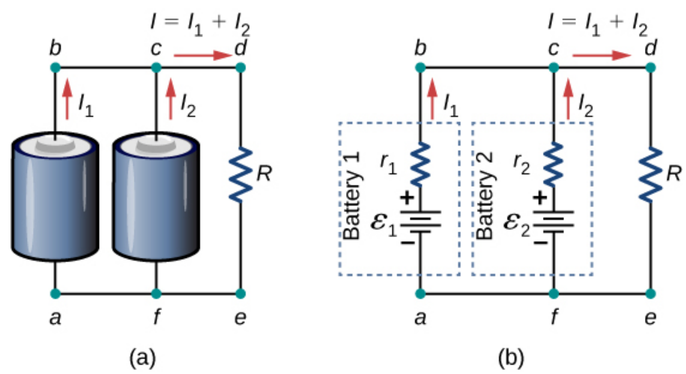


Figure 6.4.15: (a) Two batteries connect in parallel to a load resistor. (b) The circuit diagram shows the shows battery modeled as a source voltage and an internal resistor. The two batteries have identical source voltages (each labeled by ε) connected in parallel that produce the same source voltage.

Consider the Kirchhoff analysis of the circuit in Figure $6.4.15b$ There are two loops and a node at point **b** and $\varepsilon = \varepsilon_1 = \varepsilon_2$.

Node b: $I_1 + I_2 - I = 0$.

Loop abcfa: $\varepsilon_2 - I_1 r_1 + I_2 r_2 - \varepsilon = 0,\ I_1 r_1 = I_2 r_2$.

Loop fcdef: $\varepsilon_2 - I_2 r_2 - IR = 0,\ \ \varepsilon - I_2 r_2 - IR = 0.$

Solving for the current through the load resistor results in $I = \frac{\varepsilon}{r_{eq} + R}$, where $r_{eq} = \left(\frac{1}{r_1} + \frac{1}{r_2}\right)^{-1}$. The terminal voltage is equal to the potential drop across the load resistor $IR = \left(\frac{\varepsilon}{r_{eq} + R}\right)$.

The parallel connection reduces the internal resistance and thus can produce a larger current.

Any number of batteries can be connected in parallel. For **N** batteries in parallel, the terminal voltage is equal to

> **Note**
>
> $$V_{terminal} = \varepsilon - I\left(\frac{1}{r_1} + \frac{1}{r_2} + \ldots + \frac{1}{r_{N-1}} + \frac{1}{r_N}\right)^{-1} = \varepsilon - Ir_{eq} \tag{6.4.49}$$

where the equivalent resistance is

$$r_{eq} = \left(\sum_{i=1}^{N} \frac{1}{r_i}\right)^{-1} \tag{6.4.50}$$

As an example, some diesel trucks use two 12-V batteries in parallel; they produce a total source voltage of 12 V but can deliver the larger current needed to start a diesel engine.

In summary, the terminal voltage of batteries in series is equal to the sum of the individual source voltages minus the sum of the internal resistances times the current. When batteries are connected in parallel, they usually have equal source voltages and the terminal voltage is equal to the source voltage minus the equivalent internal resistance times the current, where the equivalent internal resistance is smaller than the individual internal resistances. Batteries are connected in series to increase the terminal voltage to the load. Batteries are connected in parallel to increase the current to the load.

Solar Cell Arrays

Another example dealing with multiple voltage sources is that of combinations of **solar cells** - wired in both series and parallel combinations to yield a desired voltage and current. Photovoltaic generation, which is the conversion of sunlight directly into electricity, is based upon the photoelectric effect. The photoelectric effect is beyond the scope of this chapter and is covered in Photons and Matter Waves, but in general, photons hitting the surface of a solar cell create an electric current in the cell.

Most solar cells are made from pure silicon. Most single cells have a voltage output of about 0.5 V, while the current output is a function of the amount of sunlight falling on the cell (the incident solar radiation known as the insolation). Under bright noon sunlight, a current per unit area of about $100\,mA/cm^2$ of cell surface area is produced by typical single-crystal cells.

Individual solar cells are connected electrically in modules to meet electrical energy needs. They can be wired together in series or in parallel - connected like the batteries discussed earlier. A solar-cell array or module usually consists of between 36 and 72 cells, with a power output of 50 W to 140 W.

Solar cells, like batteries, provide a direct current (dc) voltage. Current from a dc voltage source is unidirectional. Most household appliances need an alternating current (ac) voltage.

6.5: Application - Electrical Meters

Learning Objectives

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

- Describe how to connect a voltmeter in a circuit to measure voltage
- Describe how to connect an ammeter in a circuit to measure current
- Describe the use of an ohmmeter

Ohm’s law and Kirchhoff’s method are useful to analyze and design electrical circuits, providing you with the voltages across, the current through, and the resistance of the components that compose the circuit. To measure these parameters require instruments, and these instruments are described in this section.

DC Voltmeters and Ammeters

Whereas voltmeter**s** measure voltage, ammeter**s** measure current. Some of the meters in automobile dashboards, digital cameras, cell phones, and tuner-amplifiers are actually voltmeters or ammeters (Figure $6.5.1$). The internal construction of the simplest of these meters and how they are connected to the system they monitor give further insight into applications of series and parallel connections.

Figure $6.5.1$: The fuel and temperature gauges (far right and far left, respectively) in this 1996 Volkswagen are voltmeters that register the voltage output of “sender” units. These units are proportional to the amount of gasoline in the tank and to the engine temperature. (credit: Christian Giersing)

Measuring Current with an Ammeter

To measure the current through a device or component, the ammeter is placed in series with the device or component. A series connection is used because objects in series have the same current passing through them. (See Figure $6.5.2$, where the ammeter is represented by the symbol A.)

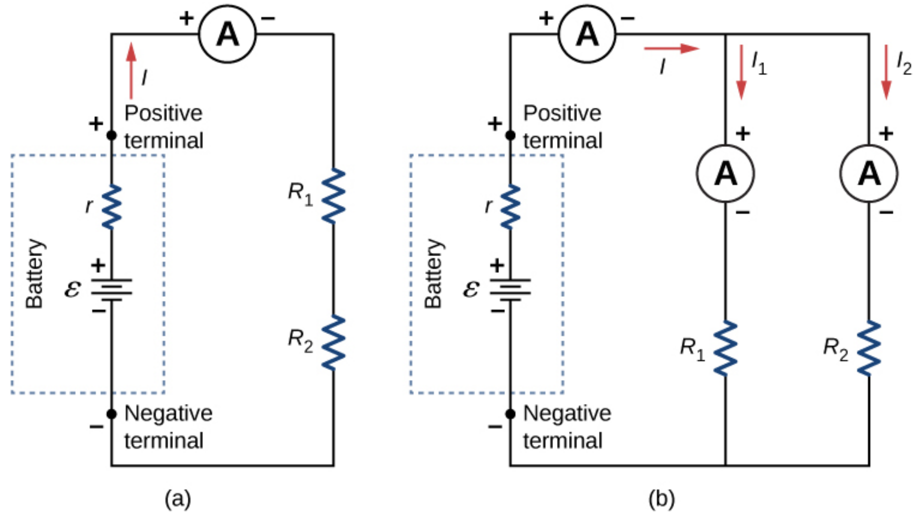


Figure 6.5.2: (a) When an ammeter is used to measure the current through two resistors connected in series to a battery, a single ammeter is placed in series with the two resistors because the current is the same through the two resistors in series. (b) When two resistors are connected in parallel with a battery, three meters, or three separate ammeter readings, are necessary to measure the current from the battery and through each resistor. The ammeter is connected in series with the component in question.

Ammeters need to have a very low resistance, a fraction of a milliohm. If the resistance is not negligible, placing the ammeter in the circuit would change the equivalent resistance of the circuit and modify the current that is being measured. Since the current in the circuit travels through the meter, ammeters normally contain a fuse to protect the meter from damage from currents which are too high.

Measuring Voltage with a Voltmeter

A voltmeter is connected in parallel with whatever device it is measuring. A parallel connection is used because objects in parallel experience the same potential difference. (See Figure 6.5.3, where the voltmeter is represented by the symbol V.)

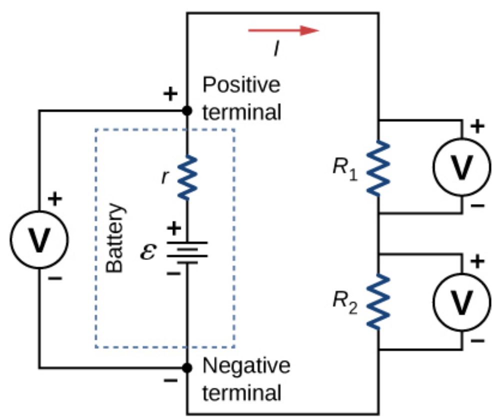


Figure 6.5.3: To measure potential differences in this series circuit, the voltmeter (V) is placed in parallel with the voltage source or either of the resistors. Note that terminal voltage is measured between the positive terminal and the negative terminal of the battery or voltage source. It is not possible to connect a voltmeter directly across the emf without including the internal resistance **r** of the battery.

Since voltmeters are connected in parallel, the voltmeter must have a very large resistance. Digital voltmeters convert the analog voltage into a digital value to display on a digital readout (Figure 6.5.4). Inexpensive voltmeters have resistances on the order of $R_M = 10\,M\Omega$, whereas high-precision voltmeters have resistances on the order of $R_M = 10\,G\Omega$. The value of the resistance may vary, depending on which scale is used on the meter.

(a) (b)

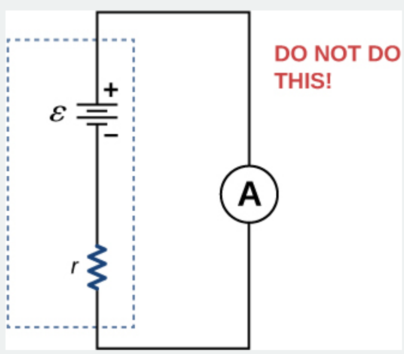
Figure 6.5.4: (a) An analog voltmeter uses a galvanometer to measure the voltage. (b) Digital meters use an analog-to-digital converter to measure the voltage. (credit a and credit b: Joseph J. Trout)

Analog and Digital Meters

You may encounter two types of meters in the physics lab: analog and digital. The term 'analog' refers to signals or information represented by a continuously variable physical quantity, such as voltage or current. An analog meter uses a galvanometer, which is essentially a coil of wire with a small resistance, in a magnetic field, with a pointer attached that points to a scale. Current flows through the coil, causing the coil to rotate. To use the galvanometer as an ammeter, a small resistance is placed in parallel with the coil. For a voltmeter, a large resistance is placed in series with the coil. A digital meter uses a component called an analog-to-digital (A to D) converter and expresses the current or voltage as a series of the digits 0 and 1, which are used to run a digital display. Most analog meters have been replaced by digital meters.

Check Your Understanding

Digital meters are able to detect smaller currents than analog meters employing galvanometers. How does this explain their ability to measure voltage and current more accurately than analog meters?



[Hide Solution]

Since digital meters require less current than analog meters, they alter the circuit less than analog meters. Their resistance as a voltmeter can be far greater than an analog meter, and their resistance as an ammeter can be far less than an analog meter. Consult Figure 6.5.3 and Figure 6.5.2 and their discussion in the text

Note

In this virtual lab simulation, you may construct circuits with resistors, voltage sources, ammeters and voltmeters to test your knowledge of circuit design.

Ohmmeters

An ohmmeter is an instrument used to measure the resistance of a component or device. The operation of the ohmmeter is based on Ohm's law. Traditional ohmmeters contained an internal voltage source (such as a battery) that would be connected across the component to be tested, producing a current through the component. A galvanometer was then used to measure the current and the resistance was deduced using Ohm's law. Modern digital meters use a constant current source to pass current through the component, and the voltage difference across the component is measured. In either case, the resistance is measured using Ohm's law $(R = V/I)$ where the voltage is known and the current is measured, or the current is known and the voltage is measured.

The component of interest should be isolated from the circuit; otherwise, you will be measuring the equivalent resistance of the circuit. An ohmmeter should never be connected to a "live" circuit, one with a voltage source connected to it and current running through it. Doing so can damage the meter.

6.6: Application - Grounding and Electrical Safety

Learning Objectives

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

- Define thermal hazard, shock hazard, and short circuit.
- Explain what effects various levels of current have on the human body.

There are two known hazards of electricity—thermal and shock. A **thermal hazard** is one where excessive electric power causes undesired thermal effects, such as starting a fire in the wall of a house. A **shock hazard** occurs when electric current passes through a person. Shocks range in severity from painful, but otherwise harmless, to heart-stopping lethality. This section considers these hazards and the various factors affecting them in a quantitative manner. Electrical Safety: Systems and Devices will consider systems and devices for preventing electrical hazards.

Thermal Hazards

Electric power causes undesired heating effects whenever electric energy is converted to thermal energy at a rate faster than it can be safely dissipated. A classic example of this is the **short circuit**, a low-resistance path between terminals of a voltage source. An example of a short circuit is shown in Figure 1. Insulation on wires leading to an appliance has worn through, allowing the two wires to come into contact. Such an undesired contact with a high voltage is called a *short*. Since the resistance of the short, r, is very small, the power dissipated in the short, $P=V^2/r$, is very large. for example, if V is 120 V and r is 0.100Ω, then the power is 144kW, *much* greater than that used by a typical household appliance. Thermal energy delivered at this rate will very quickly raise the temperature of surrounding materials, melting or perhaps igniting them.

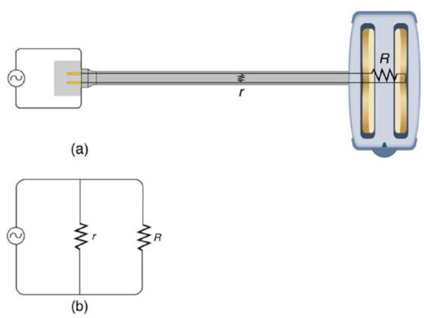


Figure 6.6.1: A short circuit is an undesired low-resistance path across a voltage source. (a) Worn insulation on the wires of a toaster allow them to come into contact with a low resistance r. Since $P=V^2/r$, thermal power is created so rapidly that the cord melts or burns. (b) A schematic of the short circuit.

One particularly insidious aspect of a short circuit is that its resistance may actually be decreased due to the increase in temperature. This can happen if the short creates ionization. These charged atoms and molecules are free to move and, thus, lower the resistance r. Since $P=V^2/r$, the power dissipated in the short rises, possibly causing more ionization, more power, and so on. High voltages, such as the 480-V AC used in some industrial applications, lend themselves to this hazard, because higher voltages create higher initial power production in a short.

Another serious, but less dramatic, thermal hazard occurs when wires supplying power to a user are overloaded with too great a current. As discussed in the previous section, the power dissipated in the supply wires is $P=I^2R_W$, where R_W is the resistance of the wires and I the current flowing through them. If either I or R_W is too large, the wires overheat. For example, a worn appliance cord (with some of its braided wires broken) may have $R_W=2.00\Omega$ rather than the 0.100Ω it should be. If 10.0 A of current passes through the cord, then $P=I^2R_W=200W$ is dissipated in the cord—much more than is safe. Similarly, if a wire with a $0.100-\Omega$ resistance is meant to carry a few amps, but is instead carrying 100 A, it will severely overheat. The power dissipated in the wire will in that case be $P=1000W$. Fuses and circuit breakers are used to limit excessive currents. (See Figure 2 and Figure 3.) Each device opens the circuit automatically when a sustained current exceeds safe limits.

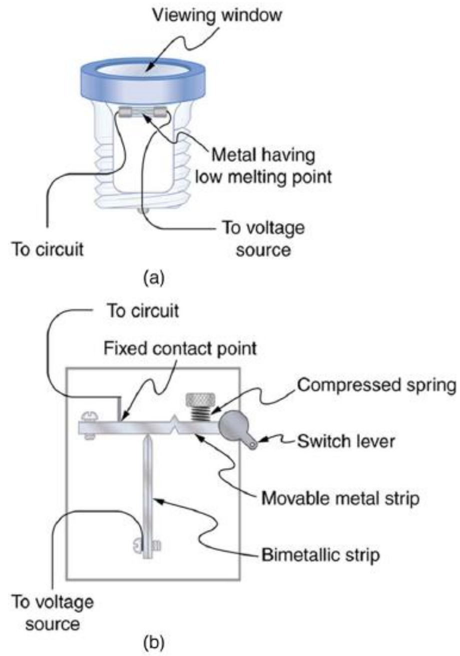


Figure 6.6.2: (a) A fuse has a metal strip with a low melting point that, when overheated by an excessive current, permanently breaks the connection of a circuit to a voltage source. (b) A circuit breaker is an automatic but restorable electric switch. The one shown here has a bimetallic strip that bends to the right and into the notch if overheated. The spring then forces the metal strip downward, breaking the electrical connection at the points.

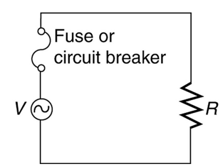


Figure 6.6.3: Schematic of a circuit with a fuse or circuit breaker in it. Fuses and circuit breakers act like automatic switches that open when sustained current exceeds desired limits.

Fuses and circuit breakers for typical household voltages and currents are relatively simple to produce, but those for large voltages and currents experience special problems. For example, when a circuit breaker tries to interrupt the flow of high-voltage electricity, a spark can jump across its points that ionizes the air in the gap and allows the current to continue flowing. Large circuit breakers found in power-distribution systems employ insulating gas and even use jets of gas to blow out such sparks. Here AC is safer than DC, since AC current goes through zero 120 times per second, giving a quick opportunity to extinguish these arcs.

Shock Hazards

Electrical currents through people produce tremendously varied effects. An electrical current can be used to block back pain. The possibility of using electrical current to stimulate muscle action in paralyzed limbs, perhaps allowing paraplegics to walk, is under study. TV dramatizations in which electrical shocks are used to bring a heart attack victim out of ventricular fibrillation (a massively irregular, often fatal, beating of the heart) are more than common. Yet most electrical shock fatalities occur because a current put the heart into fibrillation. A pacemaker uses electrical shocks to stimulate the heart to beat properly. Some fatal shocks do not produce burns, but warts can be safely burned off with electric current (though freezing using liquid nitrogen is now more common). Of course, there are consistent explanations for these disparate effects. The major factors upon which the effects of electrical shock depend are

1. The amount current I
2. The path taken by the current
3. The duration of the shock
4. The frequency f of the current ($f=0$ for DC)

The table below gives the effects of electrical shocks as a function of current for a typical accidental shock. The effects are for a shock that passes through the trunk of the body, has a duration of 1 s, and is caused by 60-Hz power.

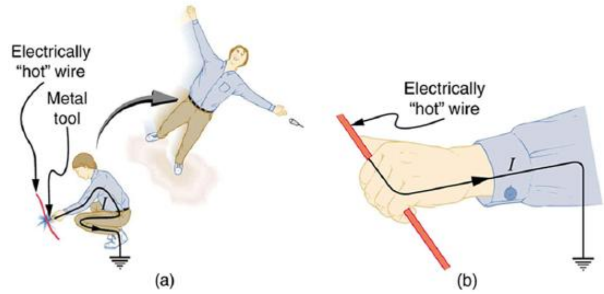


Figure 6.6.4: An electric current can cause muscular contractions with varying effects. (a) The victim is "thrown" backward by involuntary muscle contractions that extend the legs and torso. (b) The victim can't let go of the wire that is stimulating all the muscles in the hand. Those that close the fingers are stronger than those that open them.

Current (mA)	**Effect**
1	Threshold of sensation
5	Maximum harmless current
10-20	Onset of sustained muscular contraction; cannot let go for duration of shock; contraction of chest muscles may stop breathing during shock
50	Onset of pain
100-300+	Ventricular fibrillation possible; often fatal
300	Onset of burns depending on concentration of current
6000 (6A)	Onset of sustained ventricular contraction and respiratory paralysis; both cease when shock ends; heartbeat may return to normal; used to defibrillate the heart

Effects of Electrical Shock as a Function of Current

Our bodies are relatively good conductors due to the water in our bodies. Given that larger currents will flow through sections with lower resistance (to be further discussed in the next chapter), electric currents preferentially flow through paths in the human body that have a minimum resistance in a direct path to earth. The earth is a natural electron sink. Wearing insulating shoes, a requirement in many professions, prohibits a pathway for electrons by providing a large resistance in that path. Whenever working with high-power tools (drills), or in risky situations, ensure that you do not provide a pathway for current flow (especially through the heart).

Very small currents pass harmlessly and unfelt through the body. This happens to you regularly without your knowledge. The threshold of sensation is only 1 mA and, although unpleasant, shocks are apparently harmless for currents less than 5 mA. A great number of safety rules take the 5-mA value for the maximum allowed shock. At 10 to 20 mA and above, the current can stimulate sustained muscular contractions much as regular nerve impulses do. People sometimes say they were knocked across the room by a shock, but what really happened was that certain muscles contracted, propelling them in a manner not of their own choosing. (See Figure 4a.) More frightening, and potentially more dangerous, is the "can't let go" effect illustrated in Figure 4b.

The muscles that close the fingers are stronger than those that open them, so the hand closes involuntarily on the wire shocking it. This can prolong the shock indefinitely. It can also be a danger to a person trying to rescue the victim, because the rescuer's hand may close about the victim's wrist. Usually the best way to help the victim is to give the fist a hard knock/blow/jar with an insulator or to throw an insulator at the fist. Modern electric fences, used in animal enclosures, are now pulsed on and off to allow people who touch them to get free, rendering them less lethal than in the past.

Greater currents may affect the heart. Its electrical patterns can be disrupted, so that it beats irregularly and ineffectively in a condition called "ventricular fibrillation." This condition often lingers after the shock and is fatal due to a lack of blood circulation. The threshold for ventricular fibrillation is between 100 and 300 mA. At about 300 mA and above, the shock can cause burns, depending on the concentration of current—the more concentrated, the greater the likelihood of burns.

Very large currents cause the heart and diaphragm to contract for the duration of the shock. Both the heart and breathing stop. Interestingly, both often return to normal following the shock. The electrical patterns on the heart are completely erased in a manner that the heart can start afresh with normal beating, as opposed to the permanent disruption caused by smaller currents that can put the heart into ventricular fibrillation. The latter is something like scribbling on a blackboard, whereas the former completely erases it. TV dramatizations of electric shock used to bring a heart attack victim out of ventricular fibrillation also show large paddles. These are used to spread out current passed through the victim to reduce the likelihood of burns.

Current is the major factor determining shock severity (given that other conditions such as path, duration, and frequency are fixed, such as in the table and preceding discussion). A larger voltage is more hazardous, but since $I = V/R$, the severity of the shock depends on the combination of voltage and resistance. For example, a person with dry skin has a resistance of about $200 k\Omega$. If he comes into contact with 120-V AC, a current $I = (120 V)/(200 k\Omega) = .6 mA$ passes harmlessly through him. The same person soaking wet may have a resistance of $10.0 k\Omega$ and the same 120 V will produce a current of 12 mA—above the "can't let go" threshold and potentially dangerous.

Most of the body's resistance is in its dry skin. When wet, salts go into ion form, lowering the resistance significantly. The interior of the body has a much lower resistance than dry skin because of all the ionic solutions and fluids it contains. If skin resistance is bypassed, such as by an intravenous infusion, a catheter, or exposed pacemaker leads, a person is rendered **microshock sensitive**. In this condition, currents about 1/1000 those listed in the table above produce similar effects. During open-heart surgery, currents as small as $20 \mu A$ can be used to still the heart. Stringent electrical safety requirements in hospitals, particularly in surgery and intensive care, are related to the doubly disadvantaged microshock-sensitive patient. The break in the skin has reduced his resistance, and so the same voltage causes a greater current, and a much smaller current has a greater effect.

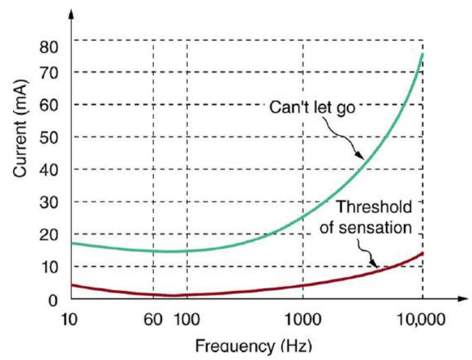


Figure 6.6.5: Graph of average values for the threshold of sensation and the "can't let go" current as a function of frequency. The lower the value, the more sensitive the body is at that frequency.

Factors other than current that affect the severity of a shock are its path, duration, and AC frequency. Path has obvious consequences. For example, the heart is unaffected by an electric shock through the brain, such as may be used to treat manic depression. And it is a general truth that the longer the duration of a shock, the greater its effects. Figure 5 presents a graph that illustrates the effects of frequency on a shock. The curves show the minimum current for two different effects, as a function of frequency. The lower the current needed, the more sensitive the body is at that frequency. Ironically, the body is most sensitive to frequencies near the 50- or 60-Hz frequencies in common use. The body is slightly less sensitive for DC ($f = 0$), mildly confirming Edison's claims that AC presents a greater hazard. At higher and higher frequencies, the body becomes progressively less sensitive to any effects that involve nerves. This is related to the maximum rates at which nerves can fire or be stimulated. At very high frequencies, electrical current travels only on the surface of a person. Thus a wart can be burned off with very high frequency current without causing the heart to stop. (Do not try this at home with 60-Hz AC!) Some of the spectacular demonstrations of electricity, in which high-voltage arcs are passed through the air and over people's bodies, employ high frequencies and low currents. (See Figure 6.) Electrical safety safety devices and techniques are discussed in detail in Electrical Safety: Systems and Devices.

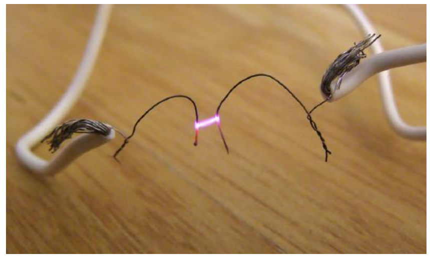


Figure 6.6.6: Is this electric arc dangerous? The answer depends on the AC frequency and the power involved. (credit: Khimich Alex, Wikimedia Commons)

Electrical Safety: Systems and Devices

Figure 6.6.7(a) shows the schematic for a simple ac circuit with no safety features. This is not how power is distributed in practice. Modern household and industrial wiring requires the **three-wire system**, shown schematically in part (b), which has several safety features, with live, neutral, and ground wires. First is the familiar circuit breaker (or fuse) to prevent thermal overload. Second is a protective case around the appliance, such as a toaster or refrigerator. The case's safety feature is that it prevents a person from touching exposed wires and coming into electrical contact with the circuit, helping prevent shocks.

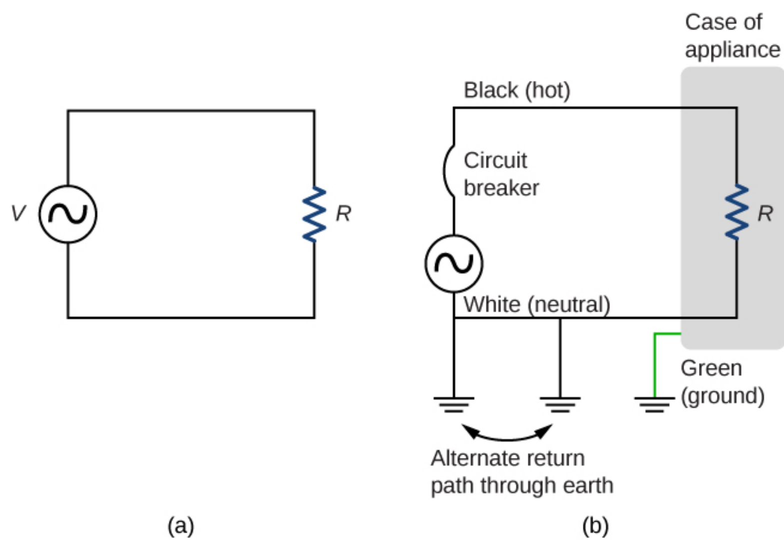


Figure 6.6.7: (a) Schematic of a simple ac circuit with a voltage source and a single appliance represented by the resistance **R**. There are no safety features in this circuit. (b) The three-wire system connects the neutral wire to ground at the voltage source and user location, forcing it to be at zero volts and supplying an alternative return path for the current through ground. Also grounded to zero volts is the case of the appliance. A circuit breaker or fuse protects against thermal overload and is in series on the active (live/hot) wire.

There are three connections to ground shown in 6.6.7(b). Recall that a ground connection is a low-resistance path directly to ground. The two ground connections on the neutral wire force it to be at zero volts relative to ground, giving the wire its name. This wire is therefore safe to touch even if its insulation, usually white, is missing. The neutral wire is the return path for the current to follow to complete the circuit. Furthermore, the two ground connections supply an alternative path through ground (a good conductor) to complete the circuit. The ground connection closest to the power source could be at the generating plant, whereas the other is at the user's location. The third ground is to the case of the appliance, through the green ground wire, forcing the case, too, to be at zero volts. The live or hot wire (hereafter referred to as "live/hot") supplies voltage and current to operate the appliance. Figure 6.6.8 shows a more pictorial version of how the three-wire system is connected through a three-prong plug to an appliance.

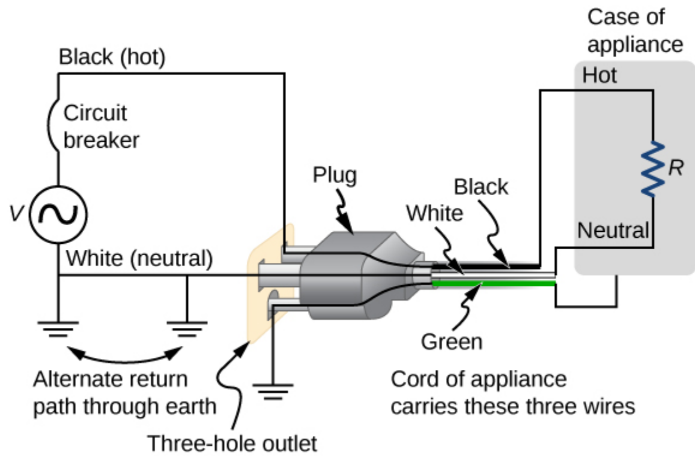


Figure 6.6.8: The standard three-prong plug can only be inserted in one way, to ensure proper function of the three-wire system.

Insulating plastic is color-coded to identify live/hot, neutral, and ground wires, but these codes vary around the world. It is essential to determine the color code in your region. Striped coatings are sometimes used for the benefit of those who are colorblind.

Grounding the case solves more than one problem. The simplest problem is worn insulation on the live/hot wire that allows it to contact the case, as shown in Figure (\PageIndex{9}\). Lacking a ground connection, a severe shock is possible. This is particularly dangerous in the kitchen, where a good connection to ground is available through water on the floor or a water faucet. With the ground connection intact, the circuit breaker will trip, forcing repair of the appliance.

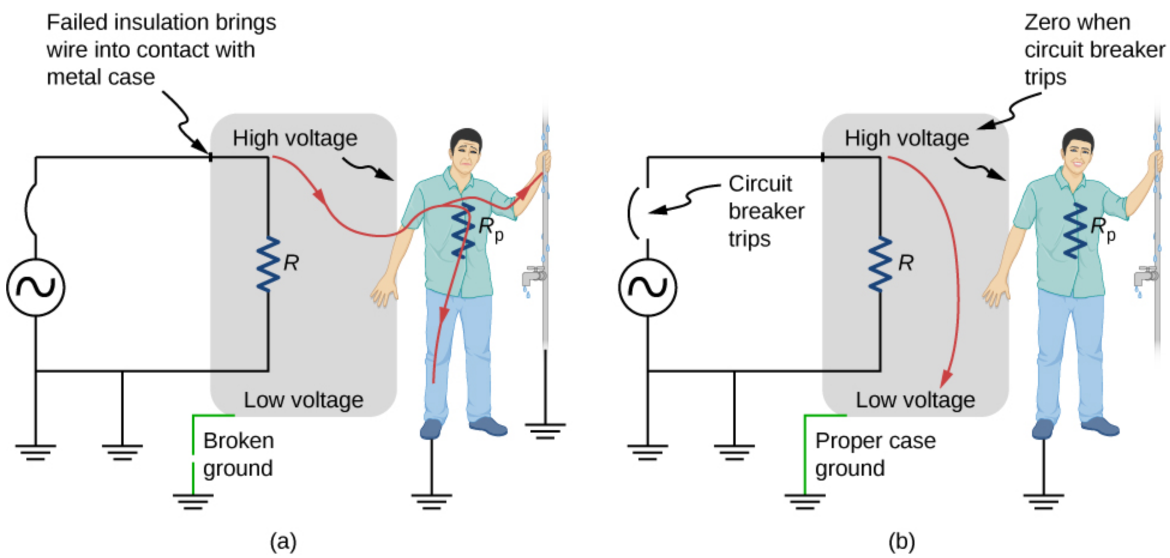


Figure 6.6.9: Worn insulation allows the live/hot wire to come into direct contact with the metal case of this appliance. (a) The ground connection being broken, the person is severely shocked. The appliance may operate normally in this situation. (b) With a proper ground, the circuit breaker trips, forcing repair of the appliance.

A **ground fault circuit interrupter** (GFCI) is a safety device found in updated kitchen and bathroom wiring that works based on electromagnetic induction. GFCIs compare the currents in the live/hot and neutral wires. When live/hot and neutral currents are not equal, it is almost always because current in the neutral is less than in the live/hot wire. Then some of the current, called a leakage current, is returning to the voltage source by a path other than through the neutral wire. It is assumed that this path presents a hazard. GFCIs are usually set to interrupt the circuit if the leakage current is greater than 5 mA, the accepted maximum harmless shock. Even if the leakage current goes safely to ground through an intact ground wire, the GFCI will trip, forcing repair of the leakage.

Footnotes

1 For an average male shocked through trunk of body for 1 s by 60-Hz AC. Values for females are 60–80% of those listed.

6.7: Direct-Current Circuits (Summary)

Key Terms

ammeter	instrument that measures current
equivalent resistance	resistance of a combination of resistors; it can be thought of as the resistance of a single resistor that can replace a combination of resistors in a series and/or parallel circuit
internal resistance	amount of resistance to the flow of current within the voltage source
junction rule	sum of all currents entering a junction must equal the sum of all currents leaving the junction
Kirchhoff's rules	set of two rules governing current and changes in potential in an electric circuit
loop rule	algebraic sum of changes in potential around any closed circuit path (loop) must be zero
potential difference	difference in electric potential between two points in an electric circuit, measured in volts
potential drop	loss of electric potential energy as a current travels across a resistor, wire, or other component
RC circuit	circuit that contains both a resistor and a capacitor
shock hazard	hazard in which an electric current passes through a person
source votage	energy produced per unit charge, drawn from a source that produces an electrical current
terminal voltage	potential difference measured across the terminals of a source when there is no load attached
thermal hazard	hazard in which an excessive electric current causes undesired thermal effects
three-wire system	wiring system used at present for safety reasons, with live, neutral, and ground wires
voltmeter	instrument that measures voltage

Key Equations

Terminal voltage of a single voltage source	$V_{terminal} = \varepsilon - Ir_{eq}$
Equivalent resistance of a series circuit	$R_{eq} = R_1 + R_2 + R_3 + \cdots + R_{N-1} + R_N = \sum_{i=1}^{N} R_i$
Equivalent resistance of a parallel circuit	$R_{eq} = (\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_N})^{-1} = (\sum_{i=1}^{N} \frac{1}{R_i})^{-1}$
Junction rule	$\sum I_{in} = \sum I_{out}$
Loop rule	$\sum V = 0$
Terminal voltage of N voltage sources in series	$V_{terminal} = \sum_{i=1}^{N} \varepsilon_i - I \sum_{i=1}^{N} r_i = \sum_{i=1}^{N} \varepsilon_i - Ir_{eq}$
Terminal voltage of N voltage sources in parallel	$V_{terminal} = \varepsilon - I \sum_{i=1}^{N} (\frac{1}{r_i})^{-1} = \varepsilon - Ir_{eq}$
Charge on a charging capacitor	$q(t) = C\varepsilon(1 - e^{-\frac{t}{RC}}) = Q(1 - e^{-\frac{t}{\tau}})$
Time constant	$\tau = RC$

Current during charging of a capacitor	$I=\frac{\varepsilon}{R}e^{-\frac{t}{RC}}=I_oe^{-\frac{t}{RC}}$
Charge on a discharging capacitor	$q(t)=Qe^{-\frac{t}{\tau}}$
Current during discharging of a capacitor	$I(t)=-\frac{Q}{RC}e^{-\frac{t}{\tau}}$

Summary

Source Voltage

- All voltage sources have two fundamental parts: a source of electrical energy that has a characteristic source voltage, and an internal resistance r. The source voltage is the work done per charge to keep the potential difference of a source constant. The source voltage is equal to the potential difference across the terminals when no current is flowing. The internal resistance **r** of a voltage source affects the output voltage when a current flows.
- The voltage output of a device is called its terminal voltage $V_{terminal}$ and is given by $V_{terminal}=\varepsilon-Ir$, where I is the electric current and is positive when flowing away from the positive terminal of the voltage source and r is the internal resistance.

Resistors in Series and Parallel

- The equivalent resistance of an electrical circuit with resistors wired in a series is the sum of the individual resistances:

$$R_s=R_1+R_2+R_3+\cdots=\sum_{i=1}^{N}R_i \quad .$$

- Each resistor in a series circuit has the same amount of current flowing through it.
- The potential drop, or power dissipation, across each individual resistor in a series is different, and their combined total is the power source input.
- The equivalent resistance of an electrical circuit with resistors wired in parallel is less than the lowest resistance of any of the components and can be determined using the formula

$$R_{eq}=\left(\frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3}+\cdots\right)^{-1}=\left(\sum_{i=1}^{N}\frac{1}{R_i}\right)^{-1} \quad .$$

- Each resistor in a parallel circuit has the same full voltage of the source applied to it.
- The current flowing through each resistor in a parallel circuit is different, depending on the resistance.
- If a more complex connection of resistors is a combination of series and parallel, it can be reduced to a single equivalent resistance by identifying its various parts as series or parallel, reducing each to its equivalent, and continuing until a single resistance is eventually reached.

Kirchhoff's Rules and Resistor Circuits

- Kirchhoff's rules can be used to analyze any circuit, simple or complex. The simpler series and parallel connection rules are special cases of Kirchhoff's rules.
- Kirchhoff's first rule, also known as the junction rule, applies to the charge to a junction. Current is the flow of charge; thus, whatever charge flows into the junction must flow out.
- Kirchhoff's second rule, also known as the loop rule, states that the voltage drop around a loop is zero.
- When calculating potential and current using Kirchhoff's rules, a set of conventions must be followed for determining the correct signs of various terms.
- When multiple voltage sources are in series, their internal resistances add together and their source voltages add together to get the total values.
- When multiple voltage sources are in parallel, their internal resistances combine to an equivalent resistance that is less than the individual resistance and provides a higher current than a single cell.
- Solar cells can be wired in series or parallel to provide increased voltage or current, respectively.

Application - Electrical Meters

- Voltmeters measure voltage, and ammeters measure current. Analog meters are based on the combination of a resistor and a galvanometer, a device that gives an analog reading of current or voltage. Digital meters are based on analog-to-digital converters and provide a discrete or digital measurement of the current or voltage.
- A voltmeter is placed in parallel with the voltage source to receive full voltage and must have a large resistance to limit its effect on the circuit.

- An ammeter is placed in series to get the full current flowing through a branch and must have a small resistance to limit its effect on the circuit.
- Standard voltmeters and ammeters alter the circuit they are connected to and are thus limited in accuracy.
- Ohmmeters are used to measure resistance. The component in which the resistance is to be measured should be isolated (removed) from the circuit.

Application - Grounding and Electrical Safety

- The two types of electric hazards are thermal (excessive power) and shock (current through a person). Electrical safety systems and devices are employed to prevent thermal and shock hazards.
- Shock severity is determined by current, path, duration, and ac frequency.
- Circuit breakers and fuses interrupt excessive currents to prevent thermal hazards.
- The three-wire system guards against thermal and shock hazards, utilizing live/hot, neutral, and ground wires, and grounding the neutral wire and case of the appliance.
- A ground fault circuit interrupter (GFCI) prevents shock by detecting the loss of current to unintentional paths.

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).

6.8: Direct-Current Circuits (Exercise)

Conceptual Questions

Source Voltage

1. What effect will the internal resistance of a rechargeable battery have on the energy being used to recharge the battery?

2. A battery with an internal resistance of r and an source voltage of 10.00 V is connected to a load resistor **R=r**. As the battery ages, the internal resistance triples. How much is the current through the load resistor reduced?

3. Show that the power dissipated by the load resistor is maximum when the resistance of the load resistor is equal to the internal resistance of the battery.

Resistors in Series and Parallel

4. A voltage occurs across an open switch. What is the power dissipated by the open switch?

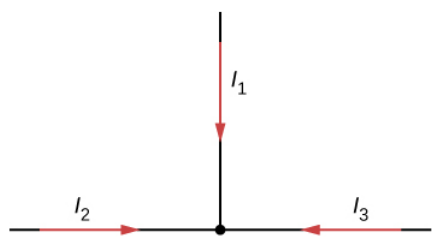
5. The severity of a shock depends on the magnitude of the current through your body. Would you prefer to be in series or in parallel with a resistance, such as the heating element of a toaster, if you were shocked by it? Explain.

6. Suppose you are doing a physics lab that asks you to put a resistor into a circuit, but all the resistors supplied have a larger resistance than the requested value. How would you connect the available resistances to attempt to get the smaller value asked for?

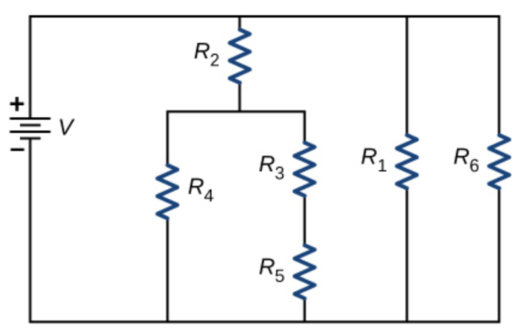
7. Some light bulbs have three power settings (not including zero), obtained from multiple filaments that are individually switched and wired in parallel. What is the minimum number of filaments needed for three power settings?

Kirchhoff's Rules and Resistor Circuits

8. Can all of the currents going into the junction shown below be positive? Explain.



9. Consider the circuit shown below. Does the analysis of the circuit require Kirchhoff's method, or can it be redrawn to simplify the circuit? If it is a circuit of series and parallel connections, what is the equivalent resistance?



10. Do batteries in a circuit always supply power to a circuit, or can they absorb power in a circuit? Give an example.

11. What are the advantages and disadvantages of connecting batteries in series? In parallel?

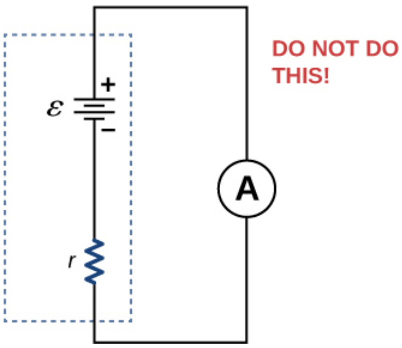
12. Semi-tractor trucks use four large 12-V batteries. The starter system requires 24 V, while normal operation of the truck's other electrical components utilizes 12 V. How could the four batteries be connected to produce 24 V? To produce 12 V? Why is 24 V better than 12 V for starting the truck's engine (a very heavy load)?

Application - Electrical Meters

13. What would happen if you placed a voltmeter in series with a component to be tested?

14. What is the basic operation of an ohmmeter as it measures a resistor?

15. Why should you not connect an ammeter directly across a voltage source as shown below?



Application - Grounding and Electrical Safety

18. Why isn't a short circuit necessarily a shock hazard?

19. We are often advised to not flick electric switches with wet hands, dry your hand first. We are also advised to never throw water on an electric fire. Why?

Problems

Source Voltage

20. A car battery with a 12-V source voltage and an internal resistance of **0.050Ω** is being charged with a current of 60 A. Note that in this process, the battery is being charged.

(a) What is the potential difference across its terminals?

(b) At what rate is thermal energy being dissipated in the battery?

(c) At what rate is electric energy being converted into chemical energy?

21. The label on a battery-powered radio recommends the use of rechargeable nickel-cadmium cells (nicads), although they have a 1.25-V source voltage, whereas alkaline cells have a 1.58-V source voltage. The radio has a **3.20Ω** resistance.

(a) Draw a circuit diagram of the radio and its batteries. Now, calculate the power delivered to the radio

(b) when using nicad cells, each having an internal resistance of **0.0400Ω**, and

(c) when using alkaline cells, each having an internal resistance of **0.200Ω**.

(d) Does this difference seem significant, considering that the radio's effective resistance is lowered when its volume is turned up?

22. An automobile starter motor has an equivalent resistance of **0.0500Ω** and is supplied by a 12.0-V battery with a **0.0100-Ω** internal resistance.

(a) What is the current to the motor?

(b) What voltage is applied to it?

(c) What power is supplied to the motor?

(d) Repeat these calculations for when the battery connections are corroded and add **0.0900Ω** to the circuit. (Significant problems are caused by even small amounts of unwanted resistance in low-voltage, high-current applications.)

23. (a) What is the internal resistance of a voltage source if its terminal potential drops by 2.00 V when the current supplied increases by 5.00 A?

(b) Can the source voltage of the voltage source be found with the information supplied?

24. A person with body resistance between his hands of **10.0kΩ** accidentally grasps the terminals of a 20.0-kV power supply. (Do NOT do this!)

(a) Draw a circuit diagram to represent the situation.

(b) If the internal resistance of the power supply is **2000Ω**, what is the current through his body?

(c) What is the power dissipated in his body?

(d) If the power supply is to be made safe by increasing its internal resistance, what should the internal resistance be for the maximum current in this situation to be 1.00 mA or less?

(e) Will this modification compromise the effectiveness of the power supply for driving low-resistance devices? Explain your reasoning.

25. A 12.0-V source voltageautomobile battery has a terminal voltage of 16.0 V when being charged by a current of 10.0 A.

(a) What is the battery's internal resistance?

(b) What power is dissipated inside the battery?

(c) At what rate (in **°C/min**) will its temperature increase if its mass is 20.0 kg and it has a specific heat of **0.300kcal/kg⋅°C**, assuming no heat escapes?

Resistors in Series and Parallel

26. (a) What is the resistance of a $1.00 \times 10^2 - \Omega$, a $2.50 - k\Omega$, and a $4.00 - k\Omega$ resistor connected in series?

(b) In parallel?

27. What are the largest and smallest resistances you can obtain by connecting a **36.0-Ω**, a **50.0-Ω**, and a **700-Ω** resistor together?

28. An 1800-W toaster, a 1400-W speaker, and a 75-W lamp are plugged into the same outlet in a 15-A fuse and 120-V circuit. (The three devices are in parallel when plugged into the same socket.)

(a) What current is drawn by each device?

(b) Will this combination blow the 15-A fuse?

29. Your car's 30.0-W headlight and 2.40-kW starter are ordinarily connected in parallel in a 12.0-V system. What power would one headlight and the starter consume if connected in series to a 12.0-V battery? (Neglect any other resistance in the circuit and any change in resistance in the two devices.)

30. (a) Given a 48.0-V battery and **24.0-Ω** and **96.0-Ω**9 resistors, find the current and power for each when connected in series.

(b) Repeat when the resistances are in parallel.

31. Referring to the example combining series and parallel circuits and Figure 10.16, calculate I_3 in the following two different ways:

(a) from the known values of I and I_2;

(b) using Ohm's law for R_3. In both parts, explicitly show how you follow the steps in the Figure 10.17.

32. Referring to Figure 10.16,

(a) Calculate P_3 and note how it compares with P_3 found in the first two example problems in this module.

(b) Find the total power supplied by the source and compare it with the sum of the powers dissipated by the resistors.

33. Refer to Figure 10.17 and the discussion of lights dimming when a heavy appliance comes on.

(a) Given the voltage source is 120 V, the wire resistance is **0.800Ω**, and the bulb is nominally 75.0 W, what power will the bulb dissipate if a total of 15.0 A passes through the wires when the motor comes on? Assume negligible change in bulb resistance.

(b) What power is consumed by the motor?

34. Show that if two resistors R_1 and R_2 are combined and one is much greater than the other ($R_1 \gg R_2$),

(a) their series resistance is very nearly equal to the greater resistance R_1 and

(b) their parallel resistance is very nearly equal to the smaller resistance R_2.

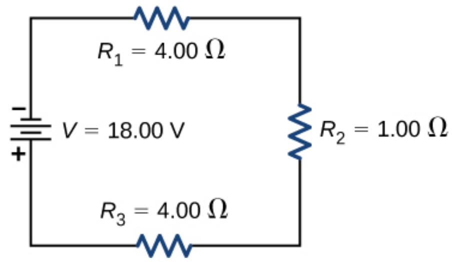
35. Consider the circuit shown below. The terminal voltage of the battery is **V=18.00V**.

(a) Find the equivalent resistance of the circuit.

(b) Find the current through each resistor.

(c) Find the potential drop across each resistor.

(d) Find the power dissipated by each resistor. (e) Find the power supplied by the battery.

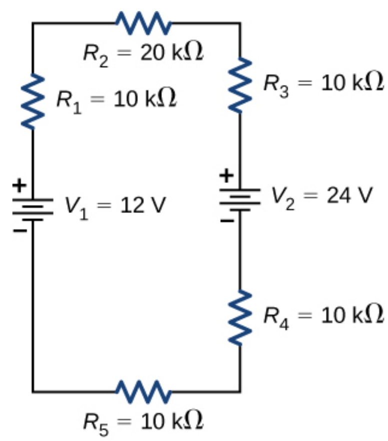


Kirchhoff's Rules and Resistor Circuits

36. Consider the circuit shown below.

(a) Find the voltage across each resistor.

(b)What is the power supplied to the circuit and the power dissipated or consumed by the circuit?



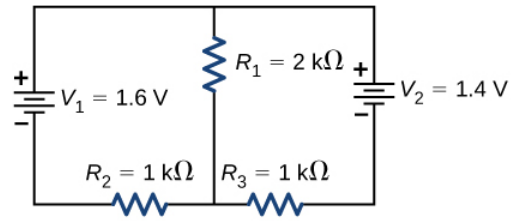
37. Consider the circuits shown below.

(a) What is the current through each resistor in part (a)?

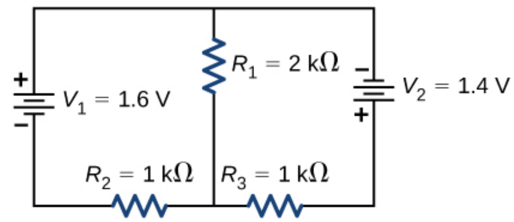
(b) What is the current through each resistor in part (b)?

(c) What is the power dissipated or consumed by each circuit?

(d) What is the power supplied to each circuit?

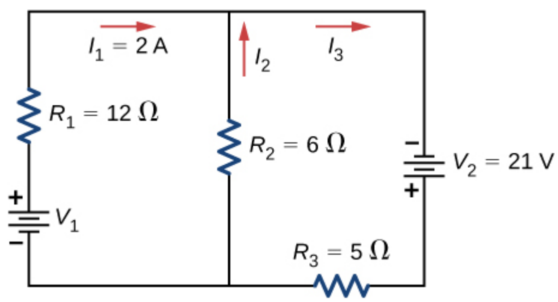


(a)

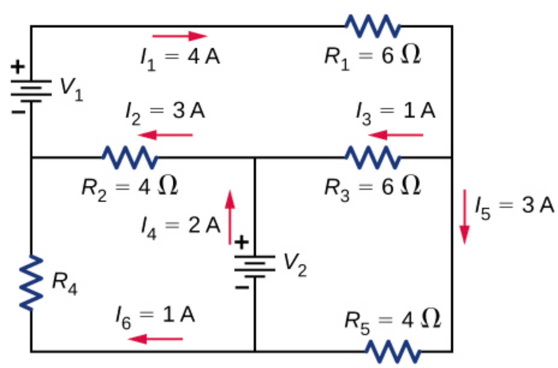


(b)

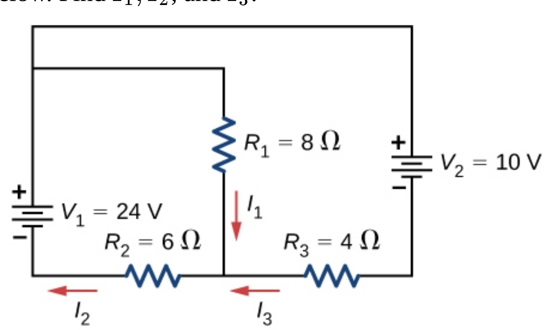
38. Consider the circuit shown below. Find V_1, I_2, and I_3.



39. Consider the circuit shown below. Find V_1, V_2, and R_4.



40. Consider the circuit shown below. Find I_1, I_2, and I_3.

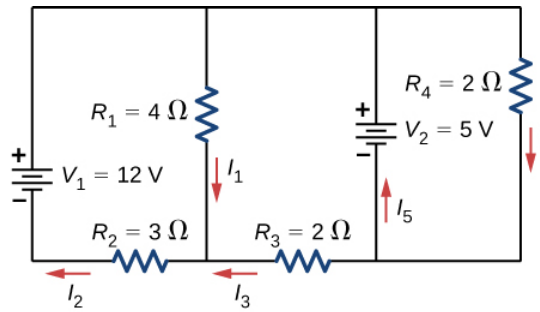


41. Consider the circuit shown below.

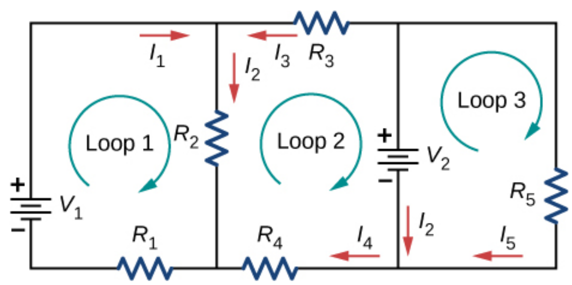
(a) Find I_1, I_2, I_3, I_4, and I_5.

(b) Find the power supplied by the voltage sources.

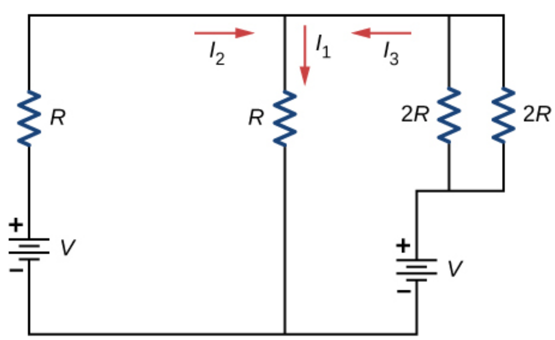
(c) Find the power dissipated by the resistors



42. Consider the circuit shown below. Write the three loop equations for the loops shown.



43. Consider the circuit shown below. Write equations for the three currents in terms of **R** and **V**.



44. Consider the circuit shown in the preceding problem. Write equations for the power supplied by the voltage sources and the power dissipated by the resistors in terms of R and V.

45. A child's electronic toy is supplied by three 1.58-V alkaline cells having internal resistances of **0.0200Ω** in series with a 1.53-V carbon-zinc dry cell having a **0.100-Ω** internal resistance. The load resistance is **10.0Ω**.

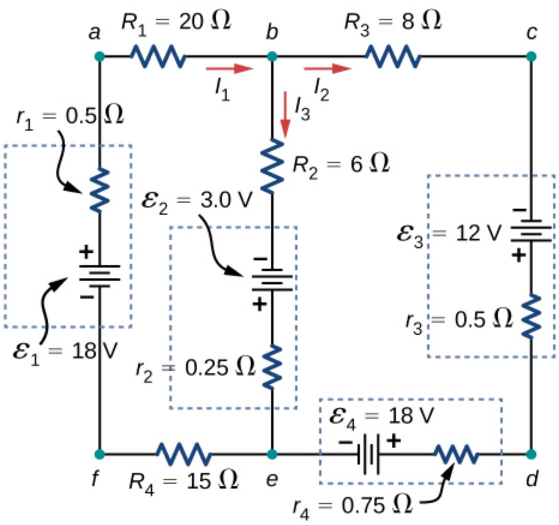
(a) Draw a circuit diagram of the toy and its batteries.

(b) What current flows?

(c) How much power is supplied to the load?

(d) What is the internal resistance of the dry cell if it goes bad, resulting in only 0.500 W being supplied to the load?

46. Apply the junction rule to Junction b shown below. Is any new information gained by applying the junction rule at e?



47. Apply the loop rule to Loop afedcba in the preceding problem.

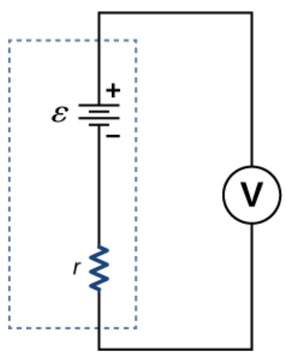
Application - Electrical Meters

48. Suppose you measure the terminal voltage of a 1.585-V alkaline cell having an internal resistance of **0.100Ω** by placing a **1.00-kΩ** voltmeter across its terminals (see below).

(a) What current flows?

(b) Find the terminal voltage.

(c) To see how close the measured terminal voltage is to the source voltage, calculate their ratio.



10.7 Household Wiring and Electrical Safety

58. (a) How much power is dissipated in a short circuit of 240-V ac through a resistance of **0.250Ω**? (b) What current flows?

59. What voltage is involved in a 1.44-kW short circuit through a **0.100-Ω** resistance?

60. Find the current through a person and identify the likely effect on her if she touches a 120-V ac source:

(a) if she is standing on a rubber mat and offers a total resistance of **300kΩ**;

(b) if she is standing barefoot on wet grass and has a resistance of only **4000kΩ**.

61. While taking a bath, a person touches the metal case of a radio. The path through the person to the drainpipe and ground has a resistance of **4000Ω**. What is the smallest voltage on the case of the radio that could cause ventricular fibrillation?

62. A man foolishly tries to fish a burning piece of bread from a toaster with a metal butter knife and comes into contact with 120-V ac. He does not even feel it since, luckily, he is wearing rubber-soled shoes. What is the minimum resistance of the path the current follows through the person?

63. (a) During surgery, a current as small as **20.0μA** applied directly to the heart may cause ventricular fibrillation. If the resistance of the exposed heart is **300Ω**, what is the smallest voltage that poses this danger?

(b) Does your answer imply that special electrical safety precautions are needed?

64. (a) What is the resistance of a 220-V ac short circuit that generates a peak power of 96.8 kW?

(b) What would the average power be if the voltage were 120 V ac?

65. A heart defibrillator passes 10.0 A through a patient's torso for 5.00 ms in an attempt to restore normal beating.

(a) How much charge passed?

(b) What voltage was applied if 500 J of energy was dissipated?

(c) What was the path's resistance? (d) Find the temperature increase caused in the 8.00 kg of affected tissue.

66. A short circuit in a 120-V appliance cord has a **0.500-Ω** resistance. Calculate the temperature rise of the 2.00 g of surrounding materials, assuming their specific heat capacity is **0.200cal/g·°C** and that it takes 0.0500 s for a circuit breaker to interrupt the current. Is this likely to be damaging?

Challenge Problems

88. Some camera flashes use flash tubes that require a high voltage. They obtain a high voltage by charging capacitors in parallel and then internally changing the connections of the capacitors to place them in series. Consider a circuit that uses four AAA batteries connected in series to charge six 10-mF capacitors through an equivalent resistance of **100Ω**. The connections are then switched internally to place the capacitors in series. The capacitors discharge through a lamp with a resistance of **100Ω**.

(a) What is the RC time constant and the initial current out of the batteries while they are connected in parallel?

(b) How long does it take for the capacitors to charge to **90%** of the terminal voltages of the batteries?

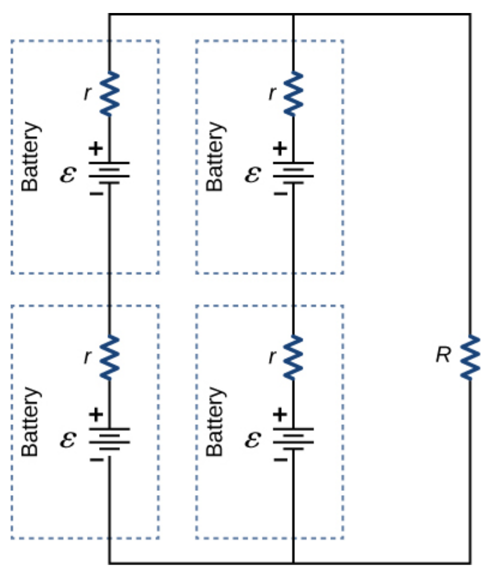
(c) What is the RC time constant and the initial current of the capacitors connected in series assuming it discharges at 90%90% of full charge?

(d) How long does it take the current to decrease to **10%** of the initial value?

89. Consider the circuit shown below. Each battery has a source voltage of 1.50 V and an internal resistance of **1.00Ω**.

(a) What is the current through the external resistor, which has a resistance of 10.00 ohms?

(b) What is the terminal voltage of each battery?



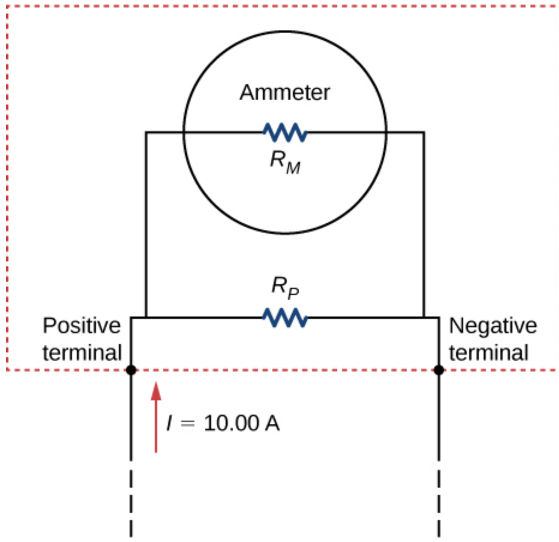
90. Analog meters use a galvanometer, which essentially consists of a coil of wire with a small resistance and a pointer with a scale attached. When current runs through the coil, the pointer turns; the amount the pointer turns is proportional to the amount of current running through the coil. Galvanometers can be used to make an ammeter if a resistor is placed in parallel with the galvanometer. Consider a galvanometer that has a resistance of **25.00Ω** and gives a full scale reading when a **50-μA**

current runs through it. The galvanometer is to be used to make an ammeter that has a full scale reading of 10.00 A, as shown below. Recall that an ammeter is connected in series with the circuit of interest, so all 10 A must run through the meter.

(a) What is the current through the parallel resistor in the meter?

(b) What is the voltage across the parallel resistor?

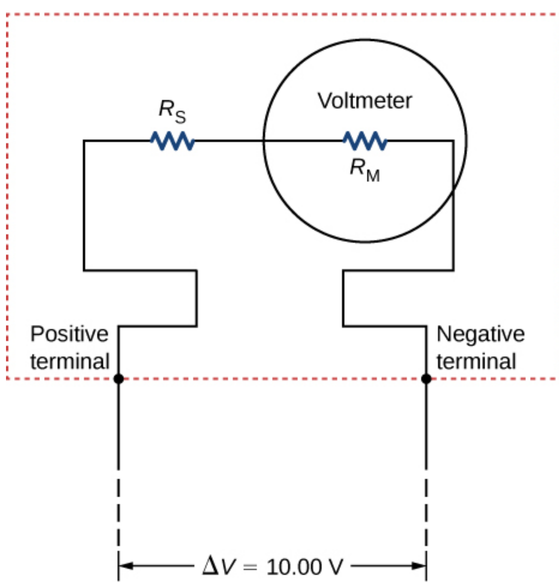
(c) What is the resistance of the parallel resistor?



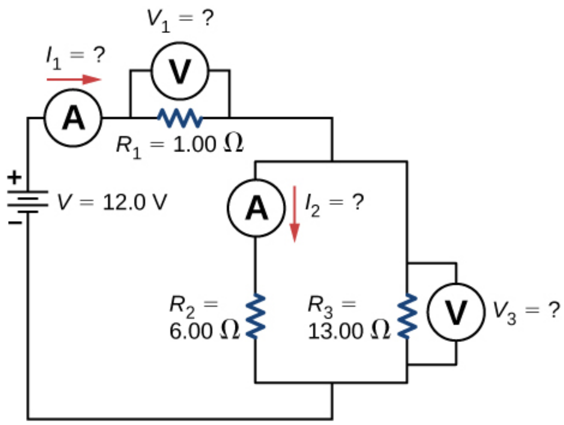
91. Analog meters use a galvanometer, which essentially consists of a coil of wire with a small resistance and a pointer with a scale attached. When current runs through the coil, the point turns; the amount the pointer turns is proportional to the amount of current running through the coil. Galvanometers can be used to make a voltmeter if a resistor is placed in series with the galvanometer. Consider a galvanometer that has a resistance of **25.00Ω** and gives a full scale reading when a **50-μA** current runs through it. The galvanometer is to be used to make an voltmeter that has a full scale reading of 10.00 V, as shown below. Recall that a voltmeter is connected in parallel with the component of interest, so the meter must have a high resistance or it will change the current running through the component.

(a) What is the potential drop across the series resistor in the meter?

(b) What is the resistance of the parallel resistor?



92. Consider the circuit shown below. Find I_1, V_1, I_2, and V_3.

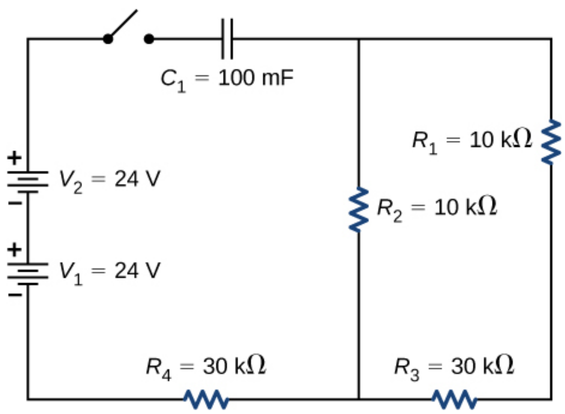


93. Consider the circuit below.

(a) What is the RC time constant of the circuit?

(b) What is the initial current in the circuit once the switch is closed?

(c) How much time passes between the instant the switch is closed and the time the current has reached half of the initial current?



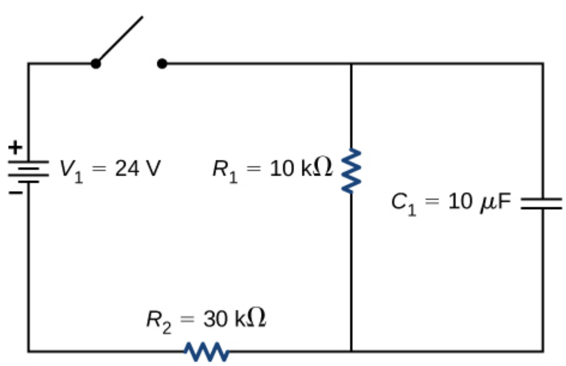
94. Consider the circuit below.

(a) What is the initial current through resistor R_2 when the switch is closed?

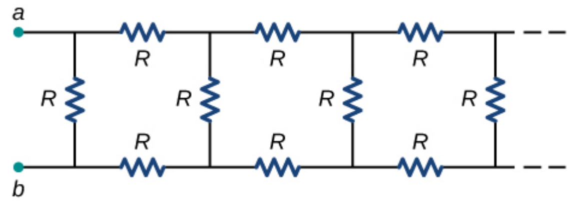
(b) What is the current through resistor R_2 when the capacitor is fully charged, long after the switch is closed?

(c) What happens if the switch is opened after it has been closed for some time?

(d) If the switch has been closed for a time period long enough for the capacitor to become fully charged, and then the switch is opened, how long before the current through resistor R_1 reaches half of its initial value?



95. Consider the infinitely long chain of resistors shown below. What is the resistance between terminals **a** and **b**?



96. Consider the circuit below. The capacitor has a capacitance of 10 mF. The switch is closed and after a long time the capacitor is fully charged.

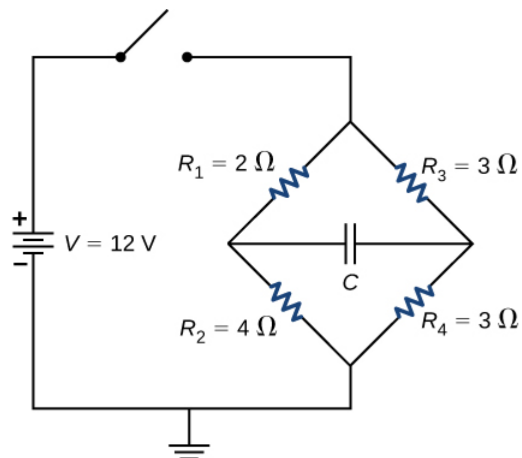
(a) What is the current through each resistor a long time after the switch is closed?

(b) What is the voltage across each resistor a long time after the switch is closed?

(c) What is the voltage across the capacitor a long time after the switch is closed?

(d) What is the charge on the capacitor a long time after the switch is closed?

(e) The switch is then opened. The capacitor discharges through the resistors. How long from the time before the current drops to one fifth of the initial value?



97. A 120-V immersion heater consists of a coil of wire that is placed in a cup to boil the water. The heater can boil one cup of **20.00°C** water in 180.00 seconds. You buy one to use in your dorm room, but you are worried that you will overload the circuit and trip the 15.00-A, 120-V circuit breaker, which supplies your dorm room. In your dorm room, you have four 100.00-W incandescent lamps and a 1500.00-W space heater.

(a) What is the power rating of the immersion heater?

(b) Will it trip the breaker when everything is turned on?

(c) If it you replace the incandescent bulbs with 18.00-W LED, will the breaker trip when everything is turned on?

98. Find the resistance that must be placed in series with a **25.0-Ω** galvanometer having a **50.0-μA** sensitivity (the same as the one discussed in the text) to allow it to be used as a voltmeter with a 3000-V full-scale reading. Include a circuit diagram with your solution.

99. Find the resistance that must be placed in parallel with a **60.0-Ω** galvanometer having a 1.00-mA sensitivity (the same as the one discussed in the text) to allow it to be used as an ammeter with a 25.0-A full-scale reading. Include a circuit diagram with your solution.

Contributors and Attributions

- **10.9: Direct-Current Circuits (Exercise)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.

6.9: Direct-Current Circuits (Answers)

Conceptual Questions

1. Some of the energy being used to recharge the battery will be dissipated as heat by the internal resistance.

3. $P = I^2R = (\frac{\varepsilon}{r+R})^2 R = \varepsilon^2 R(r+R)^{-2}, \frac{dP}{dR} = \varepsilon^2[(r+R)^{-2} - 2R(r+R)^{-3}] = 0 \qquad [\frac{(r+R)-2R}{(r+R)^3}] = 0, r = R$

5. It would probably be better to be in series because the current will be less than if it were in parallel.

7. two filaments, a low resistance and a high resistance, connected in parallel

9. It can be redrawn.

$$R_{eq} = [\frac{1}{R_6} + \frac{1}{R_1} + \frac{1}{R_2 + (\frac{1}{R_4} + \frac{1}{R_3+R+5})^{-1}}]^{-1} \ .$$

11. In series the voltages add, but so do the internal resistances, because the internal resistances are in series. In parallel, the terminal voltage is the same, but the equivalent internal resistance is smaller than the smallest individual internal resistance and a higher current can be provided.

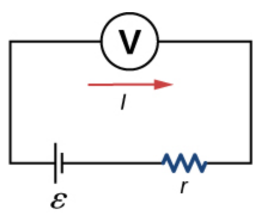
13. The voltmeter would put a large resistance in series with the circuit, significantly changing the circuit. It would probably give a reading, but it would be meaningless.

15. The ammeter has a small resistance; therefore, a large current will be produced and could damage the meter and/or overheat the battery.

19. Not only might water drip into the switch and cause a shock, but also the resistance of your body is lower when you are wet.

Problems

21. a.



b. 0.476W;

c. 0.691 W;

d. As R_L is lowered, the power difference decreases; therefore, at higher volumes, there is no significant difference.

23. a. 0.400Ω;

b. No, there is only one independent equation, so only r can be found.

25. a. 0.400Ω;

b. 40.0 W;

c. $0.0956^\circ C/min$

27. largest, 786Ω, smallest, 20.32Ω

29. 29.6 W

31. a. 0.74 A;

b. 0.742 A

33. a. 60.8 W;

b. 3.18 kW

35. a. $R_s = 9.00\Omega$;

b.$I_1 = I_2 = I_3 = 2.00A$;

c. $V_1 = 8.00V, V_2 = 2.00V, V_3 = 8.00V$;

d. $P_1 = 16.00W, P_2 = 4.00W, P_3 = 16.00W$;

e. $P = 36.00W$

37. a. $I_1 = 0.6mA, I_2 = 0.4mA, I_3 = 0.2m$;

b. $I_1 = 0.04mA, I_2 = 1.52mA, I_3 = -1.48mA$;

c. $P_{out} = 0.92mW, P_{out} = 4.50mW$;

d. $P_{in} = 0.92mW, P_{in} = 4.50mW$

39. $V_1 = 42V, V_2 = 6V, R_4 = 18\Omega$

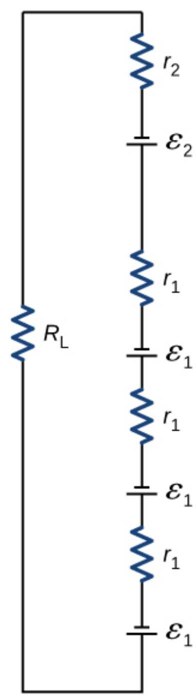
41. a. $I_1 = 1.5A, I_2 = 2A, I_3 = 0.5A, I_4 = 2.5A, I_5 = 2A$;

b. $P_{in} = I_2V_1 + I_5V_5 = 34W$;

c. $P_{out} = I_1^2R_1 + I_2^2R_2 + I_3^2R_3 + I_4^2R_4 = 34W$

43. $I_1 = \frac{2}{3}\frac{V}{R}, I_2 = \frac{1}{3}\frac{V}{R}, I_3 = \frac{1}{3}\frac{V}{R}$

45. a.



b. 0.617 A;

c. 3.81 W;

d. 18.0Ω

47. $I_1r_1 - \varepsilon_1 + I_1R_4 + \varepsilon_4 + I_2r_4 + I_4r_3 - \varepsilon_3 + I_2R_3 + I_1R_1 = 0$

59. 12.0 V

61. 400 V

63. a. 6.00 mV; .

b. It would not be necessary to take extra precautions regarding the power coming from the wall. However, it is possible to generate voltages of approximately this value from static charge built up on gloves, for instance, so some precautions are necessary.

65. a. $5.00 \times 10^{-2} C$;

b. 10.0 kV;

c. $1.00 k\Omega$;

d. $1.79 \times 10^{-2} {}^{\circ}C$

Challenge Problems

89. a. 0.273 A; b. $V_T = 1.36 V$

91. a. $V_s = V - I_M R_M = 9.99875 V$;

b. $R_S = \frac{V_P}{I_M} = 199.975 k\Omega$

93. a. $\tau = 3800 s$;

b. 1.26 A;

c. $t = 2633.96 s$

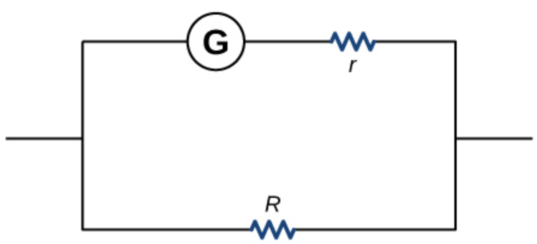
95. $R_{eq} = (\sqrt{3} - 1) R$

97. a. $P_{imheater} = \frac{1cup(\frac{0.000237m^3}{cup})(\frac{1000kg}{m^3})(4186\frac{J}{kg^{\circ}C})(100^{\circ}C - 20^{\circ}C)}{180.00s} \approx 441W$,

b. $I = \frac{441W}{120V} + 4(\frac{100W}{120V}) + \frac{1500W}{120V} = 19.51A$; Yes, the breaker will trip.

c. $I = \frac{441W}{120V} + 4(\frac{18W}{120V}) + \frac{1500W}{120V} = 13.47$; No, the breaker will not trip.

99. $2.40 \times 10^{-3} \Omega$



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).

CHAPTER OVERVIEW

7: Capacitance

7.1: Introduction

Capacitors are important components of electrical circuits in many electronic devices, including pacemakers, cell phones, and computers. In this chapter, we study their properties, and, over the next few chapters, we examine their function in combination with other circuit elements. By themselves, capacitors are often used to store electrical energy and release it when needed; with other circuit components, capacitors often act as part of a filter that allows some electrical signals to pass while blocking others. You can see why capacitors are considered one of the fundamental components of electrical circuits.

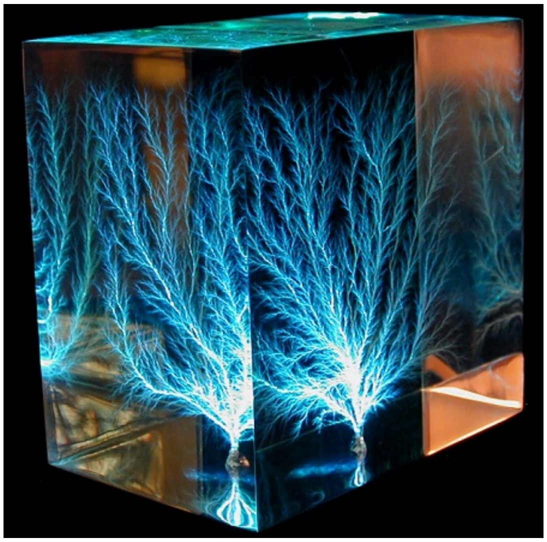


Figure 7.1.1: The tree-like branch patterns in this clear acrylic block are created by irradiating the block with an electron beam. This tree is known as a Lichtenberg figure, named for the German physicist Georg Christof Lichtenberg (1742–1799), who was the first to study these patterns. The “branches” are created by the dielectric breakdown produced by a strong electric field. (Bert Hickman).

7.2: Capacitors and Capacitance

Learning Objectives

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

- Explain the concepts of a capacitor and its capacitance
- Describe how to evaluate the capacitance of a system of conductors

A **capacitor** is a device used to store electrical charge and electrical energy. It consists of at least two electrical conductors separated by a distance. (Note that such electrical conductors are sometimes referred to as “electrodes,” but more correctly, they are “capacitor plates.”) The space between capacitors may simply be a vacuum, and, in that case, a capacitor is then known as a “vacuum capacitor.” However, the space is usually filled with an insulating material known as a **dielectric**. (You will learn more about dielectrics in the sections on dielectrics later in this chapter.) The amount of storage in a capacitor is determined by a property called **capacitance**, which you will learn more about a bit later in this section.

Capacitors have applications ranging from filtering static from radio reception to energy storage in heart defibrillators. Typically, commercial capacitors have two conducting parts close to one another but not touching, such as those in Figure $7.2.1$. Most of the time, a dielectric is used between the two plates. When battery terminals are connected to an initially uncharged capacitor, the battery potential moves a small amount of charge of magnitude Q from the positive plate to the negative plate. The capacitor remains neutral overall, but with charges $+Q$ and $-Q$ residing on opposite plates.

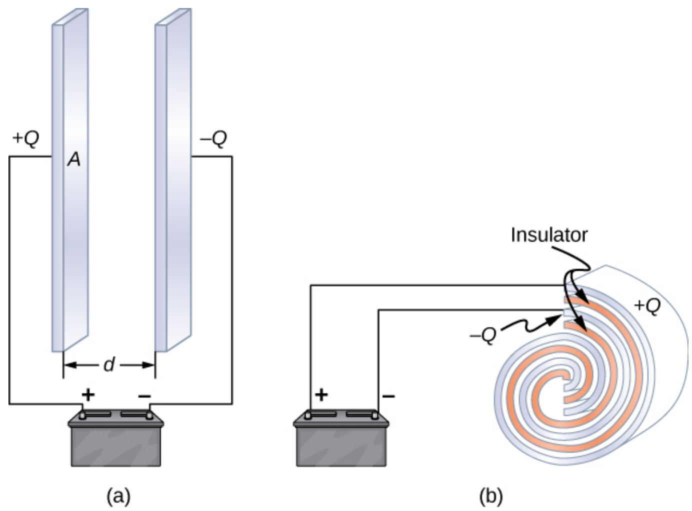


Figure $7.2.1$: Both capacitors shown here were initially uncharged before being connected to a battery. They now have charges of $+Q$ and $-Q$ (respectively) on their plates. (a) A parallel-plate capacitor consists of two plates of opposite charge with area A separated by distance d. (b) A rolled capacitor has a dielectric material between its two conducting sheets (plates).

A system composed of two identical parallel-conducting plates separated by a distance is called a **parallel-plate capacitor** (Figure $7.2.2$). The magnitude of the electrical field in the space between the parallel plates is $E=\sigma/\epsilon_0$, where σ denotes the surface charge density on one plate (recall that σ is the charge **Q** per the surface area **A**). Thus, the magnitude of the field is directly proportional to **Q**.

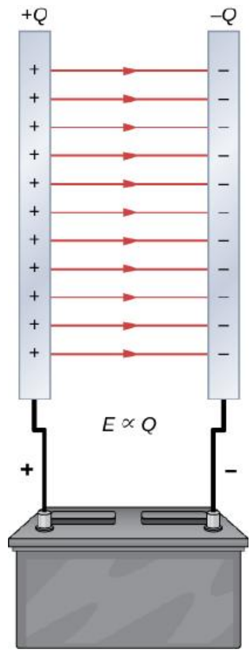


Figure $7.2.2$: The charge separation in a capacitor shows that the charges remain on the surfaces of the capacitor plates. Electrical field lines in a parallel-plate capacitor begin with positive charges and end with negative charges. The magnitude of the electrical field in the space between the plates is in direct proportion to the amount of charge on the capacitor.

Capacitors with different physical characteristics (such as shape and size of their plates) store different amounts of charge for the same applied voltage V across their plates. The **capacitance** C of a capacitor is defined as the ratio of the maximum charge Q that can be stored in a capacitor to the applied voltage V across its plates. In other words, capacitance is the largest amount of charge per volt that can be stored on the device:

$$C = \frac{Q}{V} \tag{7.2.1}$$

The SI unit of capacitance is the **farad** (F), named after Michael Faraday (1791–1867). Since capacitance is the charge per unit voltage, one farad is one coulomb per one volt, or

$$1\,F = \frac{1\,C}{1\,V}.$$

By definition, a 1.0-F capacitor is able to store 1.0 C of charge (a very large amount of charge) when the potential difference between its plates is only 1.0 V. One farad is therefore a very large capacitance. Typical capacitance values range from picofarads ($1\,pF = 10{-}12F$) to millifarads $(1\,mF = 10^{-3}F)$, which also includes microfarads $(1\,\mu C = 10^{-6}F)$.. Capacitors can be produced in various shapes and sizes (Figure $7.2.3$).

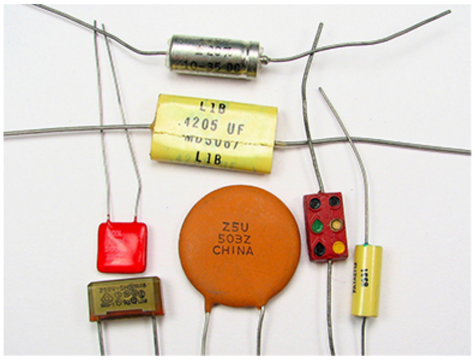


Figure $7.2.3$: These are some typical capacitors used in electronic devices. A capacitor's size is not necessarily related to its capacitance value.

Calculation of Capacitance

We can calculate the capacitance of a pair of conductors with the standard approach that follows.

> **Problem-Solving Strategy: Calculating Capacitance**
>
> 1. Assume that the capacitor has a charge Q.
> 2. Determine the electrical field $\vec{E}$ between the conductors. If symmetry is present in the arrangement of conductors, you may be able to use Gauss's law for this calculation.
> 3. Find the potential difference between the conductors from
>
> $$V_B - V_A = -\int_A^B \vec{E} \cdot d\vec{l}, \tag{7.2.2}$$
>
> where the path of integration leads from one conductor to the other. The magnitude of the potential difference is then $V = |V_B - V_A|$.
> 4. With V known, obtain the capacitance directly from Equation 7.2.1.

To show how this procedure works, we now calculate the capacitances of parallel-plate, spherical, and cylindrical capacitors. In all cases, we assume vacuum capacitors (empty capacitors) with no dielectric substance in the space between conductors.

Parallel-Plate Capacitor

The parallel-plate capacitor (Figure $7.2.4$) has two identical conducting plates, each having a surface area A, separated by a distance d. When a voltage V is applied to the capacitor, it stores a charge Q, as shown. We can see how its capacitance may depend on A and d by considering characteristics of the Coulomb force. We know that force between the charges increases with charge values and decreases with the distance between them. We should expect that the bigger the plates are, the more charge they can store. Thus, C should be greater for a larger value of A. Similarly, the closer the plates are together, the greater the attraction of the opposite charges on them. Therefore, C should be greater for a smaller d.

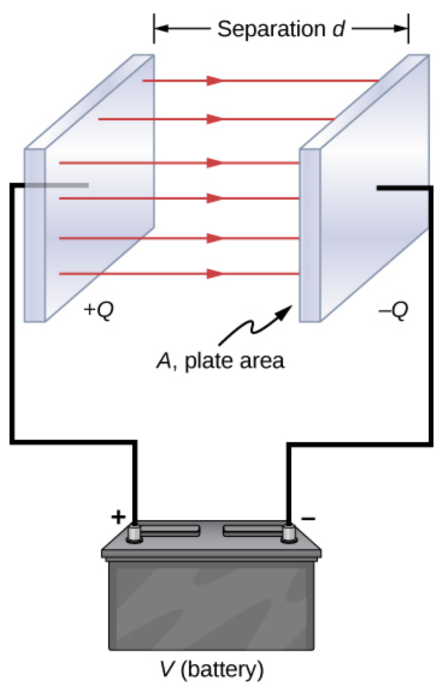


Figure $7.2.4$: In a parallel-plate capacitor with plates separated by a distance \(**d**\), each plate has the same surface area \(**A**\).

We define the surface charge density σ on the plates as

$$\sigma = \frac{Q}{A}.$$

We know from previous chapters that when d is small, the electrical field between the plates is fairly uniform (ignoring edge effects) and that its magnitude is given by

$$E = \frac{\sigma}{\epsilon_0},$$

where the constant ϵ_0 is the permittivity of free space, $\epsilon_0 = 8.85 \times 10^{-12} F/m$. The SI unit of F/m is equivalent to $C^2/N \cdot m^2$. Since the electrical field $\vec{E}$ between the plates is uniform, the potential difference between the plates is

$$V = Ed = \frac{\sigma d}{\epsilon_0} = \frac{Qd}{\epsilon_0 A}.$$

Therefore Equation 7.2.1 gives the capacitance of a parallel-plate capacitor as

$$C = \frac{Q}{V} = \frac{Q}{Qd/\epsilon_0 A} = \epsilon_0 \frac{A}{d}. \tag{7.2.3}$$

Notice from this equation that capacitance is a function **only of the geometry** and what material fills the space between the plates (in this case, vacuum) of this capacitor. In fact, this is true not only for a parallel-plate capacitor, but for all capacitors: The capacitance is independent of Q or V. If the charge changes, the potential changes correspondingly so that Q/V remains constant.

Example 7.2.1A: Capacitance and Charge Stored in a Parallel-Plate Capacitor

a. What is the capacitance of an empty parallel-plate capacitor with metal plates that each have an area of $1.00\, m^2$, separated by 1.00 mm?
b. How much charge is stored in this capacitor if a voltage of $3.00 \times 10^3 V$ is applied to it?

Strategy

Finding the capacitance C is a straightforward application of Equation 7.2.3. Once we find C, we can find the charge stored by using Equation 7.2.1.

Solution

1. Entering the given values into Equation 7.2.3 yields

$$C = \epsilon_0 \frac{A}{d} = \left(8.85 \times 10^{-12} \frac{F}{m}\right) \frac{1.00\, m^2}{1.00 \times 10^{-3} m} = 8.85 \times 10^{-9} F = 8.85\, nF.$$

This small capacitance value indicates how difficult it is to make a device with a large capacitance.

2. Inverting Equation 7.2.1 and entering the known values into this equation gives

$$Q = CV = (8.85 \times 10^{-9} F)(3.00 \times 10^3 V) = 26.6\, \mu C.$$

Significance

This charge is only slightly greater than those found in typical static electricity applications. Since air breaks down (becomes conductive) at an electrical field strength of about 3.0 MV/m, no more charge can be stored on this capacitor by increasing the voltage.

Example 7.2.1B: A 1-F Parallel-Plate Capacitor

Suppose you wish to construct a parallel-plate capacitor with a capacitance of 1.0 F. What area must you use for each plate if the plates are separated by 1.0 mm?

Solution

Rearranging Equation 7.2.3, we obtain

$$A = \frac{Cd}{\epsilon_0} = \frac{(1.0\, F)(1.0 \times 10^{-3} m)}{8.85 \times 10^{-12} F/m} = 1.1 \times 10^8\, m^2.$$

Each square plate would have to be 10 km across. It used to be a common prank to ask a student to go to the laboratory stockroom and request a 1-F parallel-plate capacitor, until stockroom attendants got tired of the joke.

Exercise $7.2.1A$

The capacitance of a parallel-plate capacitor is 2.0 pF. If the area of each plate is $2.4\ cm^2$, what is the plate separation?

Answer

$1.1 \times 10^{-3} m$

Exercise $7.2.1B$

Verify that σ/V and ϵ_0/d have the same physical units.

Spherical Capacitor

A spherical capacitor is another set of conductors whose capacitance can be easily determined (Figure $7.2.5$). It consists of two concentric conducting spherical shells of radii R_1 (inner shell) and R_2 (outer shell). The shells are given equal and opposite charges $+Q$ and $-Q$, respectively. From symmetry, the electrical field between the shells is directed radially outward. We can obtain the magnitude of the field by applying Gauss's law over a spherical Gaussian surface of radius **r** concentric with the shells. The enclosed charge is $+Q$; therefore we have

$$\oint_S \vec{E} \cdot \hat{n} dA = E(4\pi r^2) = \frac{Q}{\epsilon_0}.$$

Thus, the electrical field between the conductors is

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \hat{r}.$$

We substitute this $\vec{E}$ into Equation 7.2.2 and integrate along a radial path between the shells:

$$V = \int_{R_1}^{R_2} \vec{E} \cdot d\vec{l} = \int_{R_1}^{R_2} \left(\frac{1}{4\pi\epsilon_0} \frac{Q}{r^2} \hat{r} \right) \cdot (\hat{r} dr) = \frac{Q}{4\pi\epsilon_0} \int_{R_1}^{R_2} \frac{dr}{r^2} = \frac{Q}{4\pi\epsilon_0} \left(\frac{1}{R_1} - \frac{1}{R_2} \right).$$

In this equation, the potential difference between the plates is

$$V = -(V_2 - V_1) = V_1 - V_2.$$

We substitute this result into Equation 7.2.1 to find the capacitance of a spherical capacitor:

$$C = \frac{Q}{V} = 4\pi\epsilon_0 \frac{R_1 R_2}{R_2 - R_1}. \tag{7.2.4}$$

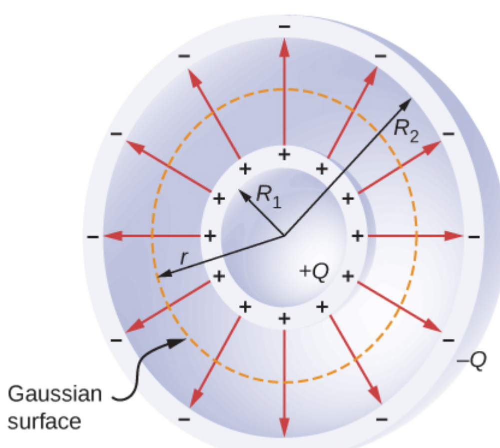


Figure $7.2.5$: A spherical capacitor consists of two concentric conducting spheres. Note that the charges on a conductor reside on its surface.

Example 7.2.2: Capacitance of an Isolated Sphere

Calculate the capacitance of a single isolated conducting sphere of radius R_1 and compare it with Equation 7.2.4 in the limit as $R_2 \to \infty$.

Strategy

We assume that the charge on the sphere is Q, and so we follow the four steps outlined earlier. We also assume the other conductor to be a concentric hollow sphere of infinite radius.

Solution

On the outside of an isolated conducting sphere, the electrical field is given by Equation 7.2.2. The magnitude of the potential difference between the surface of an isolated sphere and infinity is

$$\begin{aligned} V &= \int_{R_1}^{+\infty} \vec{E} \cdot d\vec{l} \\ &= \frac{Q}{4\pi\epsilon_0} \int_{R_1}^{+\infty} \frac{1}{r^2} \hat{r} \cdot (\hat{r}\, dr) \\ &= \frac{Q}{4\pi\epsilon_0} \int_{R_1}^{+\infty} \frac{dr}{r^2} \\ &= \frac{1}{4\pi\epsilon_0} \frac{Q}{R_1} \end{aligned}$$

The capacitance of an isolated sphere is therefore

$$C = \frac{Q}{V} = Q \frac{4\pi\epsilon_0 R_1}{Q} = 4\pi\epsilon_0 R_1.$$

Significance

The same result can be obtained by taking the limit of Equation 7.2.4 as $R_2 \to \infty$. A single isolated sphere is therefore equivalent to a spherical capacitor whose outer shell has an infinitely large radius.

Exercise 7.2.2

The radius of the outer sphere of a spherical capacitor is five times the radius of its inner shell. What are the dimensions of this capacitor if its capacitance is 5.00 pF?

Answer

3.59 cm, 17.98 cm

Cylindrical Capacitor

A cylindrical capacitor consists of two concentric, conducting cylinders (Figure $7.2.6$). The inner cylinder, of radius R_1, may either be a shell or be completely solid. The outer cylinder is a shell of inner radius R_2. We assume that the length of each cylinder is **l** and that the excess charges $+Q$ and $-Q$ reside on the inner and outer cylinders, respectively.

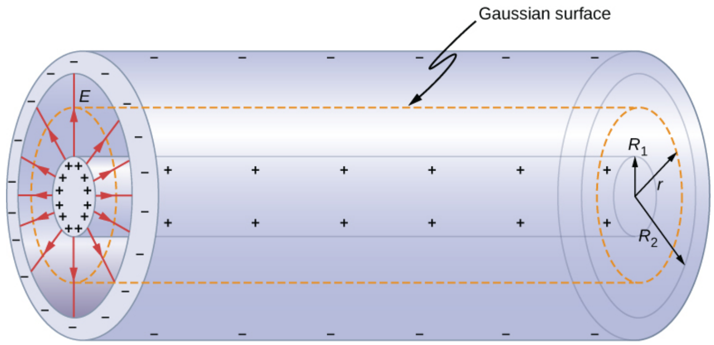


Figure 7.2.6: A cylindrical capacitor consists of two concentric, conducting cylinders. Here, the charge on the outer surface of the inner cylinder is positive (indicated by $+$) and the charge on the inner surface of the outer cylinder is negative (indicated by $-$).

With edge effects ignored, the electrical field between the conductors is directed radially outward from the common axis of the cylinders. Using the Gaussian surface shown in Figure 7.2.6, we have

$$\oint_S \vec{E} \cdot \hat{n} dA = E(2\pi r l) = \frac{Q}{\epsilon_0}.$$

Therefore, the electrical field between the cylinders is

$$\vec{E} = \frac{1}{2\pi\epsilon_0} \frac{Q}{r\,l} \hat{r}.$$

where $\hat{r}$ is the unit radial vector along the radius of the cylinder. We can substitute into Equation 7.2.2 and find the potential difference between the cylinders:

$$V = \int_{R_1}^{R_2} \vec{E} \cdot d\vec{l}_p = \frac{Q}{2\pi\epsilon_0 l} \int_{R_1}^{R_2} \frac{1}{r} \hat{r} \cdot (\hat{r}\, dr) = \frac{Q}{2\pi\epsilon_0 l} \int_{R_1}^{R_2} \frac{dr}{r} = \frac{Q}{2\pi\epsilon_0 l} \ln r \Big|_{R_1}^{R_2} = \frac{Q}{2\pi\epsilon_0 l} \frac{R_2}{R_1}.$$

Thus, the capacitance of a cylindrical capacitor is

$$C = \frac{Q}{V} = \frac{2\pi\epsilon_0\, l}{\ln(R_2/R_1)}. \tag{7.2.5}$$

As in other cases, this capacitance depends only on the geometry of the conductor arrangement. An important application of Equation 7.2.5 is the determination of the capacitance per unit length of a **coaxial cable**, which is commonly used to transmit time-varying electrical signals. A coaxial cable consists of two concentric, cylindrical conductors separated by an insulating material. (Here, we assume a vacuum between the conductors, but the physics is qualitatively almost the same when the space between the conductors is filled by a dielectric.) This configuration shields the electrical signal propagating down the inner conductor from stray electrical fields external to the cable. Current flows in opposite directions in the inner and the outer conductors, with the outer conductor usually grounded. Now, from Equation 7.2.5, the capacitance per unit length of the coaxial cable is given by

$$\frac{C}{l} = \frac{2\pi\epsilon_0}{\ln(R_2/R_1)}.$$

In practical applications, it is important to select specific values of C/l. This can be accomplished with appropriate choices of radii of the conductors and of the insulating material between them.

? Exercise 7.2.3

When a cylindrical capacitor is given a charge of 0.500 nC, a potential difference of 20.0 V is measured between the cylinders.

a. What is the capacitance of this system?
b. If the cylinders are 1.0 m long, what is the ratio of their radii?

Answer

a. 25.0 pF

b. 9.2

Several types of practical capacitors are shown in Figure $7.2.3$. Common capacitors are often made of two small pieces of metal foil separated by two small pieces of insulation (Figure $7.2.1b$). The metal foil and insulation are encased in a protective coating, and two metal leads are used for connecting the foils to an external circuit. Some common insulating materials are mica, ceramic, paper, and Teflon™ non-stick coating.

Another popular type of capacitor is an **electrolytic capacitor**. It consists of an oxidized metal in a conducting paste. The main advantage of an electrolytic capacitor is its high capacitance relative to other common types of capacitors. For example, capacitance of one type of aluminum electrolytic capacitor can be as high as 1.0 F. However, you must be careful when using an electrolytic capacitor in a circuit, because it only functions correctly when the metal foil is at a higher potential than the conducting paste. When reverse polarization occurs, electrolytic action destroys the oxide film. This type of capacitor cannot be connected across an alternating current source, because half of the time, ac voltage would have the wrong polarity, as an alternating current reverses its polarity (see Alternating-Current Circuits on alternating-current circuits).

A **variable air capacitor** (Figure $7.2.7$) has two sets of parallel plates. One set of plates is fixed (indicated as "stator"), and the other set of plates is attached to a shaft that can be rotated (indicated as "rotor"). By turning the shaft, the cross-sectional area in the overlap of the plates can be changed; therefore, the capacitance of this system can be tuned to a desired value. Capacitor tuning has applications in any type of radio transmission and in receiving radio signals from electronic devices. Any time you tune your car radio to your favorite station, think of capacitance.

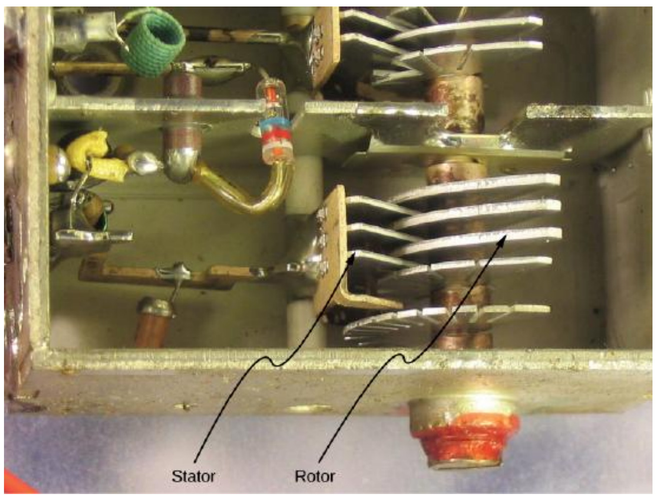


Figure $7.2.7$: In a variable air capacitor, capacitance can be tuned by changing the effective area of the plates. (credit: modification of work by Robbie Sproule)

The symbols shown in Figure $7.2.8$ are circuit representations of various types of capacitors. We generally use the symbol shown in Figure $7.2.8a$. The symbol in Figure $7.2.8c$ represents a variable-capacitance capacitor. Notice the similarity of these symbols to the symmetry of a parallel-plate capacitor. An electrolytic capacitor is represented by the symbol in part Figure $7.2.8b$, where the curved plate indicates the negative terminal.

Figure a shows two vertical lines. Figure b shows a vertical line to the left and another, slightly curved vertical line to the right. Figure c shows two vertical lines and an arrow cutting across them diagonally. In all figures, each line is connected to a horizontal line on the outside.

Figure $7.2.8$: This shows three different circuit representations of capacitors. The symbol in (a) is the most commonly used one. The symbol in (b) represents an electrolytic capacitor. The symbol in (c) represents a variable-capacitance capacitor.

An interesting applied example of a capacitor model comes from cell biology and deals with the electrical potential in the plasma membrane of a living cell (Figure $7.2.9$). **Cell membranes** separate cells from their surroundings, but allow some selected ions to pass in or out of the cell. The potential difference across a membrane is about 70 mV. The cell membrane may be 7 to 10 nm thick. Treating the cell membrane as a nano-sized capacitor, the estimate of the smallest electrical field strength across its 'plates' yields the value

$$E = \frac{V}{d} = \frac{70 \times 10^{-3} V}{10 \times 10^{-9} m} = 7 \times 10^{6} V/m > 3\, MV/m.$$

This magnitude of electrical field is great enough to create an electrical spark in the air.

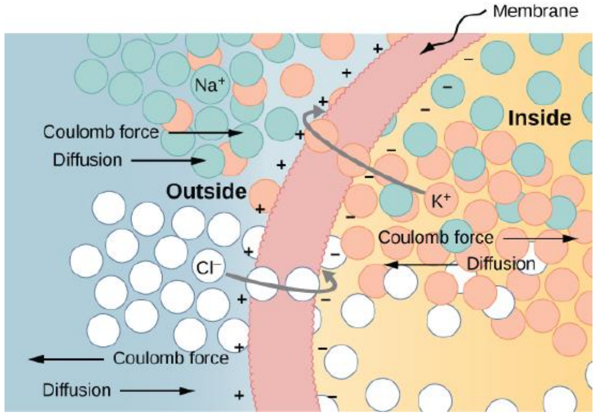


Figure 7.2.9: The semipermeable membrane of a biological cell has different concentrations of ions on its interior surface than on its exterior. Diffusion moves the K^+ (potassium) and Cl^- (chloride) ions in the directions shown, until the Coulomb force halts further transfer. In this way, the exterior of the membrane acquires a positive charge and its interior surface acquires a negative charge, creating a potential difference across the membrane. The membrane is normally impermeable to Na^+ (sodium ions).

Simulation

Visit the PhET Explorations: Capacitor Lab to explore how a capacitor works. Change the size of the plates and add a dielectric to see the effect on capacitance. Change the voltage and see charges built up on the plates. Observe the electrical field in the capacitor. Measure the voltage and the electrical field.

7.3: Capacitors in Series and in Parallel

Learning Objectives

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

- Explain how to determine the equivalent capacitance of capacitors in series and in parallel combinations
- Compute the potential difference across the plates and the charge on the plates for a capacitor in a network and determine the net capacitance of a network of capacitors

Several capacitors can be connected together to be used in a variety of applications. Multiple connections of capacitors behave as a single equivalent capacitor. The total capacitance of this equivalent single capacitor depends both on the individual capacitors and how they are connected. Capacitors can be arranged in two simple and common types of connections, known as **series** and **parallel**, for which we can easily calculate the total capacitance. These two basic combinations, series and parallel, can also be used as part of more complex connections.

The Series Combination of Capacitors

Figure $7.3.1$ illustrates a series combination of three capacitors, arranged in a row within the circuit. As for any capacitor, the capacitance of the combination is related to both charge and voltage:

$$C=\frac{Q}{V}. \tag{7.3.1}$$

When this series combination is connected to a battery with voltage **V**, each of the capacitors acquires an identical charge **Q**. To explain, first note that the charge on the plate connected to the positive terminal of the battery is $+Q$ and the charge on the plate connected to the negative terminal is $-Q$. Charges are then induced on the other plates so that the sum of the charges on all plates, and the sum of charges on any pair of capacitor plates, is zero. However, the potential drop $V_1=Q/C_1$ on one capacitor may be different from the potential drop $V_2=Q/C_2$ on another capacitor, because, generally, the capacitors may have different capacitances. The series combination of two or three capacitors resembles a single capacitor with a smaller capacitance. Generally, any number of capacitors connected in series is equivalent to one capacitor whose capacitance (called the **equivalent capacitance**) is smaller than the smallest of the capacitances in the series combination. Charge on this equivalent capacitor is the same as the charge on any capacitor in a series combination: That is, **all capacitors of a series combination have the same charge**. This occurs due to the conservation of charge in the circuit. When a charge **Q** in a series circuit is removed from a plate of the first capacitor (which we denote as $-Q$), it must be placed on a plate of the second capacitor (which we denote as $+Q$), and so on.

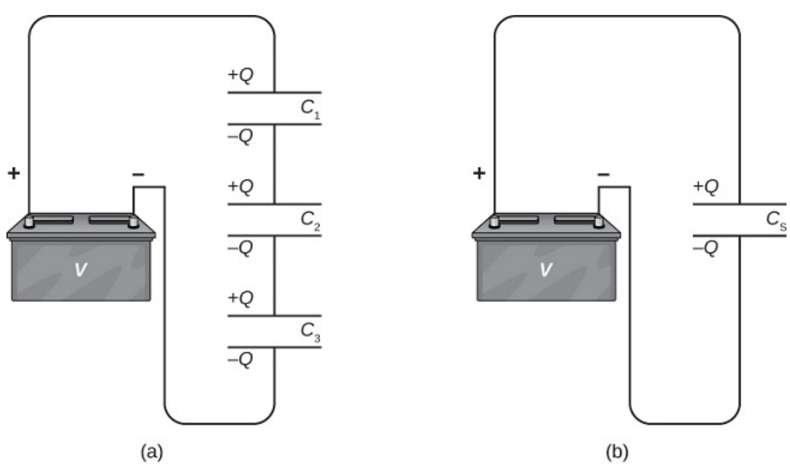


Figure $7.3.1$: (a) Three capacitors are connected in series. The magnitude of the charge on each plate is Q. (b) The network of capacitors in (a) is equivalent to one capacitor that has a smaller capacitance than any of the individual capacitances in (a), and the charge on its plates is Q.

We can find an expression for the total (equivalent) capacitance by considering the voltages across the individual capacitors. The potentials across capacitors 1, 2, and 3 are, respectively, $V_1=Q/C_1$, $V_2=Q/C_2$, and $V_3=Q/C_3$. These potentials must sum up to the voltage of the battery, giving the following potential balance:

$$V = V_1 + V_2 + V_3. \tag{7.3.2}$$

Potential V is measured across an equivalent capacitor that holds charge Q and has an equivalent capacitance C_S. Entering the expressions for V_1, V_2, and V_3, we get

$$\frac{Q}{C_S} = \frac{Q}{C_1} + \frac{Q}{C_2} + \frac{Q}{C_3}. \tag{7.3.3}$$

Canceling the charge **Q**, we obtain an expression containing the equivalent capacitance, C_S, of three capacitors connected in series:

$$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3}. \tag{7.3.4}$$

This expression can be generalized to any number of capacitors in a series network.

Series Combination

For capacitors connected in a series combination, the reciprocal of the equivalent capacitance is the sum of reciprocals of individual capacitances:

$$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \dots \tag{7.3.5}$$

Example 7.3.1: Equivalent Capacitance of a Series Network

Find the total capacitance for three capacitors connected in series, given their individual capacitances are $1.000 \mu F$, $5.000 \mu F$, and $8.000 \mu F$.

Strategy

Because there are only three capacitors in this network, we can find the equivalent capacitance by using Equation 7.3.5 with three terms.

Solution

We enter the given capacitances into Equation 7.3.5:

$$\begin{aligned} \frac{1}{C_S} &= \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \\ &= \frac{1}{1.000 \mu F} + \frac{1}{5.000 \mu F} + \frac{1}{8.000 \mu F} \\ &= \frac{1.325}{\mu F}. \end{aligned}$$

Now we invert this result and obtain

$$\begin{aligned} C_S &= \frac{\mu F}{1.325} \\ &= 0.755 \mu F. \end{aligned}$$

Significance

Note that in a series network of capacitors, the equivalent capacitance is always less than the smallest individual capacitance in the network.

The Parallel Combination of Capacitors

A parallel combination of three capacitors, with one plate of each capacitor connected to one side of the circuit and the other plate connected to the other side, is illustrated in Figure $7.3.2a$. Since the capacitors are connected in parallel, **they all have the same voltage V across their plates**. However, each capacitor in the parallel network may store a different charge. To find the equivalent capacitance C_p of the parallel network, we note that the total charge **Q** stored by the network is the sum of all the individual charges:

$$Q = Q_1 + Q_2 + Q_3. \tag{7.3.6}$$

On the left-hand side of this equation, we use the relation $Q = C_pV$, which holds for the entire network. On the right-hand side of the equation, we use the relations $Q_1 = C_1V$, $Q_2 = C_2V$, and $Q_3 = C_3V$ for the three capacitors in the network. In this way we obtain

$$C_pV = C_1V + C_2V + C_3V. \tag{7.3.7}$$

This equation, when simplified, is the expression for the equivalent capacitance of the parallel network of three capacitors:

$$C_p = C_1 + C_2 + C_3. \tag{7.3.8}$$

This expression is easily generalized to any number of capacitors connected in parallel in the network.

Parallel Combination

For capacitors connected in a parallel combination, the equivalent (net) capacitance is the sum of all individual capacitances in the network,

$$C_p = C_1 + C_2 + C_3 + \ldots \tag{7.3.9}$$

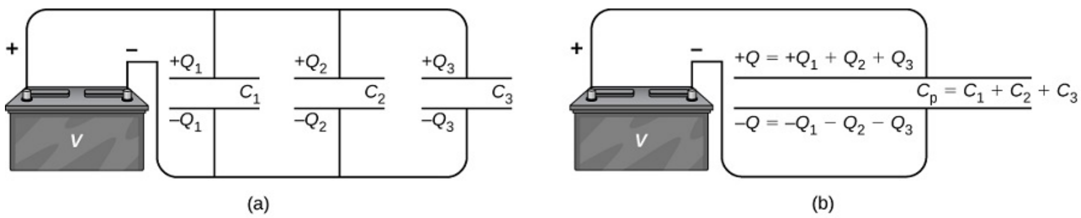


Figure 7.3.2: (a) Three capacitors are connected in parallel. Each capacitor is connected directly to the battery. (b) The charge on the equivalent capacitor is the sum of the charges on the individual capacitors.

Example 7.3.2: Equivalent Capacitance of a Parallel Network

Find the net capacitance for three capacitors connected in parallel, given their individual capacitances are $1.0 \mu F$, $5.0 \mu F$, and $8.0 \mu F$.

Strategy

Because there are only three capacitors in this network, we can find the equivalent capacitance by using Equation 7.3.9 with three terms.

Solution

Entering the given capacitances into Equation 7.3.9 yields

$$\begin{aligned} C_p &= C_1 + C_2 + C_3 \\ &= 1.0 \mu F + 5.0 \mu F + 8.0 \mu F \\ &= 14.0 \mu F. \end{aligned}$$

Significance

Note that in a parallel network of capacitors, the equivalent capacitance is always larger than any of the individual capacitances in the network.

Capacitor networks are usually some combination of series and parallel connections, as shown in Figure 7.3.3. To find the net capacitance of such combinations, we identify parts that contain only series or only parallel connections, and find their equivalent capacitances. We repeat this process until we can determine the equivalent capacitance of the entire network. The following example illustrates this process.

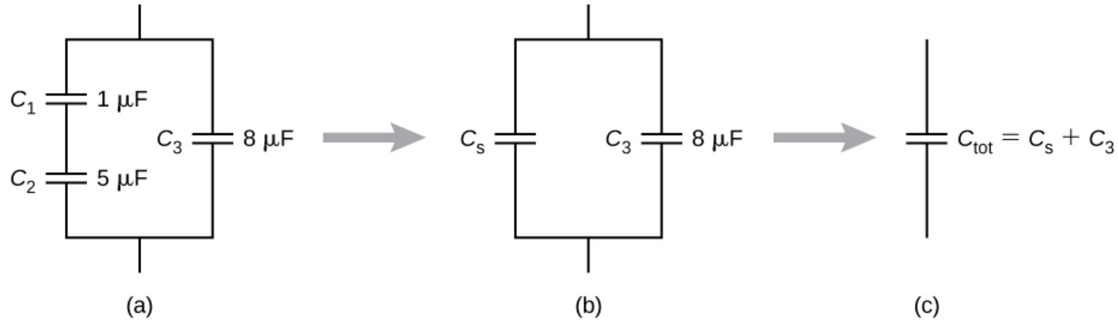


Figure 7.3.3: (a) This circuit contains both series and parallel connections of capacitors. (b) C_1 and C_2 are in series; their equivalent capacitance is C_S c) The equivalent capacitance C_S is connected in parallel with C_3. Thus, the equivalent capacitance of the entire network is the sum of C_S and C_3.

Example 7.3.3: Equivalent Capacitance of a Network

Find the total capacitance of the combination of capacitors shown in Figure 7.3.3. Assume the capacitances are known to three decimal places ($C_1 = 1.000 \mu F, C_2 = 5.000 \mu F, C_3 = 8.000 \mu F$). Round your answer to three decimal places.

Strategy

We first identify which capacitors are in series and which are in parallel. Capacitors C_1 and C_2 are in series. Their combination, labeled C_S is in parallel with C_3.

Solution

Since C_1 and C_2 are in series, their equivalent capacitance C_S is obtained with Equation 7.3.5:

$$\begin{aligned} \frac{1}{C_S} &= \frac{1}{C_1} + \frac{1}{C_2} \\ &= \frac{1}{1.000 \mu F} + \frac{1}{5.000 \mu F} \\ &= \frac{1.200}{\mu F} \end{aligned}$$

Therefor

$$C_S = 0.833 \mu F.$$

Capacitance C_S is connected in parallel with the third capacitance C_3, so we use Equation 7.3.9 find the equivalent capacitance **C** of the entire network:

$$\begin{aligned} C &= C_S + C_3 \\ &= 0.833 \mu F + 8.000 \mu F \\ &= 8.833 \mu F. \end{aligned}$$

Network of Capacitors

Determine the net capacitance **C** of the capacitor combination shown in Figure 7.3.4 when the capacitances are $C_1 = 12.0 \mu F, C_2 = 2.0 \mu F$, and $C_3 = 4.0 \mu F$. When a 12.0-V potential difference is maintained across the combination, find the charge and the voltage across each capacitor.

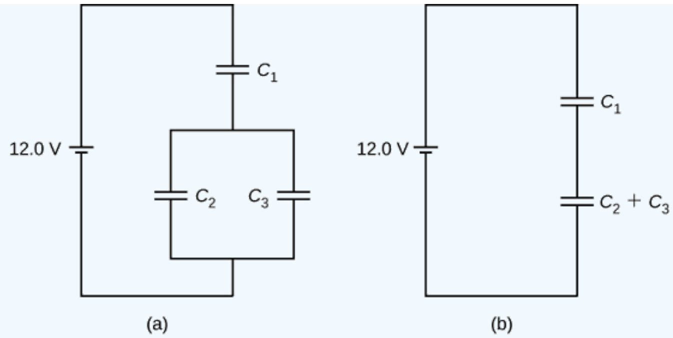


Figure 7.3.4: (a) A capacitor combination. (b) An equivalent two-capacitor combination.

Strategy We first compute the net capacitance C_{23} of the parallel connection C_2 and C_3. Then **C** is the net capacitance of the series connection C_1 and C_{23}. We use the relation $C = Q/V$ to find the charges Q_1, Q_2, and Q_3, and the voltages V_1, V_2, and V_3 across capacitors 1, 2, and 3, respectively.

Solution The equivalent capacitance for C_2 and C_3 is

$$C_{23} = C_2 + C_3 = 2.0\mu F + 4.0\mu F = 6.0\mu F. \tag{7.3.10}$$

The entire three-capacitor combination is equivalent to two capacitors in series,

$$\frac{1}{C} = \frac{1}{12.0\mu F} + \frac{1}{6.0\mu F} = \frac{1}{4.0\mu F} \Rightarrow C = 4.0\mu F. \tag{7.3.11}$$

Consider the equivalent two-capacitor combination in Figure $7.3.2b$. Since the capacitors are in series, they have the same charge, $Q_1 = Q_{23}$. Also, the capacitors share the 12.0-V potential difference, so

$$12.0V = V_1 + V_{23} = \frac{Q_1}{C_1} + \frac{Q_{23}}{C_{23}} = \frac{Q_1}{12.0\mu F} + \frac{Q_1}{6.0\mu F} \Rightarrow Q_1 = 48.0\mu C. \tag{7.3.12}$$

Now the potential difference across capacitor 1 is

$$V_1 = \frac{Q_1}{C_1} = \frac{48.0\mu C}{12.0\mu F} = 4.0V. \tag{7.3.13}$$

Because capacitors 2 and 3 are connected in parallel, they are at the same potential difference:

$$V_2 = V_3 = 12.0V - 4.0V = 8.0V. \tag{7.3.14}$$

Hence, the charges on these two capacitors are, respectively,

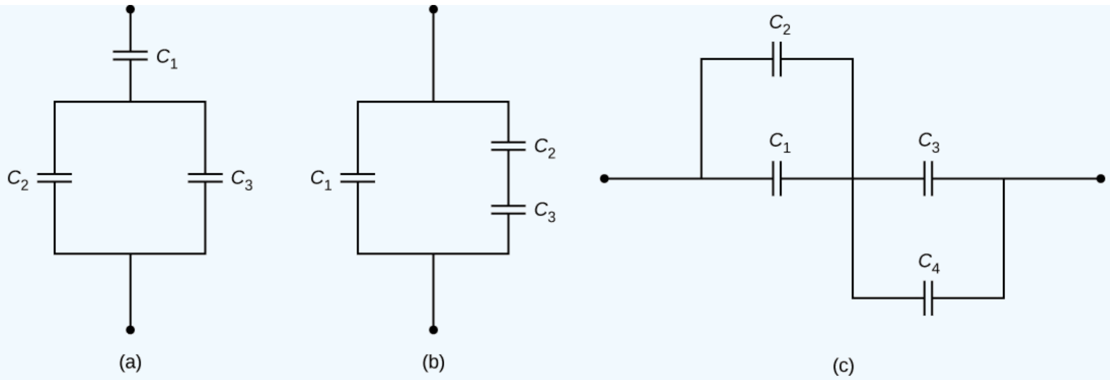
$$Q_2 = C_2V_2 = (2.0\mu F)(8.0V) = 16.0\mu C, \tag{7.3.15}$$

$$Q_3 = C_3V_3 = (4.0\mu F)(8.0V) = 32.0\mu C. \tag{7.3.16}$$

Significance As expected, the net charge on the parallel combination of C_2 and C_3 is $Q_{23} = Q_2 + Q_3 = 48.0\mu C$.

? Exercise 7.3.1

Determine the net capacitance **C** of each network of capacitors shown below. Assume that $C_1 = 1.0pF, C_2 = 2.0pF, C_3 = 4.0pF$, and $C_4 = 5.0pF$. Find the charge on each capacitor, assuming there is a potential difference of 12.0 V across each network.



Answer a

$$C = 0.86pF, Q_1 = 10pC, Q_2 = 3.4pC, Q_3 = 6.8pC$$

Answer b

$$C = 2.3pF, Q_1 = 12pC, Q_2 = Q_3 = 16pC$$

Answer c

$$C = 2.3pF, Q_1 = 9.0pC, Q_2 = 18pC, Q_3 = 12pC, Q_4 = 15pC$$

Circuits often contain both capacitors and resistors. Table $7.3.1$ summarizes the equations used for the equivalent resistance and equivalent capacitance for series and parallel connections.

Table $7.3.1$: Summary for Equivalent Resistance and Capacitance in Series and Parallel Combinations

	Series combination	Parallel combination
Equivalent capacitance	$\frac{1}{C_S} = \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \ldots$	$C_P = C_1 + C_2 + C_3 + \ldots$
Equivalent resistance	$R_S = R_1 + R_2 + R_3 + \ldots = \sum_{i=1}^{N} R_i$	$\frac{1}{R_P} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \ldots$

7.4: Electrical Energy Stored in a Capacitor

Learning Objectives

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

- Explain how energy is stored in a capacitor
- Use energy relations to determine the energy stored in a capacitor network

Most of us have seen dramatizations of medical personnel using a defibrillator to pass an electrical current through a patient's heart to get it to beat normally. Often realistic in detail, the person applying the shock directs another person to "make it 400 joules this time." The energy delivered by the defibrillator is stored in a capacitor and can be adjusted to fit the situation. SI units of joules are often employed. Less dramatic is the use of capacitors in microelectronics to supply energy when batteries are charged (Figure $7.4.1$). Capacitors are also used to supply energy for flash lamps on cameras.

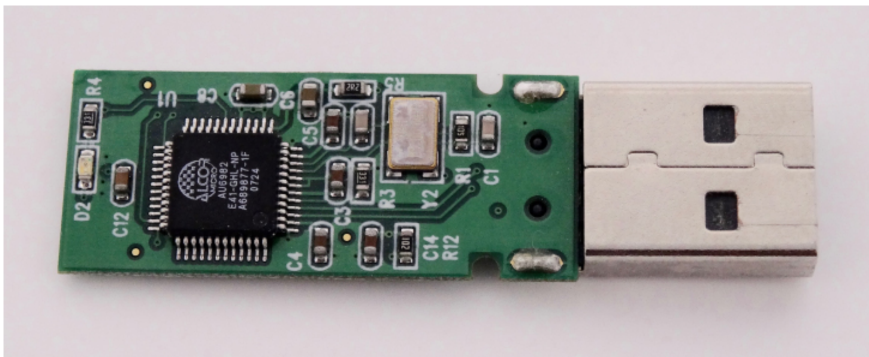


Figure $7.4.1$: The capacitors on the circuit board for an electronic device follow a labeling convention that identifies each one with a code that begins with the letter "C."

The energy U_C stored in a capacitor is electrostatic potential energy and is thus related to the charge **Q** and voltage **V** between the capacitor plates. A charged capacitor stores energy in the electrical field between its plates. As the capacitor is being charged, the electrical field builds up. When a charged capacitor is disconnected from a battery, its energy remains in the field in the space between its plates.

To gain insight into how this energy may be expressed (in terms of **Q** and **V**), consider a charged, empty, parallel-plate capacitor; that is, a capacitor without a dielectric but with a vacuum between its plates. The space between its plates has a volume **Ad**, and it is filled with a uniform electrostatic field **E**. The total energy U_C of the capacitor is contained within this space. The **energy density** u_E in this space is simply U_C divided by the volume **Ad**. If we know the energy density, the energy can be found as $U_C = u_E(Ad)$. We will learn in Electromagnetic Waves (after completing the study of Maxwell's equations) that the energy density u_E in a region of free space occupied by an electrical field **E** depends only on the magnitude of the field and is

$$u_E = \frac{1}{2}\epsilon_0 E^2.$$

If we multiply the energy density by the volume between the plates, we obtain the amount of energy stored between the plates of a parallel-plate capacitor $U_C = u_E(Ad) = \frac{1}{2}\epsilon_0 E^2 Ad = \frac{1}{2}\epsilon_0 \frac{V^2}{d^2} Ad = \frac{1}{2}V^2\epsilon_0 \frac{A}{d} = \frac{1}{2}V^2 C$.

In this derivation, we used the fact that the electrical field between the plates is uniform so that $E = V/d$ and $C = \epsilon_0 A/d$. Because $C = Q/V$, we can express this result in other equivalent forms:

$$U_C = \frac{1}{2}V^2 C = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}QV. \tag{7.4.1}$$

The expression in Equation 7.4.1 for the energy stored in a parallel-plate capacitor is generally valid for all types of capacitors. To see this, consider any uncharged capacitor (not necessarily a parallel-plate type). At some instant, we connect it across a battery, giving it a potential difference $V = q/C$ between its plates. Initially, the charge on the plates is $Q = 0$. As the capacitor is being charged, the charge gradually builds up on its plates, and after some time, it reaches the value **Q**. To move an infinitesimal charge

dq from the negative plate to the positive plate (from a lower to a higher potential), the amount of work **dW** that must be done on **dq** is $dW = W\,dq = \frac{q}{C}dq$.

This work becomes the energy stored in the electrical field of the capacitor. In order to charge the capacitor to a charge **Q**, the total work required is

$$W = \int_0^{W(Q)} dW = \int_0^Q \frac{q}{C}dq = \frac{1}{2}\frac{Q^2}{C}.$$

Since the geometry of the capacitor has not been specified, this equation holds for any type of capacitor. The total work **W** needed to charge a capacitor is the electrical potential energy U_C stored in it, or $U_C = W$. When the charge is expressed in coulombs, potential is expressed in volts, and the capacitance is expressed in farads, this relation gives the energy in joules.

Knowing that the energy stored in a capacitor is $U_C = Q^2/(2C)$, we can now find the energy density u_E stored in a vacuum between the plates of a charged parallel-plate capacitor. We just have to divide U_C by the volume **Ad** of space between its plates and take into account that for a parallel-plate capacitor, we have $E = \sigma/\epsilon_0$ and $C = \epsilon_0 A/d$. Therefore, we obtain

$$u_E = \frac{U_C}{Ad} = \frac{1}{2}\frac{Q^2}{C}\frac{1}{Ad} = \frac{1}{2}\frac{Q^2}{\epsilon_0 A/d}\frac{1}{Ad} = \frac{1}{2}\frac{1}{\epsilon_0}\left(\frac{Q}{A}\right)^2 = \frac{\sigma^2}{2\epsilon_0} = \frac{(E\epsilon_0)^2}{2\epsilon_0} = \frac{\epsilon_0}{2}E^2$$

We see that this expression for the density of energy stored in a parallel-plate capacitor is in accordance with the general relation expressed in Equation ???. We could repeat this calculation for either a spherical capacitor or a cylindrical capacitor—or other capacitors—and in all cases, we would end up with the general relation given by Equation ???.

Energy Stored in a Capacitor

Calculate the energy stored in the capacitor network in Figure 8.3.4a when the capacitors are fully charged and when the capacitances are $C_1 = 12.0\,\mu F$, $C_2 = 2.0\,\mu F$, and $C_3 = 4.0\,\mu F$, respectively.

Strategy

We use Equation 7.4.1 to find the energy U_1, U_2, and U_3 stored in capacitors 1, 2, and 3, respectively. The total energy is the sum of all these energies.

Solution We identify $C_1 = 12.0\,\mu F$ and $V_1 = 4.0\,V$, $C_2 = 2.0\,\mu F$ and $V_2 = 8.0\,V$, $C_3 = 4.0\,\mu F$ and $V_3 = 8.0\,V$. The energies stored in these capacitors are

$$U_1 = \frac{1}{2}C_1V_1^2 = \frac{1}{2}(12.0\,\mu F)(4.0\,V)^2 = 96\,\mu J,$$

$$U_2 = \frac{1}{2}C_2V_2^2 = \frac{1}{2}(2.0\,\mu F)(8.0\,V)^2 = 64\,\mu J,$$

$$U_3 = \frac{1}{2}C_3V_3^2 = \frac{1}{2}(4.0\,\mu F)(8.0\,V)^2 = 130\,\mu J,$$

The total energy stored in this network is

$$U_C = U_1 + U_2 + U_3 = 96\,\mu J + 64\,\mu J + 130\,\mu J = 0.29\,mJ.$$

Significance

We can verify this result by calculating the energy stored in the single $4.0 - \mu F$ capacitor, which is found to be equivalent to the entire network. The voltage across the network is 12.0 V. The total energy obtained in this way agrees with our previously obtained result, $U_C = \frac{1}{2}CV^2 = \frac{1}{2}(4.0\,\mu F)(12.0\,V)^2 = 0.29\,mJ$

Exercise 7.4.1

The potential difference across a 5.0-pF capacitor is 0.40 V. (a) What is the energy stored in this capacitor? (b) The potential difference is now increased to 1.20 V. By what factor is the stored energy increased?

Answer

a. $4.0 \times 10^{-13} J$; b. 9 times

In a cardiac emergency, a portable electronic device known as an automated external defibrillator (AED) can be a lifesaver. A **defibrillator** (Figure $7.4.2$) delivers a large charge in a short burst, or a shock, to a person's heart to correct abnormal heart rhythm (an arrhythmia). A heart attack can arise from the onset of fast, irregular beating of the heart—called cardiac or ventricular fibrillation. Applying a large shock of electrical energy can terminate the arrhythmia and allow the body's natural pacemaker to resume its normal rhythm. Today, it is common for ambulances to carry AEDs. AEDs are also found in many public places. These are designed to be used by lay persons. The device automatically diagnoses the patient's heart rhythm and then applies the shock with appropriate energy and waveform. CPR (cardiopulmonary resuscitation) is recommended in many cases before using a defibrillator.

Figure $7.4.2$: Automated external defibrillators are found in many public places. These portable units provide verbal instructions for use in the important first few minutes for a person suffering a cardiac attack.

Example $7.4.2$: Capacitance of a Heart Defibrillator

A heart defibrillator delivers $4.00 \times 10^2 J$ of energy by discharging a capacitor initially at $1.00 \times 10^4 V$. What is its capacitance?

Strategy

We are given U_C and **V**, and we are asked to find the capacitance **C**. We solve Equation 7.4.1 for **C** and substitute.

Solution

Solving this expression for **C** and entering the given values yields $C = 2\frac{U_C}{V^2} = 2\frac{4.00 \times 10^2 J}{(1.00 \times 10^4 V)^2} = 8.00\, \mu F$.

7.5: Capacitor with a Dielectric

Learning Objectives

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

- Describe the effects a dielectric in a capacitor has on capacitance and other properties
- Calculate the capacitance of a capacitor containing a dielectric

As we discussed earlier, an insulating material placed between the plates of a capacitor is called a dielectric. Inserting a dielectric between the plates of a capacitor affects its capacitance. To see why, let's consider an experiment described in Figure 7.5.1. Initially, a capacitor with capacitance C_0 when there is air between its plates is charged by a battery to voltage V_0. When the capacitor is fully charged, the battery is disconnected. A charge Q_0 then resides on the plates, and the potential difference between the plates is measured to be V_0. Now, suppose we insert a dielectric that **totally** fills the gap between the plates. If we monitor the voltage, we find that the voltmeter reading has dropped to a **smaller** value V. We write this new voltage value as a fraction of the original voltage V_0, with a positive number κ, $\kappa > 1$.

$$V = \frac{1}{\kappa} V_0.$$

The constant κ in this equation is called the **dielectric constant** of the material between the plates, and its value is characteristic for the material. A detailed explanation for why the dielectric reduces the voltage is given in the next section. Different materials have different dielectric constants (a table of values for typical materials is provided in the next section). Once the battery becomes disconnected, there is no path for a charge to flow to the battery from the capacitor plates. Hence, the insertion of the dielectric has no effect on the charge on the plate, which remains at a value of Q_0. Therefore, we find that the capacitance of the capacitor with a dielectric is

$$C = \frac{Q_0}{V} = \frac{Q_0}{V_0/\kappa} = \kappa \frac{Q_0}{V_0} = \kappa C_0. \tag{7.5.1}$$

This equation tells us that the **capacitance C_0 of an empty (vacuum) capacitor can be increased by a factor of κ when we insert a dielectric material to completely fill the space between its plates**. Note that Equation 7.5.1 can also be used for an empty capacitor by setting $\kappa = 1$. In other words, we can say that the dielectric constant of the vacuum is 1, which is a reference value.

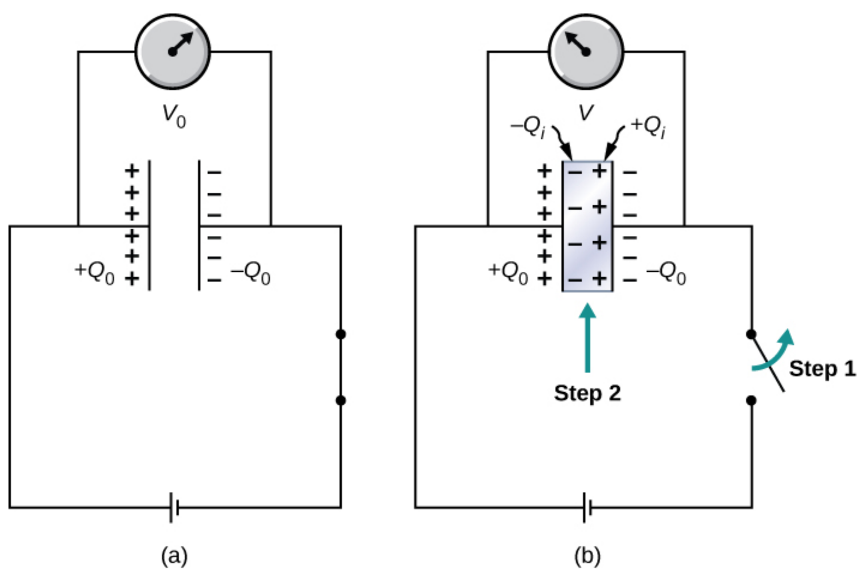


Figure 7.5.1: (a) When fully charged, a vacuum capacitor has a voltage V_0 and charge Q_0 (the charges remain on plate's inner surfaces; the schematic indicates the sign of charge on each plate). (b) In step 1, the battery is disconnected. Then, in step 2, a dielectric (that is electrically neutral) is inserted into the charged capacitor. When the voltage across the capacitor is now measured, it is found that the voltage value has decreased to $V = V_0/\kappa$. The schematic indicates the sign of the induced charge that is now present on the surfaces of the dielectric material between the plates.

The principle expressed by Equation 7.5.1 is widely used in the construction industry (Figure $7.5.2$). Metal plates in an electronic stud finder act effectively as a capacitor. You place a stud finder with its flat side on the wall and move it continually in the horizontal direction. When the finder moves over a wooden stud, the capacitance of its plates changes, because wood has a different dielectric constant than a gypsum wall. This change triggers a signal in a circuit, and thus the stud is detected.

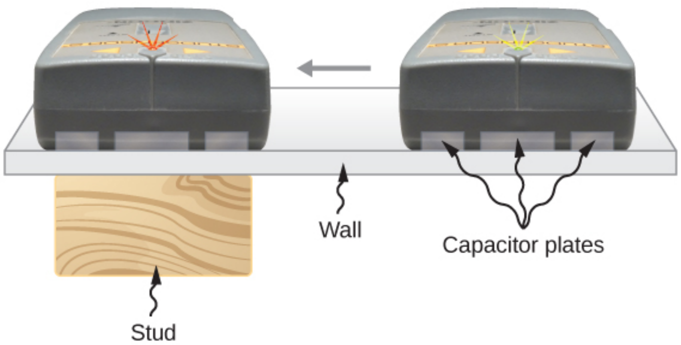


Figure $7.5.2$: An electronic stud finder is used to detect wooden studs behind drywall.

The electrical energy stored by a capacitor is also affected by the presence of a dielectric. When the energy stored in an empty capacitor is U_0, the energy U stored in a capacitor with a dielectric is smaller by a factor of κ.

$$U = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}\frac{Q_0^2}{\kappa C_0} = \frac{1}{\kappa}U_0. \tag{7.5.2}$$

As a dielectric material sample is brought near an empty charged capacitor, the sample reacts to the electrical field of the charges on the capacitor plates. Just as we learned in Electric Charges and Fields on electrostatics, there will be the induced charges on the surface of the sample; however, they are not free charges like in a conductor, because a perfect insulator does not have freely moving charges. These induced charges on the dielectric surface are of an opposite sign to the free charges on the plates of the capacitor, and so they are attracted by the free charges on the plates. Consequently, the dielectric is "pulled" into the gap, and the work to polarize the dielectric material between the plates is done at the expense of the stored electrical energy, which is reduced, in accordance with Equation 7.5.2.

✓ Example 7.5.1: Inserting a Dielectric into an Isolated Capacitor

An empty 20.0-pF capacitor is charged to a potential difference of 40.0 V. The charging battery is then disconnected, and a piece of Teflon™ with a dielectric constant of 2.1 is inserted to completely fill the space between the capacitor plates (see Figure $7.5.1$). What are the values of:

a. the capacitance,
b. the charge of the plate,
c. the potential difference between the plates, and
d. the energy stored in the capacitor with and without dielectric?

Strategy

We identify the original capacitance $C_0 = 20.0\,pF$ and the original potential difference $V_0 = 40.0\,V$ between the plates. We combine Equation 7.5.1 with other relations involving capacitance and substitute.

Solution

a. The capacitance increases to

$$C = \kappa C_0 = 2.1(20.0\,pF) = 42.0\,pF.$$

b. Without dielectric, the charge on the plates is

$$Q_0 = C_0 V_0 = (20.0\,pF)(40.0\,V) = 0.8\,nC.$$

Since the battery is disconnected before the dielectric is inserted, the plate charge is unaffected by the dielectric and remains at 0.8 nC.

c. With the dielectric, the potential difference becomes

$$V = \frac{1}{\kappa} V_0 = \frac{1}{2.1} 40.0\,V = 19.0\,V.$$

d. The stored energy without the dielectric is

$$U_0 = \frac{1}{2} C_0 V_0^2 = \frac{1}{2}(20.0\,pF)(40.0\,V)^2 = 16.0\,nJ.$$

With the dielectric inserted, we use Equation 7.5.2 to find that the stored energy decreases to

$$U = \frac{1}{\kappa} U_0 = \frac{1}{2.1} 16.0\,nJ = 7.6\,nJ.$$

Significance

Notice that the effect of a dielectric on the capacitance of a capacitor is a drastic increase of its capacitance. This effect is far more profound than a mere change in the geometry of a capacitor.

? Exercise 7.5.1

When a dielectric is inserted into an isolated and charged capacitor, the stored energy decreases to 33% of its original value.

a. What is the dielectric constant?
b. How does the capacitance change?

Answer

a. 3.0; b. $C = 3.0\,C_0$

7.6: Molecular Model of a Dielectric

Learning Objectives

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

- Explain the polarization of a dielectric in a uniform electrical field
- Describe the effect of a polarized dielectric on the electrical field between capacitor plates
- Explain dielectric breakdown

We can understand the effect of a dielectric on capacitance by looking at its behavior at the molecular level. As we have seen in earlier chapters, in general, all molecules can be classified as either **polar** or **nonpolar**. There is a net separation of positive and negative charges in an isolated polar molecule, whereas there is no charge separation in an isolated nonpolar molecule (Figure $7.6.1$). In other words, polar molecules have permanent **electric-dipole moments** and nonpolar molecules do not. For example, a molecule of water is polar, and a molecule of oxygen is nonpolar. Nonpolar molecules can become polar in the presence of an external electrical field, which is called **induced polarization**.

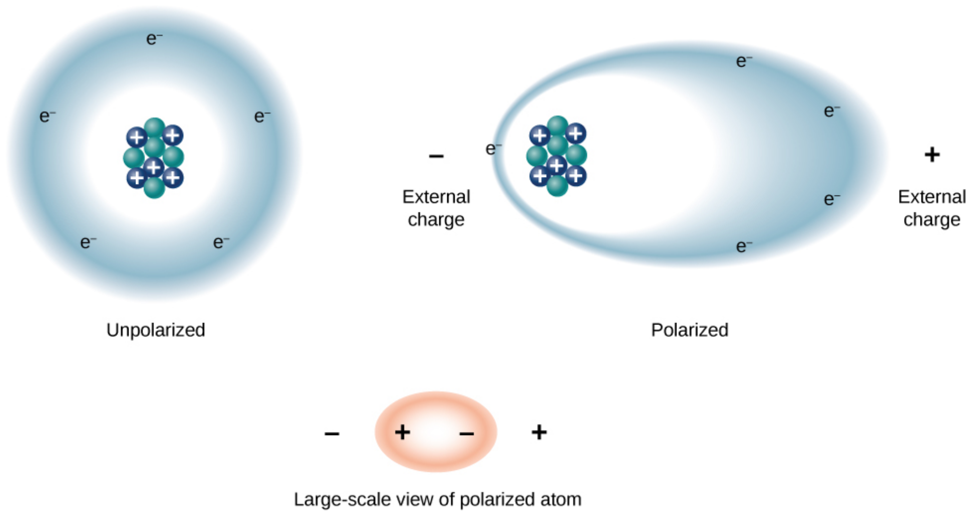


Figure $7.6.1$: The concept of polarization: In an unpolarized atom or molecule, a negatively charged electron cloud is evenly distributed around positively charged centers, whereas a polarized atom or molecule has an excess of negative charge at one side so that the other side has an excess of positive charge. However, the entire system remains electrically neutral. The charge polarization may be caused by an external electrical field. Some molecules and atoms are permanently polarized (electric dipoles) even in the absence of an external electrical field (polar molecules and atoms).

Let's first consider a dielectric composed of polar molecules. In the absence of any external electrical field, the electric dipoles are oriented randomly, as illustrated in Figure $7.6.2a$. However, if the dielectric is placed in an external electrical field $\vec{E}_0$, the polar molecules align with the external field, as shown in $7.6.2b$ of the figure. Opposite charges on adjacent dipoles within the volume of dielectric neutralize each other, so there is no net charge within the dielectric (see the dashed circles in part (b)). However, this is not the case very close to the upper and lower surfaces that border the dielectric (the region enclosed by the dashed rectangles in $7.6.2b$, where the alignment does produce a net charge. Since the external electrical field merely aligns the dipoles, the dielectric as a whole is neutral, and the surface charges induced on its opposite faces are equal and opposite. These **induced surface charges** $+Q_i$ and $-Q_i$ produce an additional electrical field $\vec{E}_i$ (an **induced electrical field**), which **opposes** the external field $\vec{E}_0$, as illustrated in part (c).

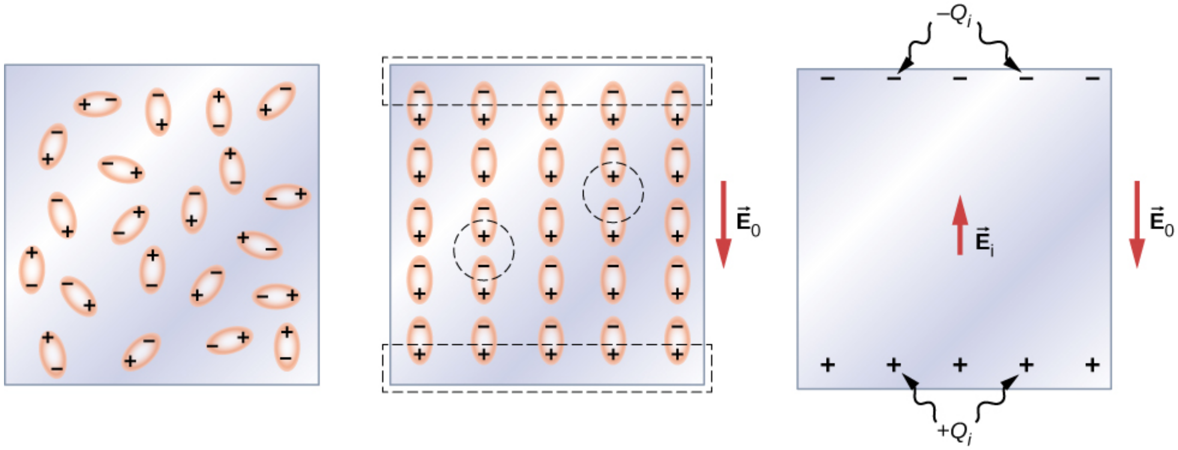


Figure 7.6.2: A dielectric with polar molecules: (a) In the absence of an external electrical field; (b) in the presence of an external electrical field $\vec{E}_0$. The dashed lines indicate the regions immediately adjacent to the capacitor plates. (c) The induced electrical field $\vec{E}_i$ inside the dielectric produced by the induced surface charge Q_i of the dielectric. Note that, in reality, the individual molecules are not perfectly aligned with an external field because of thermal fluctuations; however, the **average** alignment is along the field lines as shown.

The same effect is produced when the molecules of a dielectric are nonpolar. In this case, a nonpolar molecule acquires an induced **electric-dipole moment** because the external field $\vec{E}_0$ causes a separation between its positive and negative charges. The induced dipoles of the nonpolar molecules align with $\vec{E}_0$ in the same way as the permanent dipoles of the polar molecules are aligned (shown in part (b)). Hence, the electrical field within the dielectric is weakened regardless of whether its molecules are polar or nonpolar.

Therefore, when the region between the parallel plates of a charged capacitor, such as that shown in Figure $7.6.3a$, is filled with a dielectric, within the dielectric there is an electrical field $\vec{E}_0$ due to the **free charge** $\boldsymbol{Q_0}$ on the capacitor plates and an electrical field $\vec{E}_i$ due to the induced charge Q_i on the surfaces of the dielectric. Their vector sum gives the net electrical field $\vec{E}$ within the dielectric between the capacitor plates (shown in part (b) of the figure):

$$\vec{E}=\vec{E}_0+\vec{E}_i. \tag{7.6.1}$$

This net field can be considered to be the field produced by an **effective charge** $\boldsymbol{Q_0-Q_i}$ on the capacitor.

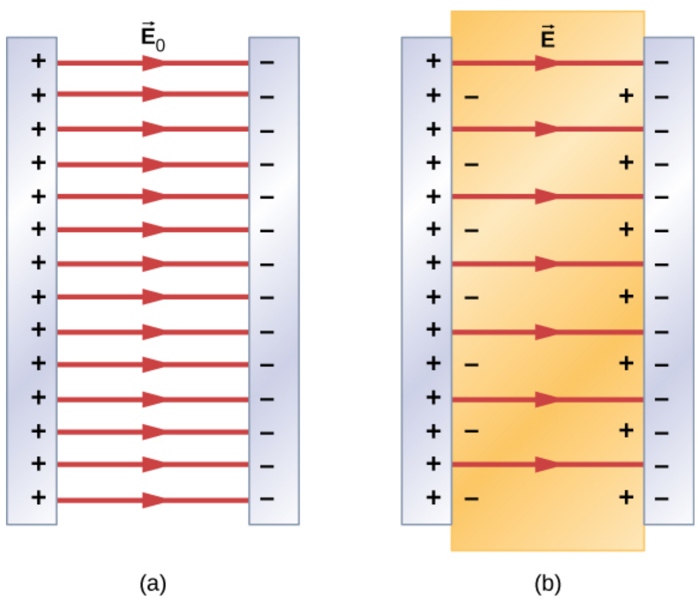


Figure 7.6.3: Electrical field: (a) In an empty capacitor, electrical field $\vec{E}_0$ (b) In a dielectric-filled capacitor, electrical field $\vec{E}$.

In most dielectrics, the net electrical field $\vec{E}$ is proportional to the field $\vec{E}_0$ produced by the free charge. In terms of these two electrical fields, the dielectric constant κ of the material is defined as

$$\kappa = \frac{E_0}{E}. \tag{7.6.2}$$

Since $\vec{E}_0$ and $\vec{E}_i$ point in opposite directions, the magnitude **E** is smaller than the magnitude $\vec{E}_0$ and therefore $\kappa > 1$. Combining Equations 7.6.1 with 7.6.2, and rearranging the terms, yields the following expression for the induced electrical field in a dielectric:

$$\vec{E}_i = \left(\frac{1}{\kappa} - 1\right)\vec{E}_0. \tag{7.6.3}$$

When the magnitude of an external electrical field becomes too large, the molecules of dielectric material start to become ionized. A molecule or an atom is ionized when one or more electrons are removed from it and become free electrons, no longer bound to the molecular or atomic structure. When this happens, the material can conduct, thereby allowing charge to move through the dielectric from one capacitor plate to the other. This phenomenon is called **dielectric breakdown**. (Figure 8.1.1 shows typical random-path patterns of electrical discharge during dielectric breakdown.) The critical value, E_c of the electrical field at which the molecules of an insulator become ionized is called the **dielectric strength** of the material. The dielectric strength imposes a limit on the voltage that can be applied for a given plate separation in a capacitor. For example, the dielectric strength of air is $E_c = 3.0\, MV/m$, so for an air-filled capacitor with a plate separation of $d = 1.00\, mm$, the limit on the potential difference that can be safely applied across its plates without causing dielectric breakdown is $V = E_c d = (3.0 \times 10^6 V/m)(1.00 \times 10^{-3} m) = 3.0\, kV$.

However, this limit becomes 60.0 kV when the same capacitor is filled with Teflon™, whose dielectric strength is about $60.0\, MV/m$. Because of this limit imposed by the dielectric strength, the amount of charge that an air-filled capacitor can store is only $Q_0 = \kappa_{air} C_0 (3.0\, kV)$ and the charge stored on the same Teflon™-filled capacitor can be as much as

$$\begin{aligned} Q &= \kappa_{teflon} C_0 (60.0\, kV) \\ &= \kappa_{teflon} \frac{Q_0}{\kappa_{air}(3.0\, kV)}(60.0\, kV) \\ &= 20 \frac{\kappa_{teflon}}{\kappa_{air}} Q_0 = 20 \frac{2.1}{1.00059} Q_0 \\ &\cong 42 Q_0. \end{aligned} \tag{7.6.4}$$

which is about 42 times greater than a charge stored on an air-filled capacitor. Typical values of dielectric constants and dielectric strengths for various materials are given in Table $7.6.1$. Notice that the dielectric constant κ is exactly 1.0 for a vacuum (the empty space serves as a reference condition) and very close to 1.0 for air under normal conditions (normal pressure at room temperature). These two values are so close that, in fact, the properties of an air-filled capacitor are essentially the same as those of an empty capacitor.

Table 7.6.1: Representative Values of Dielectric Constants and Dielectric Strengths of Various Materials at Room Temperature

Material	Dielectric constant κ	Dielectric strength $E_c[\times 10^6\ V/m]$
Vacuum	1	∞
Dry air (1 atm)	1.00059	3.0
Teflon™	2.1	60 to 173
Paraffin	2.3	11
Silicon oil	2.5	10 to 15
Polystyrene	2.56	19.7
Nylon	3.4	14
Paper	3.7	16

Material	Dielectric constant κ	Dielectric strength $E_c[\times 10^6\ V/m]$
Fused quartz	3.78	8
Glass	4 to 6	9.8 to 13.8
Concrete	4.5	–
Bakelite	4.9	24
Diamond	5.5	2,000
Pyrex glass	5.6	14
Mica	6.0	118
Neoprene rubber	6.7	15.7 to 26.7
Water	80	-
Sulfuric acid	84 to 100	-
Titanium dioxide	86 to 173	–
Strontium titanate	310	8
Barium titanate	1,200 to 10,000	–
Calcium copper titanate	> 250,000	–

Not all substances listed in the table are good insulators, despite their high dielectric constants. Water, for example, consists of polar molecules and has a large dielectric constant of about 80. In a water molecule, electrons are more likely found around the oxygen nucleus than around the hydrogen nuclei. This makes the oxygen end of the molecule slightly negative and leaves the hydrogens end slightly positive, which makes the molecule easy to align along an external electrical field, and thus water has a large dielectric constant. However, the polar nature of water molecules also makes water a good solvent for many substances, which produces undesirable effects, because any concentration of free ions in water conducts electricity.

✓ Example 7.6.1: Electrical Field and Induced Surface Charge

Suppose that the distance between the plates of the capacitor in Example 8.5.1 is 2.0 mm and the area of each plate is $4.5 \times 10^{-3} m^2$. Determine:

a. the electrical field between the plates before and after the Teflon™ is inserted, and
b. the surface charge induced on the Teflon™ surfaces.

Strategy

In part (a), we know that the voltage across the empty capacitor is $V_0 = 40\ V$, so to find the electrical fields we use the relation $V = Ed$ and Equation 7.6.3. In part (b), knowing the magnitude of the electrical field, we use the expression for the magnitude of electrical field near a charged plate $E = \sigma/\epsilon_0$, where σ is a uniform surface charge density caused by the surface charge. We use the value of free charge $Q_0 = 8.0 \times 10^{-10} C$ obtained in Example 8.5.1.

Solution

1. The electrical field E_0 between the plates of an empty capacitor is

$$E_0 = \frac{V_0}{d} = \frac{40\ V}{2.0 \times 10^{-3} m} = 2.010^4 V/m.$$

The electrical field E with the Teflon™ in place is

$$E = \frac{1}{\kappa} E_0 = \frac{1}{2.1} 2.0 \times 10^4 V/m = 9.5 \times 10^3\ V/m.$$

2. The effective charge on the capacitor is the difference between the free charge Q_0 and the induced charge Q_i. The electrical field in the Teflon™ is caused by this effective charge. Thus

$$\begin{aligned} E &= \frac{1}{\epsilon_0}\sigma \\ &= \frac{1}{\epsilon_0}\frac{Q_0 - Q_i}{A}. \end{aligned}$$

We invert this equation to obtain Q_i, which yields

$$\begin{aligned} Q_i &= Q_0 - \epsilon_0 AE \\ &= 8.0 \times 10^{-10} C - \left(8.85 \times 10^{-12} \frac{C^2}{N \cdot m^2}\right)(4.5 \times 10^{-3} m^2)\left(9.5 \times 10^3 \frac{V}{m}\right) \\ &= 4.2 \times 10^{-10} C = 0.42\, nC. \end{aligned}$$

✓ Example 7.6.2: Inserting a Dielectric into a Capacitor Connected to a Battery

When a battery of voltage V_0 is connected across an empty capacitor of capacitance C_0, the charge on its plates is Q_0, and the electrical field between its plates is E_0. A dielectric of dielectric constant κ is inserted between the plates **while the battery remains in place**, as shown in Figure $7.6.4$.

a. Find the capacitance C, the voltage V across the capacitor, and the electrical field E between the plates after the dielectric is inserted.
b. Obtain an expression for the free charge Q on the plates of the filled capacitor and the induced charge Q_i on the dielectric surface in terms of the original plate charge Q_0.

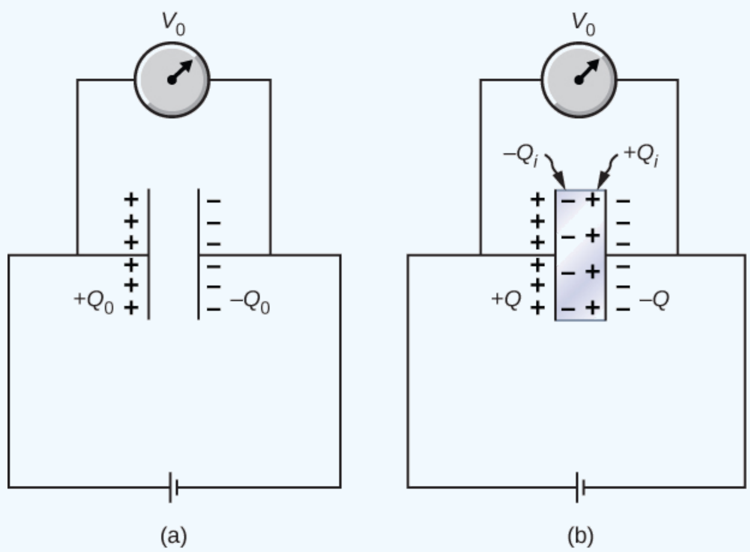


Figure 7.6.4: A dielectric is inserted into the charged capacitor while the capacitor remains connected to the battery.

Strategy

We identify the known values:V_0, C_0, E_o, κ, and Q_0. Our task is to express the unknown values in terms of these known values.

Solution

(a) The capacitance of the filled capacitor is $C = \kappa C_0$. Since the battery is always connected to the capacitor plates, the potential difference between them does not change; hence, $V = V_0$. Because of that, the electrical field in the filled capacitor is the same as the field in the empty capacitor, so we can obtain directly that

$$E = \frac{V}{d} = \frac{V_0}{d} = E_0.$$

(b) For the filled capacitor, the free charge on the plates is

$$\begin{aligned} Q &= CV \\ &= (\kappa C_0)V_0 \\ &= \kappa(C_0V_0) \\ &= \kappa Q_0. \end{aligned}$$

The electrical field E in the filled capacitor is due to the effective charge $Q - Q_i$ (Figure $7.6.4b$). Since $E = E_0$, we have

$$\frac{Q - Q_i}{\epsilon_0 A} = \frac{Q_0}{\epsilon_0 A}.$$

Solving this equation for Q_i, we obtain for the induced charge

$$\begin{aligned} Q_i &= Q - Q_0 \\ &= \kappa Q_0 - Q_0 \\ &= (\kappa - 1)Q_0. \end{aligned}$$

Significance

Notice that for materials with dielectric constants larger than 2 (see Table $7.6.1$), the induced charge on the surface of dielectric is larger than the charge on the plates of a vacuum capacitor. The opposite is true for gasses like air whose dielectric constant is smaller than 2.

? Exercise 7.6.1

Continuing with Example $7.6.2$, show that when the battery is connected across the plates the energy stored in dielectric-filled capacitor is $U = \kappa U_0$ (larger than the energy U_0 of an empty capacitor kept at the same voltage). Compare this result with the result $U = U_0/\kappa$ found previously for an isolated, charged capacitor.

? Exercise 7.6.2

Repeat the calculations of previous example for the case in which the battery remains connected while the dielectric is placed in the capacitor.

Answer

a. $C_0 = 20\,pF$, $C = 42\,pF$;

b. $Q_0 = 0.8\,nC$, $Q = 1.7\,nC$:

c. $V_0 = V = 40\,V$;

d. $U_0 = 16\,nJ$, $U = 34\,nJ$

7.7: Application - RC Circuits

Learning Objectives

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

- Describe the charging process of a capacitor
- Describe the discharging process of a capacitor
- List some applications of RC circuits

When you use a flash camera, it takes a few seconds to charge the capacitor that powers the flash. The light flash discharges the capacitor in a tiny fraction of a second. Why does charging take longer than discharging? This question and several other phenomena that involve charging and discharging capacitors are discussed in this module.

Circuits with Resistance and Capacitance

An **RC circuit** is a circuit containing resistance and capacitance. As presented in Capacitance, the capacitor is an electrical component that stores electric charge, storing energy in an electric field.

Figure $7.7.1a$ shows a simple **RC** circuit that employs a dc (direct current) voltage source ε, a resistor R, a capacitor C, and a two-position switch. The circuit allows the capacitor to be charged or discharged, depending on the position of the switch. When the switch is moved to position \(**A**\), the capacitor charges, resulting in the circuit in Figure $7.7.1b$. When the switch is moved to position **B**, the capacitor discharges through the resistor.

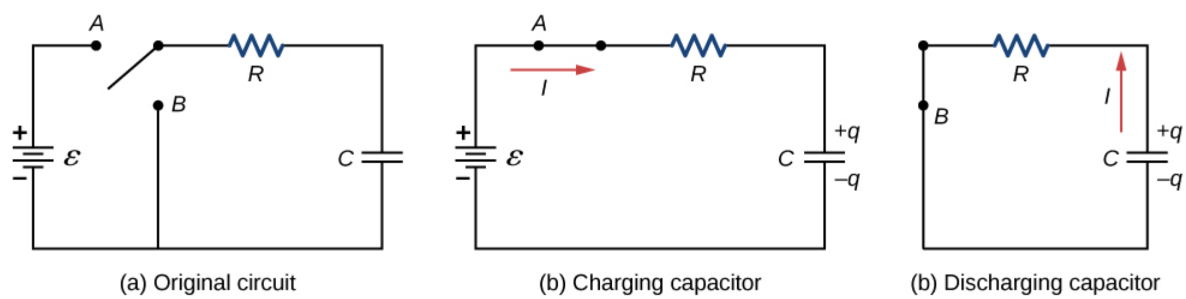


Figure $7.7.1$: (a) An **RC** circuit with a two-pole switch that can be used to charge and discharge a capacitor. (b) When the switch is moved to position **A**, the circuit reduces to a simple series connection of the voltage source, the resistor, the capacitor, and the switch. (c) When the switch is moved to position **B**, the circuit reduces to a simple series connection of the resistor, the capacitor, and the switch. The voltage source is removed from the circuit.

Charging a Capacitor

We can use Kirchhoff's loop rule to understand the charging of the capacitor. This results in the equation $\epsilon - V_R - V_C = 0$. This equation can be used to model the charge as a function of time as the capacitor charges. Capacitance is defined as $C = q/V$, so the voltage across the capacitor is $V_C = \frac{q}{C}$. Using Ohm's law, the potential drop across the resistor is $V_R = IR$, and the current is defined as $I = dq/dt$.

$$\epsilon - V_R - V_C = 0, \tag{7.7.1}$$

$$\epsilon - IR - \frac{q}{C} = 0, \tag{7.7.2}$$

$$\epsilon - R\frac{dq}{dt} - \frac{q}{C} = 0. \tag{7.7.3}$$

This differential equation can be integrated to find an equation for the charge on the capacitor as a function of time.

$$\epsilon - R\frac{dq}{dt} - \frac{q}{C} = 0. \tag{7.7.4}$$

$$\frac{dq}{dt} = \frac{\epsilon C - q}{RC}, \tag{7.7.5}$$

$$\int_0^q \frac{dq}{\epsilon C - q} = \frac{1}{RC}\int_0^t dt. \tag{7.7.6}$$

Let $u = \epsilon C - q$, then $du = -dq$. The result is

$$-\int_0^q \frac{du}{u} = \frac{1}{RC}\int_0^t dt, \tag{7.7.7}$$

$$\ln\left(\frac{\epsilon C - q}{\epsilon C}\right) = -\frac{1}{RC}t. \tag{7.7.8}$$

$$\frac{\epsilon C - q}{\epsilon C} = e^{-t/RC}. \tag{7.7.9}$$

Simplifying results in an equation for the charge on the charging capacitor as a function of time:

$$q(t) = C\epsilon\left(1 - e^{-\frac{t}{RC}}\right) = Q\left(1 - e^{-\frac{t}{\tau}}\right). \tag{7.7.10}$$

A graph of the charge on the capacitor versus time is shown in Figure $7.7.2a$. First note that as time approaches infinity, the exponential goes to zero, so the charge approaches the maximum charge $Q = C\epsilon$ and has units of coulombs. The units of **RC** are seconds, units of time. This quantity is known as the **time constant**:

$$\tau = RC. \tag{7.7.11}$$

At time $t = \tau = RC$, the charge equal to $1 - e^{-1} = 1 - 0.368 = 0.632$ of the maximum charge $Q = C\epsilon$. Notice that the time rate change of the charge is the slope at a point of the charge versus time plot. The slope of the graph is large at time $t - 0.0\ s$ and approaches zero as time increases.

As the charge on the capacitor increases, the current through the resistor decreases, as shown in Figure $7.7.2b$. The current through the resistor can be found by taking the time derivative of the charge.

$$\begin{aligned} I(t) &= \frac{dq}{dt} \\ &= \frac{d}{dt}\left[C\epsilon\left(1 - e^{-\frac{t}{RC}}\right)\right], \\ &= C\epsilon\left(\frac{1}{RC}\right)e^{-\frac{t}{RC}} \\ &= \frac{\epsilon}{R}e^{-\frac{t}{RC}} \\ &= I_0 e^{-\frac{t}{RC}}, \end{aligned}$$

$$I(t) = I_0 e^{-t/\tau}. \tag{7.7.12}$$

At time $t = 0.0\ s$, the current through the resistor is $I_0 = \frac{\epsilon}{R}$. As time approaches infinity, the current approaches zero. At time $t = \tau$, the current through the resistor is $I(t = \tau) = I_0 e^{-1} = 0.368 I_0$.

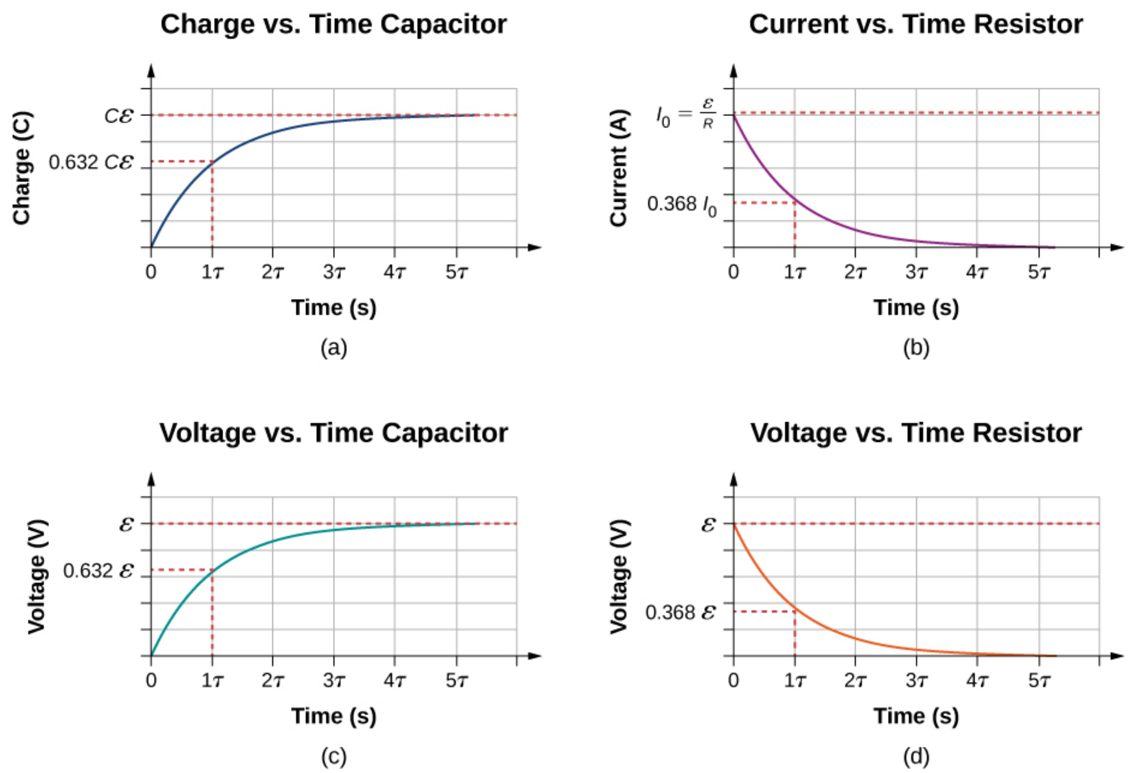


Figure 7.7.2: (a) Charge on the capacitor versus time as the capacitor charges. (b) Current through the resistor versus time. (c) Voltage difference across the capacitor. (d) Voltage difference across the resistor.

Figures $7.7.2c$ and Figure $7.7.2d$ show the voltage differences across the capacitor and the resistor, respectively. As the charge on the capacitor increases, the current decreases, as does the voltage difference across the resistor $V_R(t) = (I_0 R)e^{-t/\tau} = \epsilon e^{-t/\tau}$. The voltage difference across the capacitor increases as $V_C(t) = \epsilon(1 - e^{-t/\tau})$.

Discharging a Capacitor

When the switch in Figure $7.7.3a$ is moved to position **B**, the circuit reduces to the circuit in part (c), and the charged capacitor is allowed to discharge through the resistor. A graph of the charge on the capacitor as a function of time is shown in Figure $7.7.3a$. Using Kirchhoff's loop rule to analyze the circuit as the capacitor discharges results in the equation $-V_R - V_C = 0$, which simplifies to $IR + \frac{q}{C} = 0$. Using the definition of current $\frac{dq}{dt} R = -\frac{q}{C}$ and integrating the loop equation yields an equation for the charge on the capacitor as a function of time:

$$q(t) = Qe^{-t/\tau}. \tag{7.7.13}$$

Here, **Q** is the initial charge on the capacitor and $\tau = RC$ is the time constant of the circuit. As shown in the graph, the charge decreases exponentially from the initial charge, approaching zero as time approaches infinity.

The current as a function of time can be found by taking the time derivative of the charge:

$$I(t) = -\frac{Q}{RC} e^{-t/\tau}. \tag{7.7.14}$$

The negative sign shows that the current flows in the opposite direction of the current found when the capacitor is charging. Figure $7.7.3b$ shows an example of a plot of charge versus time and current versus time. A plot of the voltage difference across the capacitor and the voltage difference across the resistor as a function of time are shown in Figures $7.7.3c$ and $7.7.3d$. Note that the magnitudes of the charge, current, and voltage all decrease exponentially, approaching zero as time increases.

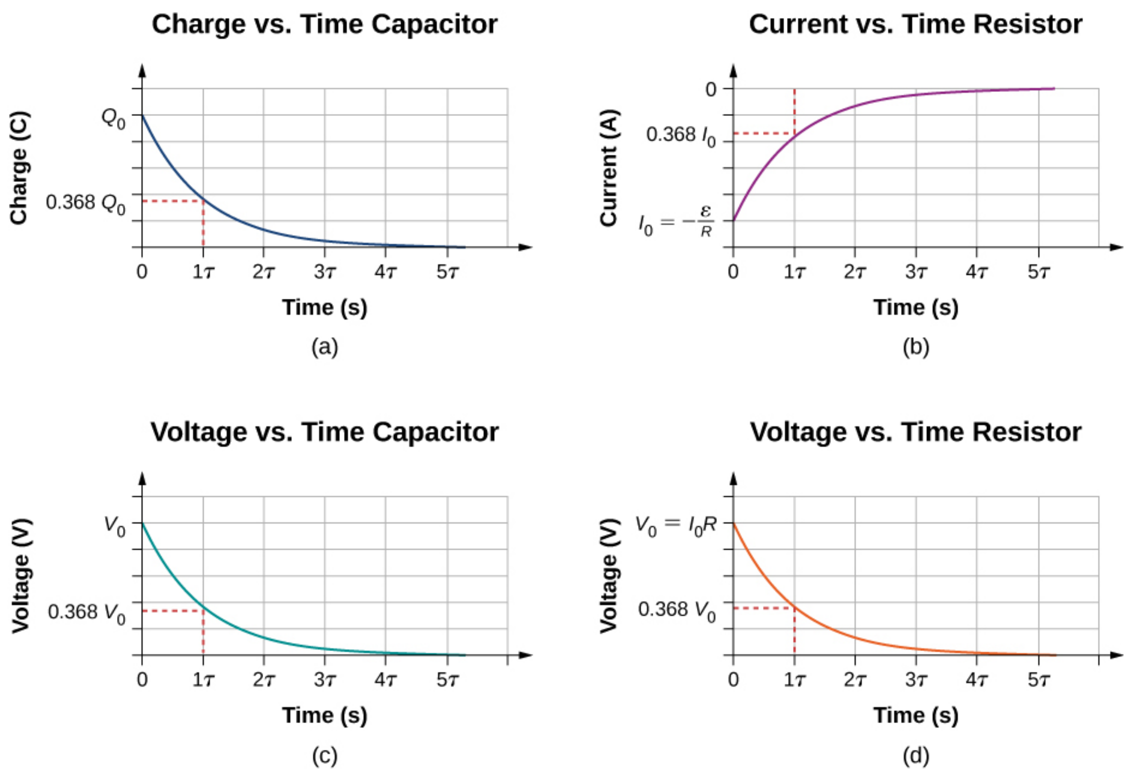
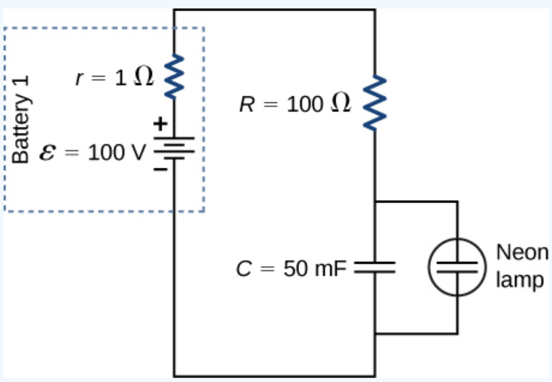


Figure 7.7.3: (a) Charge on the capacitor versus time as the capacitor discharges. (b) Current through the resistor versus time. (c) Voltage difference across the capacitor. (d) Voltage difference across the resistor.

Now we can explain why the **flash camera** mentioned at the beginning of this section takes so much longer to charge than discharge: The resistance while charging is significantly greater than while discharging. The internal resistance of the battery accounts for most of the resistance while charging. As the battery ages, the increasing internal resistance makes the charging process even slower.

Example 7.7.2: The Relaxation Oscillator

One application of an **RC** circuit is the relaxation oscillator, as shown below. The relaxation oscillator consists of a voltage source, a resistor, a capacitor, and a neon lamp. The neon lamp acts like an open circuit (infinite resistance) until the potential difference across the neon lamp reaches a specific voltage. At that voltage, the lamp acts like a short circuit (zero resistance), and the capacitor discharges through the neon lamp and produces light. In the relaxation oscillator shown, the voltage source charges the capacitor until the voltage across the capacitor is 80 V. When this happens, the neon in the lamp breaks down and allows the capacitor to discharge through the lamp, producing a bright flash. After the capacitor fully discharges through the neon lamp, it begins to charge again, and the process repeats. Assuming that the time it takes the capacitor to discharge is negligible, what is the time interval between flashes?



Strategy

The time period can be found from considering the equation $V_C(t) = \epsilon(1 - e^{=t/\tau})$. where $\tau = (R + r)C$.

Solution

The neon lamp flashes when the voltage across the capacitor reaches 80 V. The **RC** time constant is equal to $\tau = (R + r) = (101\,\Omega)(50 \times 10^{-3}\,F) = 5.05\,s$. We can solve the voltage equation for the time it takes the capacitor to reach 80 V:

$$V_C(t) = \epsilon(1 - e^{-t/\tau}), \tag{7.7.15}$$

$$e^{-t/\tau} = 1 - \frac{V_C(t)}{\epsilon}, \tag{7.7.16}$$

$$ln(e^{-t/\tau}) = ln\left(1 - \frac{V_C(t)}{\epsilon}\right), \tag{7.7.17}$$

$$t = -\tau ln\left(1 - \frac{V_C(t)}{\epsilon}\right) = -5.05\,s \cdot ln\left(1 - \frac{80\,V}{100\,V}\right) = 8.13\,s. \tag{7.7.18}$$

Significance

One application of the relaxation oscillator is for controlling indicator lights that flash at a frequency determined by the values for **R** and **C**. In this example, the neon lamp will flash every 8.13 seconds, a frequency of $f = \frac{1}{T} = \frac{1}{8.13\,s} = 0.55\,Hz$. The relaxation oscillator has many other practical uses. It is often used in electronic circuits, where the neon lamp is replaced by a transistor or a device known as a tunnel diode. The description of the transistor and tunnel diode is beyond the scope of this chapter, but you can think of them as voltage controlled switches. They are normally open switches, but when the right voltage is applied, the switch closes and conducts. The “switch” can be used to turn on another circuit, turn on a light, or run a small motor. A relaxation oscillator can be used to make the turn signals of your car blink or your cell phone to vibrate.

RC circuits have many applications. They can be used effectively as timers for applications such as intermittent windshield wipers, pace makers, and strobe lights. Some models of intermittent windshield wipers use a variable resistor to adjust the interval between sweeps of the wiper. Increasing the resistance increases the **RC** time constant, which increases the time between the operation of the wipers.

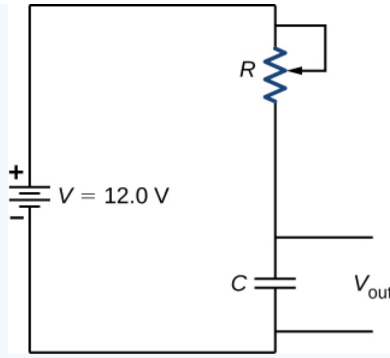
Another application is the **pacemaker**. The heart rate is normally controlled by electrical signals, which cause the muscles of the heart to contract and pump blood. When the heart rhythm is abnormal (the heartbeat is too high or too low), pace makers can be used to correct this abnormality. Pacemakers have sensors that detect body motion and breathing to increase the heart rate during physical activities, thus meeting the increased need for blood and oxygen, and an **RC** timing circuit can be used to control the time between voltage signals to the heart.

Looking ahead to the study of ac circuits (Alternating-Current Circuits), ac voltages vary as sine functions with specific frequencies. Periodic variations in voltage, or electric signals, are often recorded by scientists. These voltage signals could come from music recorded by a microphone or atmospheric data collected by radar. Occasionally, these signals can contain unwanted frequencies known as “noise.” **RC** filters can be used to filter out the unwanted frequencies.

In the study of electronics, a popular device known as a 555 timer provides timed voltage pulses. The time between pulses is controlled by an **RC** circuit. These are just a few of the countless applications of **RC** circuits.

✔ Example 7.7.2: Intermittent Windshield Wipers

A relaxation oscillator is used to control a pair of windshield wipers. The relaxation oscillator consists of a 10.00-mF capacitor and a $10.00\,k\Omega$ variable resistor known as a rheostat. A knob connected to the variable resistor allows the resistance to be adjusted from $0.00\,\Omega$ to $10.00\,k\Omega$. The output of the capacitor is used to control a voltage-controlled switch. The switch is normally open, but when the output voltage reaches 10.00 V, the switch closes, energizing an electric motor and discharging the capacitor. The motor causes the windshield wipers to sweep once across the windshield and the capacitor begins to charge again. To what resistance should the rheostat be adjusted for the period of the wiper blades be 10.00 seconds?



Strategy

The resistance considers the equation $V_{out}(t) = V(1 - e^{-t/\tau})$, where $\tau = RC$. The capacitance, output voltage, and voltage of the battery are given. We need to solve this equation for the resistance.

Solution

The output voltage will be 10.00 V and the voltage of the battery is 12.00 V. The capacitance is given as 10.00 mF. Solving for the resistance yields

$$V_{out}(t) = V(1 - e^{-t/\tau}) \tag{7.7.19}$$

$$e^{-t/RC} = 1 - \frac{V_{out}(t)}{V}, \tag{7.7.20}$$

$$ln(e^{-t/RC}) = ln\left(1 - \frac{V_{out}(t)}{V}\right), \tag{7.7.21}$$

$$-\frac{t}{RC} = ln\left(1 - \frac{V_{out}(t)}{V}\right), \tag{7.7.22}$$

$$R = \frac{-t}{C\,ln\left(1 - \frac{V_C(t)}{V}\right)} = \frac{-10.00\,s}{10\times10^{-3}F\,ln\left(1 - \frac{10\,V}{12\,V}\right)} = 558.11\,\Omega. \tag{7.7.23}$$

Significance

Increasing the resistance increases the time delay between operations of the windshield wipers. When the resistance is zero, the windshield wipers run continuously. At the maximum resistance, the period of the operation of the wipers is:

$$t = -RCln\left(1 - \frac{V_{out}(t)}{V}\right) = -(10\times10^{-3}F)(10\times10^{3}\,\Omega)ln\left(1 - \frac{10\,V}{12\,V}\right) = 179.18\,s = 2.98\,min. \tag{7.7.24}$$

The **RC** circuit has thousands of uses and is a very important circuit to study. Not only can it be used to time circuits, it can also be used to filter out unwanted frequencies in a circuit and used in power supplies, like the one for your computer, to help turn ac voltage to dc voltage.

7.8: Application - RC Circuits with AC

Learning Objectives

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

- Interpret phasor diagrams and apply them to ac circuits with resistors and capacitors
- Define the reactance for a resistor and capacitor 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 and (2) a capacitor. The power furnished by an ac voltage source has an emf given by

$$v(t) = V_0 \, sin \, \omega t, \tag{7.8.1}$$

as shown in Figure 7.8.1. This sine function assumes we start recording the voltage when it is $v = 0\,V$ at a time of $t = 0\,s$. 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.

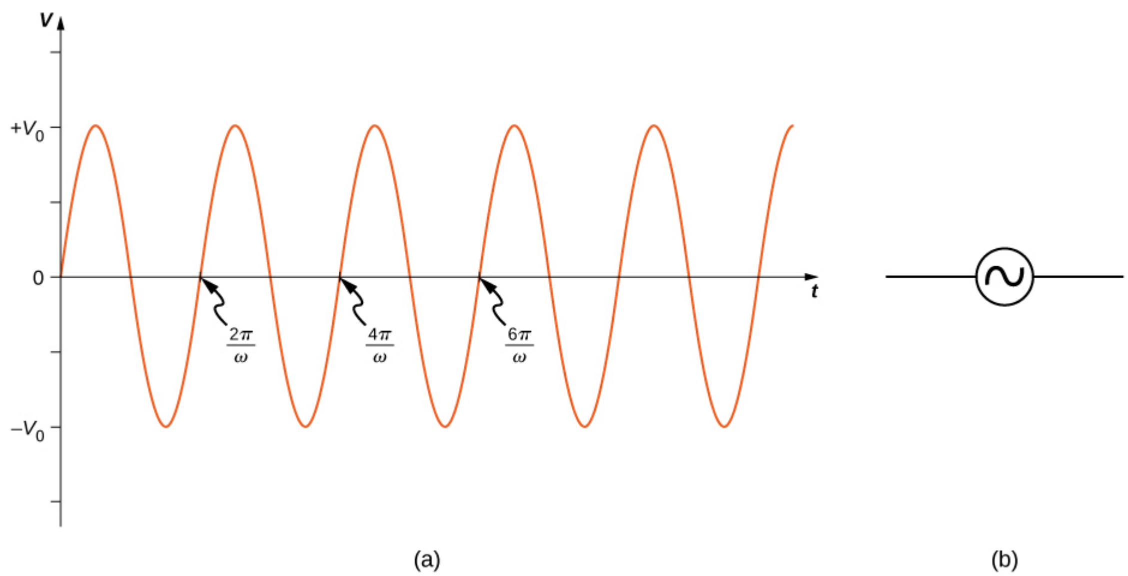


Figure 7.8.1: (a) The output $v(t) = V_0 \, sin \, \omega t$ of an ac generator. (b) Symbol used to represent an ac voltage source in a circuit diagram.

Resistor

First, consider a **resistor** connected across an ac voltage source. From Kirchhoff's loop rule, the instantaneous voltage across the resistor of Figure $7.8.2a$ is

$$v_R(t) = V_0 \, sin \, \omega t \tag{7.8.2}$$

and the instantaneous current through the resistor is

$$i_R(t) = \frac{v_R(t)}{R} = \frac{V_0}{R} \, sin \, \omega t = I_0 \, sin \, \omega t. \tag{7.8.3}$$

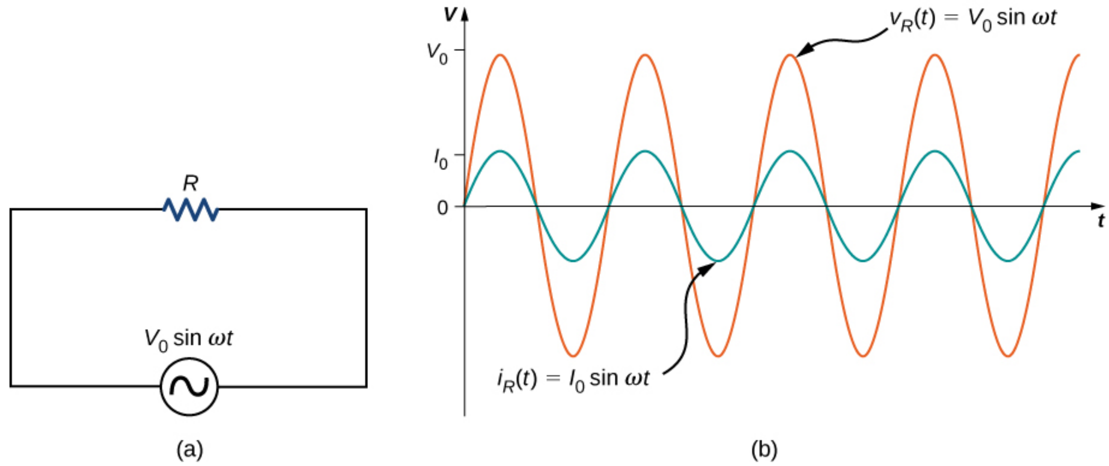


Figure 7.8.2: (a) A resistor connected across an ac voltage source. (b) The current $i_R(t)$ through the resistor and the voltage $v_R(t)$ across the resistor. The two quantities are in phase.

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 $7.8.2b$. 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 $7.8.3a$, 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.

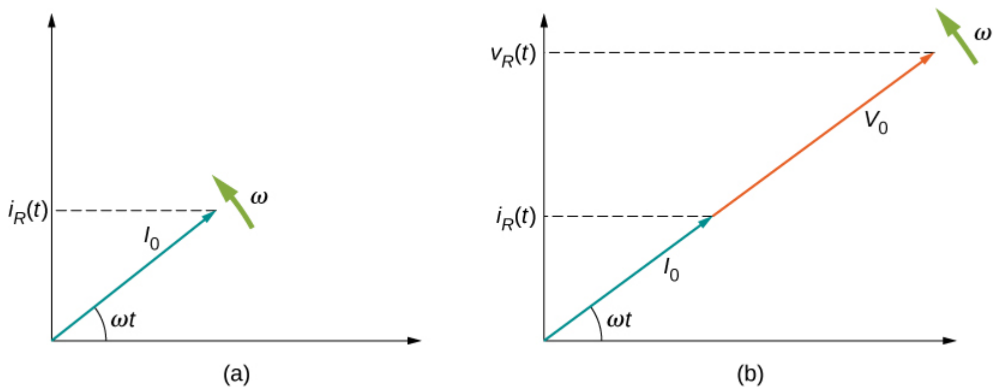


Figure 7.8.3: (a) The phasor diagram representing the current through the resistor of Figure 7.8.2 (b) The phasor diagram representing both $i_R(t)$ and $v_R(t)$.

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 $7.8.3b$. 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 $7.8.4a$ is

$$v_C(t) = V_0 \, sin \, \omega t. \tag{7.8.4}$$

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

$$q(t) = Cv_C(t) = CV_0 \, sin \, \omega t. \tag{7.8.5}$$

Since the current in the circuit is the rate at which charge enters (or leaves) the capacitor,

$$i_C(t) = \frac{dq(t)}{dt} = \omega C V_0 \cos \omega t = I_0 \cos \omega t, \tag{7.8.6}$$

where $I_0 = \omega C V_0$ is the current amplitude. Using the trigonometric relationship $\cos \omega t = \sin(\omega t + \pi/2)$, we may express the instantaneous current as

$$i_C(t) = I_0 \sin\left(\omega t + \frac{\pi}{2}\right). \tag{7.8.7}$$

Dividing V_0 by I_0, we obtain an equation that looks similar to Ohm's law:

$$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C. \tag{7.8.8}$$

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.

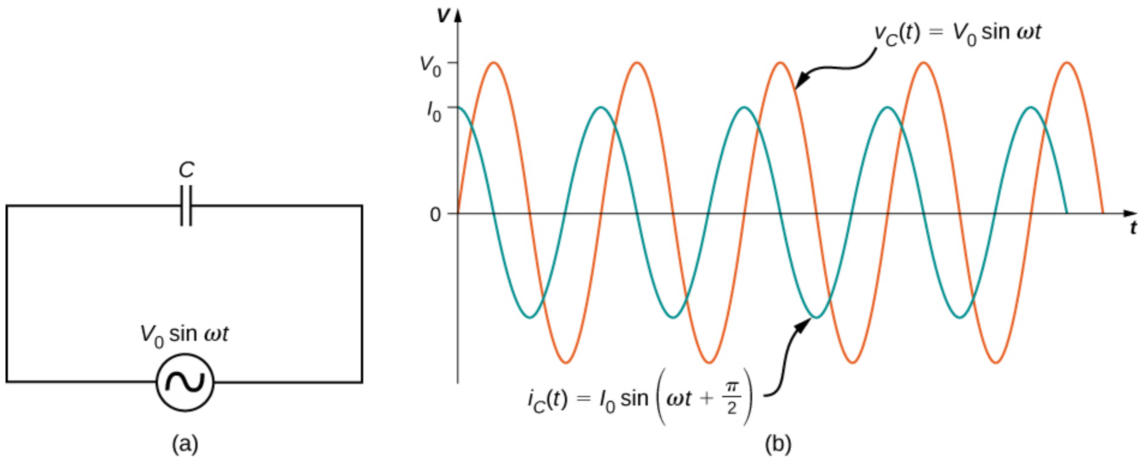


Figure 7.8.4: (a) A capacitor connected across an ac generator. (b) The current $i_C(t)$ through the capacitor and the voltage $v_C(t)$ across the capacitor. Notice that $i_C(t)$ leads $v_C(t)$ by $\pi/2$ rad.

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 7.8.4b. 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 7.8.5. Here, the relationship between $i_C(t)$ and $v_C(t)$ is represented by having their phasors rotate at the same angular frequency, with the current phasor leading by $\pi/2$ rad.

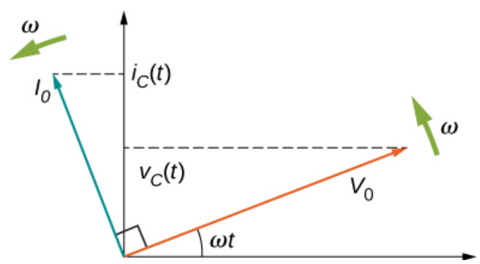


Figure 7.8.5: The current phasor leads the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

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

$$I_{rms} = \frac{I_0}{\sqrt{2}}, \tag{7.8.9}$$

where I_0 is the peak current in an ac system. The **rms voltage**, or the root mean square of the voltage, is

Note

$$V_{rms} = \frac{V_0}{\sqrt{2}}, \tag{7.8.10}$$

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).

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).

7.9: Capacitance (Summary)

Key Terms

capacitance	amount of charge stored per unit volt
capacitor	device that stores electrical charge and electrical energy
dielectric	insulating material used to fill the space between two plates
dielectric breakdown	phenomenon that occurs when an insulator becomes a conductor in a strong electrical field
dielectric constant	factor by which capacitance increases when a dielectric is inserted between the plates of a capacitor
dielectric strength	critical electrical field strength above which molecules in insulator begin to break down and the insulator starts to conduct
energy density	energy stored in a capacitor divided by the volume between the plates
induced electric-dipole moment	dipole moment that a nonpolar molecule may acquire when it is placed in an electrical field
induced electrical field	electrical field in the dielectric due to the presence of induced charges
induced surface charges	charges that occur on a dielectric surface due to its polarization
parallel combination	components in a circuit arranged with one side of each component connected to one side of the circuit and the other sides of the components connected to the other side of the circuit
parallel-plate capacitor	system of two identical parallel conducting plates separated by a distance
RC circuit	circuit that contains both a resistor and a capacitor
series combination	components in a circuit arranged in a row one after the other in a circuit

Key Equations

Capacitance	$C=\frac{Q}{V}$
Capacitance of a parallel-plate capacitor	$C=\varepsilon_0\frac{A}{d}$
Capacitance of a vacuum spherical capacitor	$C=4\pi\varepsilon_0\frac{R_1R_2}{R_2-R_1}$
Capacitance of a vacuum cylindrical capacitor	$C=\frac{2\pi\varepsilon_0 l}{ln(R_2/R_1)}$
Capacitance of a series combination	$\frac{1}{C_S}=\frac{1}{C_1}+\frac{1}{C_2}+\frac{1}{C_3}+\cdots$
Capacitance of a parallel combination	$C_P=C_1+C_2+C_3+\cdots$
Energy density	$u_E=\frac{1}{2}\varepsilon_0E^2$
Energy stored in a capacitor	$U_C=\frac{1}{2}V^2C=\frac{1}{2}\frac{Q^2}{C}=\frac{1}{2}QV$
Capacitance of a capacitor with dielectric	$C=\kappa C_0$

Energy stored in an isolated capacitor with dielectric	$U=\frac{1}{\kappa}U_0$
Dielectric constant	$\kappa=\frac{E_0}{E}$
Induced electrical field in a dielectric	$\vec{E}_i=(\frac{1}{\kappa}-1)\vec{E}_0$
Charge on a charging capacitor	$q(t)=C\varepsilon(1-e^{-\frac{t}{RC}})=Q(1-e^{-\frac{t}{\tau}})$
Time constant	$\tau=RC$
Current during charging of a capacitor	$I=\frac{\varepsilon}{R}e^{-\frac{t}{RC}}=I_oe^{-\frac{t}{RC}}$
Charge on a discharging capacitor	$q(t)=Qe^{-\frac{t}{\tau}}$
Current during discharging of a capacitor	$I(t)=-\frac{Q}{RC}e^{-\frac{t}{\tau}}$

Summary

Capacitors and Capacitance

- A capacitor is a device that stores an electrical charge and electrical energy. The amount of charge a vacuum capacitor can store depends on two major factors: the voltage applied and the capacitor’s physical characteristics, such as its size and geometry.
- The capacitance of a capacitor is a parameter that tells us how much charge can be stored in the capacitor per unit potential difference between its plates. Capacitance of a system of conductors depends only on the geometry of their arrangement and physical properties of the insulating material that fills the space between the conductors. The unit of capacitance is the farad, where $1F=1C/1V$.

Capacitors in Series and in Parallel

- When several capacitors are connected in a series combination, the reciprocal of the equivalent capacitance is the sum of the reciprocals of the individual capacitances.
- When several capacitors are connected in a parallel combination, the equivalent capacitance is the sum of the individual capacitances.
- When a network of capacitors contains a combination of series and parallel connections, we identify the series and parallel networks, and compute their equivalent capacitances step by step until the entire network becomes reduced to one equivalent capacitance.

Energy Stored in a Capacitor

- Capacitors are used to supply energy to a variety of devices, including defibrillators, microelectronics such as calculators, and flash lamps.
- The energy stored in a capacitor is the work required to charge the capacitor, beginning with no charge on its plates. The energy is stored in the electrical field in the space between the capacitor plates. It depends on the amount of electrical charge on the plates and on the potential difference between the plates.
- The energy stored in a capacitor network is the sum of the energies stored on individual capacitors in the network. It can be computed as the energy stored in the equivalent capacitor of the network.

Capacitor with a Dielectric

- The capacitance of an empty capacitor is increased by a factor of κ when the space between its plates is completely filled by a dielectric with dielectric constant κ.
- Each dielectric material has its specific dielectric constant.
- The energy stored in an empty isolated capacitor is decreased by a factor of κκ when the space between its plates is completely filled with a dielectric with dielectric constant κ.

Molecular Model of a Dielectric

- When a dielectric is inserted between the plates of a capacitor, equal and opposite surface charge is induced on the two faces of the dielectric. The induced surface charge produces an induced electrical field that opposes the field of the free charge on the capacitor plates.

- The dielectric constant of a material is the ratio of the electrical field in vacuum to the net electrical field in the material. A capacitor filled with dielectric has a larger capacitance than an empty capacitor.
- The dielectric strength of an insulator represents a critical value of electrical field at which the molecules in an insulating material start to become ionized. When this happens, the material can conduct and dielectric breakdown is observed.

Application - RC Circuits

- An **RC** circuit is one that has both a resistor and a capacitor.
- The time constant τ for an **RC** circuit is $\tau = RC$.
- When an initially uncharged ($q = 0$ at $t = 0$) capacitor in series with a resistor is charged by a dc voltage source, the capacitor asymptotically approaches the maximum charge.
- As the charge on the capacitor increases, the current exponentially decreases from the initial current: $I_0 = \varepsilon / R$.
- If a capacitor with an initial charge **Q** is discharged through a resistor starting at $t = 0$, then its charge decreases exponentially. The current flows in the opposite direction, compared to when it charges, and the magnitude of the charge decreases with time.

Contributors and Attributions

7.10: Capacitance (Exercises)

Conceptual Questions

Capacitors and Capacitance

1. Does the capacitance of a device depend on the applied voltage? Does the capacitance of a device depend on the charge residing on it?

2. Would you place the plates of a parallel-plate capacitor closer together or farther apart to increase their capacitance?

3. The value of the capacitance is zero if the plates are not charged. True or false?

4. If the plates of a capacitor have different areas, will they acquire the same charge when the capacitor is connected across a battery?

5. Does the capacitance of a spherical capacitor depend on which sphere is charged positively or negatively?

Capacitors in Series and in Parallel

6. If you wish to store a large amount of charge in a capacitor bank, would you connect capacitors in series or in parallel? Explain.

7. What is the maximum capacitance you can get by connecting three **1.0-μF** capacitors? What is the minimum capacitance?

Energy Stored in a Capacitor

8. If you wish to store a large amount of energy in a capacitor bank, would you connect capacitors in series or parallel? Explain.

Capacitor with a Dielectric

9. Discuss what would happen if a conducting slab rather than a dielectric were inserted into the gap between the capacitor plates.

10. Discuss how the energy stored in an empty but charged capacitor changes when a dielectric is inserted if (a) the capacitor is isolated so that its charge does not change; (b) the capacitor remains connected to a battery so that the potential difference between its plates does not change.

Molecular Model of a Dielectric

11. Distinguish between dielectric strength and dielectric constant.

12. Water is a good solvent because it has a high dielectric constant. Explain.

13. Water has a high dielectric constant. Explain why it is then not used as a dielectric material in capacitors.

14. Elaborate on why molecules in a dielectric material experience net forces on them in a non-uniform electrical field but not in a uniform field.

15. Explain why the dielectric constant of a substance containing permanent molecular electric dipoles decreases with increasing temperature.

16. Give a reason why a dielectric material increases capacitance compared with what it would be with air between the plates of a capacitor. How does a dielectric material also allow a greater voltage to be applied to a capacitor? (The dielectric thus increases **C** and permits a greater **V**.)

17. Elaborate on the way in which the polar character of water molecules helps to explain water's relatively large dielectric constant.

18. Sparks will occur between the plates of an air-filled capacitor at a lower voltage when the air is humid than when it is dry. Discuss why, considering the polar character of water molecules.

Application - RC Circuits

RC1. When making an ECG measurement, it is important to measure voltage variations over small time intervals. The time is limited by the RC constant of the circuit—it is not possible to measure time variations shorter than RC. How would you manipulate R and C in the circuit to allow the necessary measurements?

RC2. A battery, switch, capacitor, and lamp are connected in series. Describe what happens to the lamp when the switch is closed.

Problems

Capacitors and Capacitance

19. What charge is stored in a **180.0-μF** capacitor when 120.0 V is applied to it?

20. Find the charge stored when 5.50 V is applied to an 8.00-pF capacitor.

21. Calculate the voltage applied to a **2.00-μF** capacitor when it holds **3.10μC** of charge.

22. What voltage must be applied to an 8.00-nF capacitor to store 0.160 mC of charge?

23. What capacitance is needed to store **3.00μC** of charge at a voltage of 120 V?

24. What is the capacitance of a large Van de Graaff generator's terminal, given that it stores 8.00 mC of charge at a voltage of 12.0 MV?

25. The plates of an empty parallel-plate capacitor of capacitance 5.0 pF are 2.0 mm apart. What is the area of each plate?

26. A 60.0-pF vacuum capacitor has a plate area of $0.010m^2$. What is the separation between its plates?

27. A set of parallel plates has a capacitance of **5.0μF**. How much charge must be added to the plates to increase the potential difference between them by 100 V?

28. Consider Earth to be a spherical conductor of radius 6400 km and calculate its capacitance.

29. If the capacitance per unit length of a cylindrical capacitor is 20 pF/m, what is the ratio of the radii of the two cylinders?

30. An empty parallel-plate capacitor has a capacitance of **20μF**. How much charge must leak off its plates before the voltage across them is reduced by 100 V?

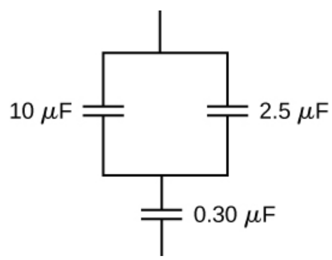
Capacitors in Series and in Parallel

31. A 4.00-pF is connected in series with an 8.00-pF capacitor and a 400-V potential difference is applied across the pair. (a) What is the charge on each capacitor?

(b) What is the voltage across each capacitor?

32. Three capacitors, with capacitances of $C_1 = 2.0\mu F, C_2 = 3.0\mu F$, and $C_3 = 6.0\mu F$, respectively, are connected in parallel. A 500-V potential difference is applied across the combination. Determine the voltage across each capacitor and the charge on each capacitor.

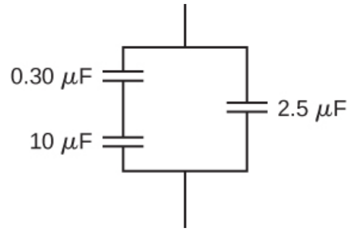
33. Find the total capacitance of this combination of series and parallel capacitors shown below.



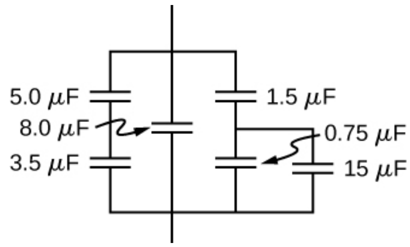
34. Suppose you need a capacitor bank with a total capacitance of 0.750 F but you have only 1.50-mF capacitors at your disposal. What is the smallest number of capacitors you could connect together to achieve your goal, and how would you connect them?

35. What total capacitances can you make by connecting a **5.00-μF** and a **8.00-μF** capacitor?

36. Find the equivalent capacitance of the combination of series and parallel capacitors shown below.



37. Find the net capacitance of the combination of series and parallel capacitors shown below.



38. A 40-pF capacitor is charged to a potential difference of 500 V. Its terminals are then connected to those of an uncharged 10-pF capacitor. Calculate:

(a) the original charge on the 40-pF capacitor;

(b) the charge on each capacitor after the connection is made; and

(c) the potential difference across the plates of each capacitor after the connection.

39. A **2.0-μF** capacitor and a **4.0-μF** capacitor are connected in series across a 1.0-kV potential. The charged capacitors are then disconnected from the source and connected to each other with terminals of like sign together. Find the charge on each capacitor and the voltage across each capacitor.

Energy Stored in a Capacitor

40. How much energy is stored in an **8.00-μF** capacitor whose plates are at a potential difference of 6.00 V?

41. A capacitor has a charge of **2.5μC** when connected to a 6.0-V battery. How much energy is stored in this capacitor?

42. How much energy is stored in the electrical field of a metal sphere of radius 2.0 m that is kept at a 10.0-V potential?

43. (a) What is the energy stored in the **10.0-μF** capacitor of a heart defibrillator charged to $9.00 \times 10^3 V$?

(b) Find the amount of the stored charge.

44. In open-heart surgery, a much smaller amount of energy will defibrillate the heart.

(a) What voltage is applied to the **8.00-μF** capacitor of a heart defibrillator that stores 40.0 J of energy?

(b) Find the amount of the stored charge.

45. A **165-μF** capacitor is used in conjunction with a dc motor. How much energy is stored in it when 119 V is applied?

46. Suppose you have a 9.00-V battery, a **2.00-μF** capacitor, and a **7.40-μF** capacitor.

(a) Find the charge and energy stored if the capacitors are connected to the battery in series.

(b) Do the same for a parallel connection.

47. An anxious physicist worries that the two metal shelves of a wood frame bookcase might obtain a high voltage if charged by static electricity, perhaps produced by friction.

(a) What is the capacitance of the empty shelves if they have area $1.00 \times 10^2 m$ and are 0.200 m apart?

(b) What is the voltage between them if opposite charges of magnitude 2.00 nC are placed on them?

(c) To show that this voltage poses a small hazard, calculate the energy stored.

(d) The actual shelves have an area 100 times smaller than these hypothetical shelves. Are his fears justified?

48. A parallel-plate capacitor is made of two square plates 25 cm on a side and 1.0 mm apart. The capacitor is connected to a 50.0-V battery. With the battery still connected, the plates are pulled apart to a separation of 2.00 mm. What are the energies stored in the capacitor before and after the plates are pulled farther apart? Why does the energy decrease even though work is done in separating the plates?

49. Suppose that the capacitance of a variable capacitor can be manually changed from 100 pF to 800 pF by turning a dial, connected to one set of plates by a shaft, from **0°** to **180°.** With the dial set at 180°180° (corresponding to **C=800pF**), the capacitor is connected to a 500-V source. After charging, the capacitor is disconnected from the source, and the dial is turned to **0°**. If friction is negligible, how much work is required to turn the dial from **180°** to **0°**?

Capacitor with a Dielectric

50. Show that for a given dielectric material, the maximum energy a parallel-plate capacitor can store is directly proportional to the volume of dielectric.

51. An air-filled capacitor is made from two flat parallel plates 1.0 mm apart. The inside area of each plate is 8.0cm28.0cm2.

(a) What is the capacitance of this set of plates?

(b) If the region between the plates is filled with a material whose dielectric constant is 6.0, what is the new capacitance?

52. A capacitor is made from two concentric spheres, one with radius 5.00 cm, the other with radius 8.00 cm.

(a) What is the capacitance of this set of conductors?

(b) If the region between the conductors is filled with a material whose dielectric constant is 6.00, what is the capacitance of the system?

53. A parallel-plate capacitor has charge of magnitude **9.00μC** on each plate and capacitance **3.00μF** when there is air between the plates. The plates are separated by 2.00 mm. With the charge on the plates kept constant, a dielectric with $\kappa = 5$ is inserted between the plates, completely filling the volume between the plates.

(a) What is the potential difference between the plates of the capacitor, before and after the dielectric has been inserted?

(b) What is the electrical field at the point midway between the plates before and after the dielectric is inserted?

54. Some cell walls in the human body have a layer of negative charge on the inside surface. Suppose that the surface charge densities are $\pm 0.50 \times 10^{-3} C/m^2$, the cell wall is $5.0 \times 10^{-9} m$ thick, and the cell wall material has a dielectric constant of **κ=5.4.**

(a) Find the magnitude of the electric field in the wall between two charge layers.

(b) Find the potential difference between the inside and the outside of the cell. Which is at higher potential?

(c) A typical cell in the human body has volume $10^{-16} m^3$. Estimate the total electrical field energy stored in the wall of a cell of this size when assuming that the cell is spherical. (**Hint:** Calculate the volume of the cell wall.)

55. A parallel-plate capacitor with only air between its plates is charged by connecting the capacitor to a battery. The capacitor is then disconnected from the battery, without any of the charge leaving the plates.

(a) A voltmeter reads 45.0 V when placed across the capacitor. When a dielectric is inserted between the plates, completely filling the space, the voltmeter reads 11.5 V. What is the dielectric constant of the material?

(b) What will the voltmeter read if the dielectric is now pulled away out so it fills only one-third of the space between the plates?

Molecular Model of a Dielectric

56. Two flat plates containing equal and opposite charges are separated by material 4.0 mm thick with a dielectric constant of 5.0. If the electrical field in the dielectric is 1.5 MV/m, what are

(a) the charge density on the capacitor plates, and

(b) the induced charge density on the surfaces of the dielectric?

57. For a Teflon™-filled, parallel-plate capacitor, the area of the plate is $50.0cm^2$ and the spacing between the plates is 0.50 mm. If the capacitor is connected to a 200-V battery, find

(a) the free charge on the capacitor plates,

(b) the electrical field in the dielectric, and

(c) the induced charge on the dielectric surfaces.

58. Find the capacitance of a parallel-plate capacitor having plates with a surface area of $5.00m^2$ and separated by 0.100 mm of Teflon™.

59. (a) What is the capacitance of a parallel-plate capacitor with plates of area $1.50m^2$ that are separated by 0.0200 mm of neoprene rubber?

(b) What charge does it hold when 9.00 V is applied to it?

60. Two parallel plates have equal and opposite charges. When the space between the plates is evacuated, the electrical field is $E = 3.20 \times 10^5 V/m$. When the space is filled with dielectric, the electrical field is $E = 2.50 \times 10^5 V/m$.

(a) What is the surface charge density on each surface of the dielectric?

(b) What is the dielectric constant?

61. The dielectric to be used in a parallel-plate capacitor has a dielectric constant of 3.60 and a dielectric strength of $1.60 \times 10^7 V/m$. The capacitor has to have a capacitance of 1.25 nF and must be able to withstand a maximum potential difference 5.5 kV. What is the minimum area the plates of the capacitor may have?

62. When a 360-nF air capacitor is connected to a power supply, the energy stored in the capacitor is **18.5μJ**. While the capacitor is connected to the power supply, a slab of dielectric is inserted that completely fills the space between the plates. This increases the stored energy by **23.2μJ**.

(a) What is the potential difference between the capacitor plates?

(b) What is the dielectric constant of the slab?

63. A parallel-plate capacitor has square plates that are 8.00 cm on each side and 3.80 mm apart. The space between the plates is completely filled with two square slabs of dielectric, each 8.00 cm on a side and 1.90 mm thick. One slab is Pyrex glass and the other slab is polystyrene. If the potential difference between the plates is 86.0 V, find how much electrical energy can be stored in this capacitor.

Application - RC Circuits

RC1 (49). The timing device in an automobile's intermittent wiper system is based on an RC time constant and utilizes a **0.500-μF** capacitor and a variable resistor. Over what range must R be made to vary to achieve time constants from 2.00 to 15.0 s?

RC2. A heart pacemaker fires 72 times a minute, each time a 25.0-nF capacitor is charged (by a battery in series with a resistor) to 0.632 of its full voltage. What is the value of the resistance?

RC3. The duration of a photographic flash is related to an RC time constant, which is **0.100μs** for a certain camera.

(a) If the resistance of the flash lamp is **0.0400Ω** during discharge, what is the size of the capacitor supplying its energy?

(b) What is the time constant for charging the capacitor, if the charging resistance is **800kΩ**?

RC4. A 2.00- and a **7.50-μF** capacitor can be connected in series or parallel, as can a 25.0- and a **100-kΩ** resistor. Calculate the four RC time constants possible from connecting the resulting capacitance and resistance in series.

RC5. A **500-Ω** resistor, an uncharged **1.50-μF** capacitor, and a 6.16-V source voltage are connected in series.

(a) What is the initial current?

(b) What is the RC time constant?

(c) What is the current after one time constant? (d) What is the voltage on the capacitor after one time constant?

RC6. A heart defibrillator being used on a patient has an RC time constant of 10.0 ms due to the resistance of the patient and the capacitance of the defibrillator.

(a) If the defibrillator has a capacitance of **8.00μF,** what is the resistance of the path through the patient? (You may neglect the capacitance of the patient and the resistance of the defibrillator.)

(b) If the initial voltage is 12.0 kV, how long does it take to decline to $6.00 \times 10^2 V$?

RC7. An ECG monitor must have an RC time constant less than $1.00 \times 10^2 \mu s$ to be able to measure variations in voltage over small time intervals.

(a) If the resistance of the circuit (due mostly to that of the patient's chest) is **1.00kΩ**, what is the maximum capacitance of the circuit?

(b) Would it be difficult in practice to limit the capacitance to less than the value found in (a)?

RC8. Using the exact exponential treatment, determine how much time is required to charge an initially uncharged 100-pF capacitor through a **75.0-MΩ** resistor to **90.0%** of its final voltage.

RC9. If you wish to take a picture of a bullet traveling at 500 m/s, then a very brief flash of light produced by an RC discharge through a flash tube can limit blurring. Assuming 1.00 mm of motion during one RC constant is acceptable, and given that the flash is driven by a **600-μF** capacitor, what is the resistance in the flash tube?

Additional Problems

64. A capacitor is made from two flat parallel plates placed 0.40 mm apart. When a charge of **0.020μC** is placed on the plates the potential difference between them is 250 V.

(a) What is the capacitance of the plates?

(b) What is the area of each plate?

(c) What is the charge on the plates when the potential difference between them is 500 V?

(d) What maximum potential difference can be applied between the plates so that the magnitude of electrical fields between the plates does not exceed 3.0 MV/m?

65. An air-filled (empty) parallel-plate capacitor is made from two square plates that are 25 cm on each side and 1.0 mm apart. The capacitor is connected to a 50-V battery and fully charged. It is then disconnected from the battery and its plates are pulled apart to a separation of 2.00 mm.

(a) What is the capacitance of this new capacitor?

(b) What is the charge on each plate?

(c) What is the electrical field between the plates?

66. Suppose that the capacitance of a variable capacitor can be manually changed from 100 to 800 pF by turning a dial connected to one set of plates by a shaft, from **0°** to **180°**. With the dial set at **180°** (corresponding to **C=800pF**), the capacitor is connected to a 500-V source. After charging, the capacitor is disconnected from the source, and the dial is turned to **0°**. (a) What is the charge on the capacitor? (b) What is the voltage across the capacitor when the dial is set to **0°**?

67. Earth can be considered as a spherical capacitor with two plates, where the negative plate is the surface of Earth and the positive plate is the bottom of the ionosphere, which is located at an altitude of approximately 70 km. The potential difference between Earth's surface and the ionosphere is about 350,000 V.

(a) Calculate the capacitance of this system.

(b) Find the total charge on this capacitor.

(c) Find the energy stored in this system.

68. A **4.00-μF** capacitor and a **6.00-μF** capacitor are connected in parallel across a 600-V supply line.

(a) Find the charge on each capacitor and voltage across each.

(b) The charged capacitors are disconnected from the line and from each other. They are then reconnected to each other with terminals of unlike sign together. Find the final charge on each capacitor and the voltage across each.

69. Three capacitors having capacitances of 8.40, 8.40, and 4.20 **μF**, respectively, are connected in series across a 36.0-V potential difference.

(a) What is the charge on the **4.20-μF** capacitor?

(b) The capacitors are disconnected from the potential difference without allowing them to discharge. They are then reconnected in parallel with each other with the positively charged plates connected together. What is the voltage across each capacitor in the parallel combination?

70. A parallel-plate capacitor with capacitance **5.0μF** is charged with a 12.0-V battery, after which the battery is disconnected. Determine the minimum work required to increase the separation between the plates by a factor of 3.

71. (a) How much energy is stored in the electrical fields in the capacitors (in total) shown below?

(b) Is this energy equal to the work done by the 400-V source in charging the capacitors?

72. Three capacitors having capacitances 8.4, 8.4, and 4.2 **μF** are connected in series across a 36.0-V potential difference.

(a) What is the total energy stored in all three capacitors?

(b) The capacitors are disconnected from the potential difference without allowing them to discharge. They are then reconnected in parallel with each other with the positively charged plates connected together. What is the total energy now stored in the capacitors?

73. (a) An **8.00-μF** capacitor is connected in parallel to another capacitor, producing a total capacitance of **5.00μF**. What is the capacitance of the second capacitor?

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

74. (a) On a particular day, it takes $9.60 \times 10^3 J$ of electrical energy to start a truck's engine. Calculate the capacitance of a capacitor that could store that amount of energy at 12.0 V.

(b) What is unreasonable about this result?

(c) Which assumptions are responsible?

75. (a) A certain parallel-plate capacitor has plates of area $4.00m^2$, separated by 0.0100 mm of nylon, and stores 0.170 C of charge. What is the applied voltage?

(b) What is unreasonable about this result?

(c) Which assumptions are responsible or inconsistent?

76. A prankster applies 450 V to an **80.0-μF** capacitor and then tosses it to an unsuspecting victim. The victim's finger is burned by the discharge of the capacitor through 0.200 g of flesh. Estimate, what is the temperature increase of the flesh? Is it reasonable to assume that no thermodynamic phase change happened?

Advanced Problems

AP1. A circuit contains a D cell battery, a switch, a **20-Ω** resistor, and four 20-mF capacitors connected in series.

(a) What is the equivalent capacitance of the circuit?

(b) What is the RC time constant?

(c) How long before the current decreases to **50%** of the initial value once the switch is closed?

AP2. A circuit contains a D-cell battery, a switch, a **20-Ω** resistor, and three 20-mF capacitors. The capacitors are connected in parallel, and the parallel connection of capacitors are connected in series with the switch, the resistor and the battery.

(a) What is the equivalent capacitance of the circuit?

(b) What is the RC time constant?

(c) How long before the current decreases to **50%** of the initial value once the switch is closed?

AP3. Consider the circuit below. The battery has an source voltage of **ε=30.00V** and an internal resistance of **r=1.00Ω**.

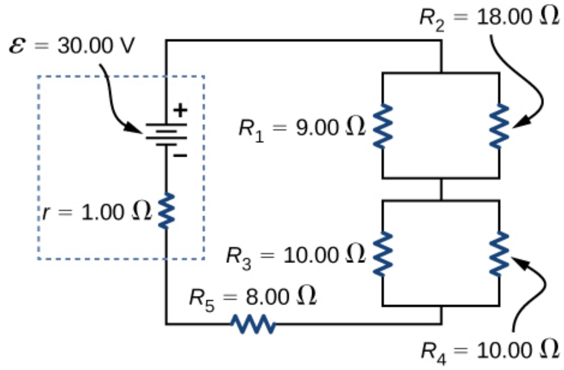
(a) Find the equivalent resistance of the circuit and the current out of the battery.

(b) Find the current through each resistor.

(c) Find the potential drop across each resistor.

(d) Find the power dissipated by each resistor.

(e) Find the total power supplied by the batteries.



AP4. A homemade capacitor is constructed of 2 sheets of aluminum foil with an area of 2.00 square meters, separated by paper, 0.05 mm thick, of the same area and a dielectric constant of 3.7. The homemade capacitor is connected in series with a **100.00-Ω** resistor, a switch, and a 6.00-V voltage source.

(a) What is the RC time constant of the circuit?

(b) What is the initial current through the circuit, when the switch is closed? (c) How long does it take the current to reach one third of its initial value?

AP5. A student makes a homemade resistor from a graphite pencil 5.00 cm long, where the graphite is 0.05 mm in diameter. The resistivity of the graphite is $\rho = 1.38 \times 10^{-5} \Omega/m$. The homemade resistor is place in series with a switch, a 10.00-mF capacitor and a 0.50-V power source.

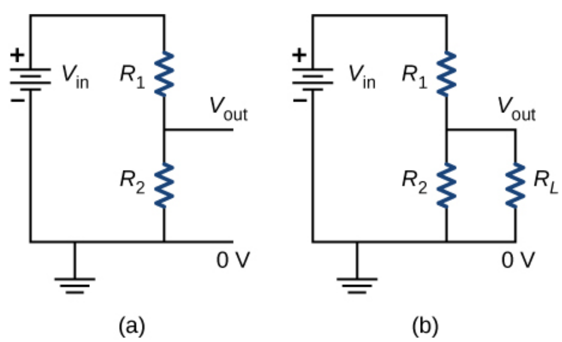
(a) What is the RC time constant of the circuit?

(b) What is the potential drop across the pencil 1.00 s after the switch is closed?

AP6. The rather simple circuit shown below is known as a voltage divider. The symbol consisting of three horizontal lines is represents "ground" and can be defined as the point where the potential is zero. The voltage divider is widely used in circuits and a single voltage source can be used to provide reduced voltage to a load resistor as shown in the second part of the figure. (a) What is the output voltage V_{out} of circuit

(a) in terms of R_1, R_2, and V_{in}?

(b) What is the output voltage V_{out} of circuit (b) in terms of R_1, R_2, R_L, and V_{in}?



AP7. Three **300-Ω** resistors are connect in series with an AAA battery with a rating of 3 AmpHours. (a) How long can the battery supply the resistors with power? (b) If the resistors are connected in parallel, how long can the battery last?

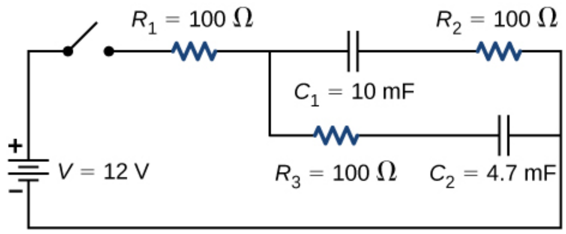
AP8. Consider a circuit that consists of a real battery with an source voltage εε and an internal resistance of r connected to a variable resistor R.

(a) In order for the terminal voltage of the battery to be equal to the source voltage of the battery, what should the resistance of the variable resistor be adjusted to?

(b) In order to get the maximum current from the battery, what should the resistance of the variable resistor be adjusted to?

(c) In order for the maximum power output of the battery to be reached, what should the resistance of the variable resistor be set to?

AP9. Consider the circuit shown below. What is the energy stored in each capacitor after the switch has been closed for a very long time?



AP10. Consider a circuit consisting of a battery with an source voltage εε and an internal resistance of r connected in series with a resistorR and a capacitor C. Show that the total energy supplied by the battery while charging the battery is equal to $\varepsilon^2 C$.

AP11. Consider the circuit shown below. The terminal voltages of the batteries are shown.

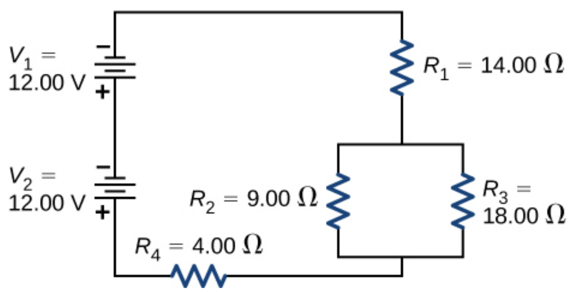
(a) Find the equivalent resistance of the circuit and the current out of the battery.

(b) Find the current through each resistor.

(c) Find the potential drop across each resistor.

(d) Find the power dissipated by each resistor.

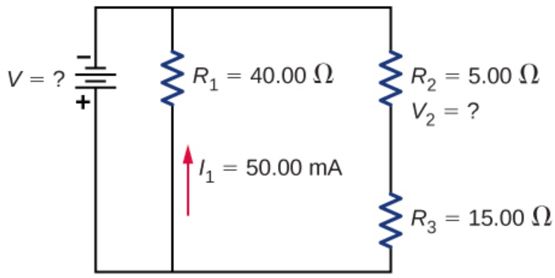
(e) Find the total power supplied by the batteries.



AP12. Consider the circuit shown below.

(a) What is the terminal voltage of the battery?

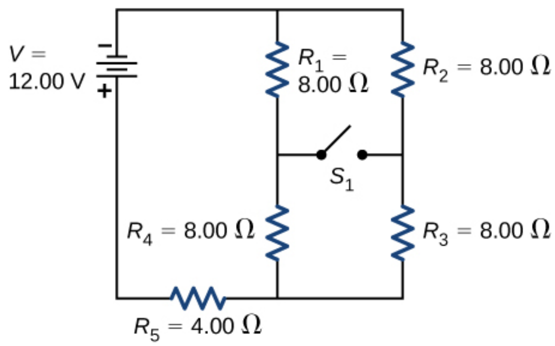
(b) What is the potential drop across resistor R_2?



AP13. Consider the circuit shown below.

(a)Determine the equivalent resistance and the current from the battery with switch S_1 open.

(b) Determine the equivalent resistance and the current from the battery with switch S_1 closed.



AP14. Two resistors, one having a resistance of **145Ω**, are connected in parallel to produce a total resistance of **150Ω**.

(a) What is the value of the second resistance?

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

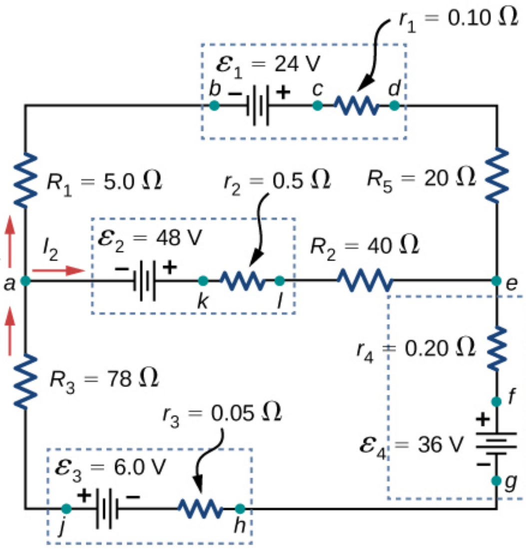
AP15. Two resistors, one having a resistance of **900kΩ**, are connected in series to produce a total resistance of **0.500MΩ**.

(a) What is the value of the second resistance?

(b) What is unreasonable about this result?

(c) Which assumptions are unreasonable or inconsistent?

AP16. Apply the junction rule at point a shown below.



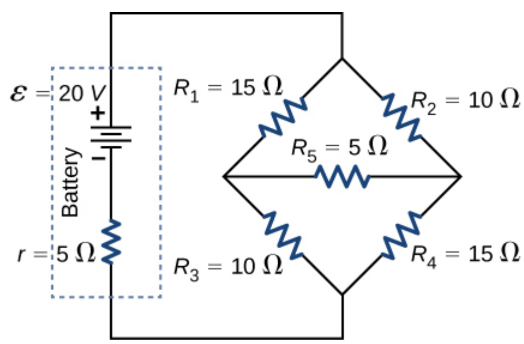
AP17. Apply the loop rule to Loop akledcba in the preceding problem.

AP18. Find the currents flowing in the circuit in the preceding problem. Explicitly show how you follow the steps in the Problem-Solving Strategy: Series and Parallel Resistors.

AP19. Consider the circuit shown below.

(a) Find the current through each resistor.

(b) Check the calculations by analyzing the power in the circuit.



AP20. A flashing lamp in a Christmas earring is based on an RC discharge of a capacitor through its resistance. The effective duration of the flash is 0.250 s, during which it produces an average 0.500 W from an average 3.00 V.

(a) What energy does it dissipate?

(b) How much charge moves through the lamp?

(c) Find the capacitance.

(d) What is the resistance of the lamp? (Since average values are given for some quantities, the shape of the pulse profile is not needed.)

AP21. A **160-μF** capacitor charged to 450 V is discharged through a **31.2-kΩ** resistor.

(a) Find the time constant.

(b) Calculate the temperature increase of the resistor, given that its mass is 2.50 g and its specific heat is **1.67kJ/kg⋅°C**, noting that most of the thermal energy is retained in the short time of the discharge.

(c) Calculate the new resistance, assuming it is pure carbon.

(d) Does this change in resistance seem significant?

Challenge Problems

77. A spherical capacitor is formed from two concentric spherical conducting spheres separated by vacuum. The inner sphere has radius 12.5 cm and the outer sphere has radius 14.8 cm. A potential difference of 120 V is applied to the capacitor.

(a) What is the capacitance of the capacitor?

(b) What is the magnitude of the electrical field at **r=12.6cm**, just outside the inner sphere?

(c) What is the magnitude of the electrical field at **r=14.7cm**, just inside the outer sphere?

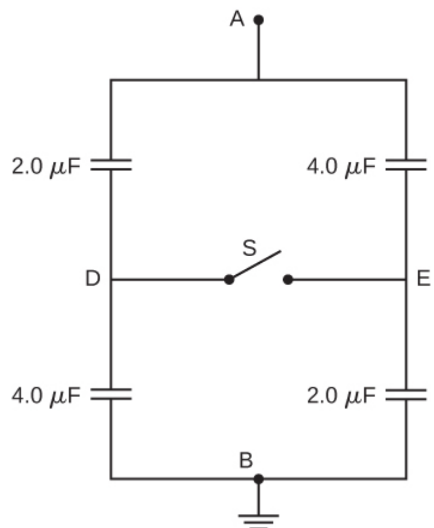
(d) For a parallel-plate capacitor the electrical field is uniform in the region between the plates, except near the edges of the plates. Is this also true for a spherical capacitor?

78. The network of capacitors shown below are all uncharged when a 300-V potential is applied between points A and B with the switch S open.

(a) What is the potential difference $V_E - V_D$?

(b) What is the potential at point E after the switch is closed?

(c) How much charge flows through the switch after it is closed?



79. Electronic flash units for cameras contain a capacitor for storing the energy used to produce the flash. In one such unit the flash lasts for 1/675 fraction of a second with an average light power output of 270 kW.

(a) If the conversion of electrical energy to light is 95% efficient (because the rest of the energy goes to thermal energy), how much energy must be stored in the capacitor for one flash?

(b) The capacitor has a potential difference between its plates of 125 V when the stored energy equals the value stored in part (a). What is the capacitance?

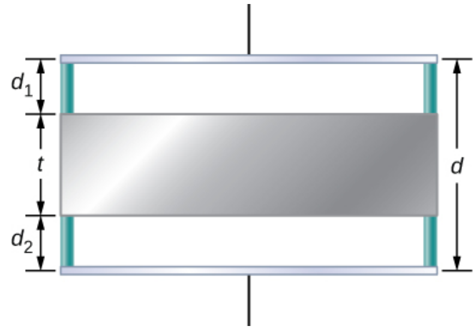
80. A spherical capacitor is formed from two concentric spherical conducting shells separated by a vacuum. The inner sphere has radius 12.5 cm and the outer sphere has radius 14.8 cm. A potential difference of 120 V is applied to the capacitor.

(a) What is the energy density at **r=12.6cm**, just outside the inner sphere?

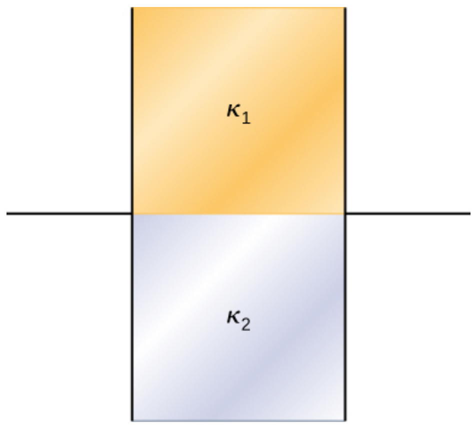
(b) What is the energy density at **r=14.7cm**, just inside the outer sphere?

(c) For the parallel-plate capacitor the energy density is uniform in the region between the plates, except near the edges of the plates. Is this also true for the spherical capacitor?

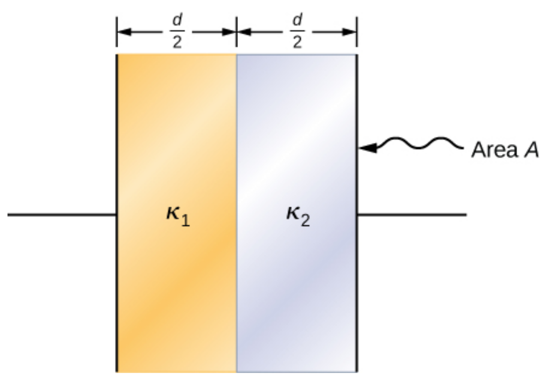
81. A metal plate of thickness t is held in place between two capacitor plates by plastic pegs, as shown below. The effect of the pegs on the capacitance is negligible. The area of each capacitor plate and the area of the top and bottom surfaces of the inserted plate are all A. What is the capacitance of this system?



82. A parallel-plate capacitor is filled with two dielectrics, as shown below. When the plate area is A and separation between plates is d, show that the capacitance is given by $C=\varepsilon_0 \frac{A}{d} \frac{\kappa_1+\kappa_2}{2}$.



83. A parallel-plate capacitor is filled with two dielectrics, as shown below. Show that the capacitance is given by $C=2\varepsilon_0 \frac{A}{d} \frac{\kappa_1 \kappa_2}{\kappa_1+\kappa_2}$.



84. A capacitor has parallel plates of area $12 cm^2$ separated by 2.0 mm. The space between the plates is filled with polystyrene.

(a) Find the maximum permissible voltage across the capacitor to avoid dielectric breakdown.

(b) When the voltage equals the value found in part (a), find the surface charge density on the surface of the dielectric.

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).

- **8.8: Capacitance (Exercises)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.
- **10.9: Direct-Current Circuits (Exercise)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.

7.11: Capacitance (Answers)

Check Your Understanding

8.1. $1.1 \times 10^{-3} m$

8.3. 3.59 cm, 17.98 cm

8.4. a. 25.0 pF;

b. 9.2

8.5. a. $C = 0.86 pF, Q_1 = 10 pC, Q_2 = 3.4 pC, Q_3 = 6.8 pC$;

b. $C = 2.3 pF, Q_1 = 12 pC, Q_2 = Q_3 = 16 pC$;

c. $C = 2.3 pF, Q_1 = 9.0 pC, Q_2 = 18 pC, Q_3 = 12 pC, Q_4 = 15 pC$

8.6. a.$4.0 \times 10^{-13} J$; b. 9 times

8.7. a. 3.0; b. $C = 3.0 C_0$

8.9. a. $C_0 = 20 pF, C = 42 pF$;

b. $Q_0 = 0.8 nC, Q = 1.7 nC$;

c. $V_0 = V = 40 V$; d. $U_0 = 16 nJ, U = 34 nJ$

Conceptual Questions

1. no; yes

3. false

5. no

7. $3.0 \mu F, 0.33 \mu F$

9. answers may vary

11. Dielectric strength is a critical value of an electrical field above which an insulator starts to conduct; a dielectric constant is the ratio of the electrical field in vacuum to the net electrical field in a material.

13. Water is a good solvent.

15. When energy of thermal motion is large (high temperature), an electrical field must be large too in order to keep electric dipoles aligned with it.

17. answers may vary

Problems

19. 21.6 mC

21. 1.55 V

23. 25.0 nF

25. $1.1 \times 10^{-3} m^2$

27. 500 μC

29. 1:16

31. a. 1.07 nC;

b. 267 V, 133 V

33. $0.29 \mu F$

34. 500 capacitors; connected in parallel

35. $3.08 \mu F$ (series) and 13.0μ (parallel)

37. $11.4 \mu F$

39. 0.89 mC; 1.78 mC; 444 V

41. $7.5 \mu J$

43. a. 405 J; b. 90.0 mC

45. 1.15 J

47. a. $4.43 \times 10^{-9} F$;

b. 0.453 V;

c. $4.53 \times 10^{-10} J$;

d. no

49. 0.7 mJ

51. a. 7.1 pF;

b. 42 pF

53. a. before 3.00 V; after 0.600 V;

b. before 1500 V/m; after 300 V/m

55. a. 3.91;

b. 22.8 V

57. a. 37 nC;

b. 0.4 MV/m;

c. 19 nC

59. a. $4.4 \mu F$;

b. $4.0 \times 10^{-5} C$

61. $0.0135 m^2$

63. $0.185 \mu J$

Additional Problems

65. a. 0.277 nF;

b. 27.7 nC;

c. 50 kV/m

67. a. 0.065 F;

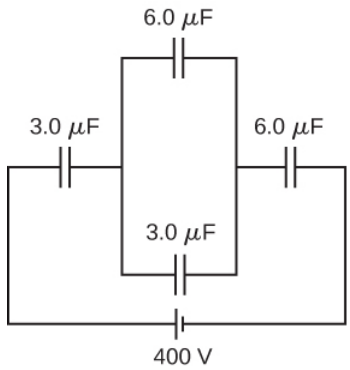
b. 23,000 C;

c. 4.0 GJ

69. a. $75.6 \mu C$; b. 10.8 V

71. a. 0.13 J;

b. no, because of resistive heating in connecting wires that is always present, but the circuit schematic does not indicate resistors



73. a. $-3.00 \mu F$;

b. You cannot have a negative C_2 capacitance.

c. The assumption that they were hooked up in parallel, rather than in series, is incorrect. A parallel connection always produces a greater capacitance, while here a smaller capacitance was assumed. This could only happen if the capacitors are connected in series.

75. a. 14.2 kV;

b. The voltage is unreasonably large, more than 100 times the breakdown voltage of nylon.

c. The assumed charge is unreasonably large and cannot be stored in a capacitor of these dimensions.

Challenge Problems

77. a. 89.6 pF;

b. 6.09 kV/m;

c. 4.47 kV/m;

d. no

79. a. 421 J;

b. 53.9 mF

81. $C = \varepsilon_0 A/(d_1 + d_2)$

83. proof

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).

CHAPTER OVERVIEW

8: The Magnetic Field

8.1: Introduction

For the past few chapters, we have been studying electrostatic forces and fields, which are caused by electric charges at rest. These electric fields can move other free charges, such as producing a current in a circuit; however, the electrostatic forces and fields themselves come from other static charges. In this chapter, we see that when an electric charge moves, it generates other forces and fields. These additional forces and fields are what we commonly call magnetism.

Figure 8.1.1: An industrial electromagnet is capable of lifting thousands of pounds of metallic waste. (credit: modification of work by “BedfordAl”/Flickr)

Before we examine the origins of magnetism, we first describe what it is and how magnetic fields behave. Once we are more familiar with magnetic effects, we can explain how they arise from the behavior of atoms and molecules, and how magnetism is related to electricity. The connection between electricity and magnetism is fascinating from a theoretical point of view, but it is also immensely practical, as shown by an industrial electromagnet that can lift thousands of pounds of metal.

8.2: Introduction to Magnetism

Learning Objectives

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

- Describe the difference between the north and south poles of a magnet.
- Describe how magnetic poles interact with each other.

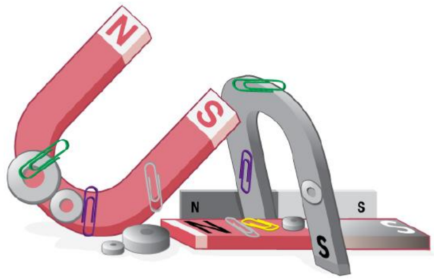


Figure 8.2.1: Magnets come in various shapes, sizes, and strengths. All have both a north pole and a south pole. There is never an isolated pole (a monopole).

All magnets attract iron, such as that in a refrigerator door. However, magnets may attract or repel other magnets. Experimentation shows that all magnets have two poles. If freely suspended, one pole will point toward the north. The two poles are thus named the north magnetic pole and the south magnetic pole (or more properly, north-seeking and south-seeking poles, for the attractions in those directions).

UNIVERSAL CHARACTERISTICS OF MAGNETS AND MAGNET POLES

It is a universal characteristic of all magnets that *like poles repel and unlike poles attract*. (Note the similarity with electrostatics: unlike charges attract and like charges repel.)

Further experimentation shows that it is *impossible to separate north and south poles* in the manner that + and − charges can be separated.

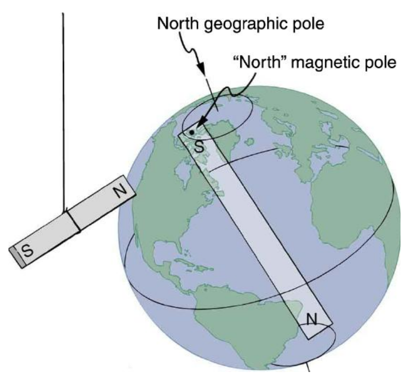


Figure 8.2.2: One end of a bar magnet is suspended from a thread that points toward north. The magnet's two poles are labeled N and S for north-seeking and south-seeking poles, respectively.

MISCONCEPTION ALERT: EARTH'S GEOGRAPHIC NORTH POLE HIDES AN S

The Earth acts like a very large bar magnet with its south-seeking pole near the geographic North Pole. That is why the north pole of your compass is attracted toward the geographic north pole of the Earth—because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole! Confusion arises because the geographic term "North Pole" has come to be used (incorrectly) for the magnetic pole that is near the North Pole. Thus, "North magnetic pole" is actually a misnomer—it should be called the South magnetic pole.

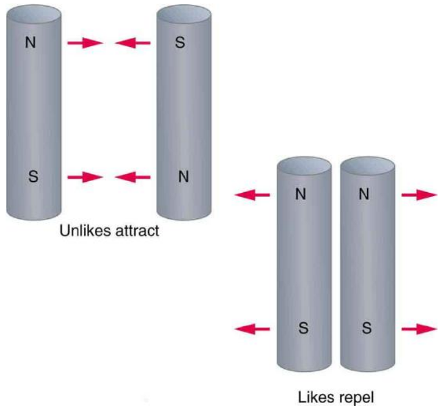


Figure 8.2.3: Unlike poles attract, whereas like poles repel.

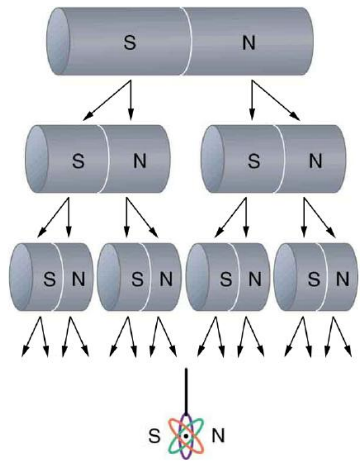


Figure 8.2.4: North and south poles always occur in pairs. Attempts to separate them result in more pairs of poles. If we continue to split the magnet, we will eventually get down to an iron atom with a north pole and a south pole—these, too, cannot be separated.

The fact that magnetic poles always occur in pairs of north and south is true from the very large scale—for example, sunspots always occur in pairs that are north and south magnetic poles—all the way down to the very small scale. Magnetic atoms have both a north pole and a south pole, as do many types of subatomic particles, such as electrons, protons, and neutrons.

MAKING CONNECTIONS: TAKE-HOME EXPERIMENT -- REFRIGERATOR MAGNETS

We know that like magnetic poles repel and unlike poles attract. See if you can show this for two refrigerator magnets. Will the magnets stick if you turn them over? Why do they stick to the door anyway? What can you say about the magnetic properties of the door next to the magnet? Do refrigerator magnets stick to metal or plastic spoons? Do they stick to all types of metal?

Summary

- Magnetism is a subject that includes the properties of magnets, the effect of the magnetic force on moving charges and currents, and the creation of magnetic fields by currents.
- There are two types of magnetic poles, called the north magnetic pole and south magnetic pole.
- North magnetic poles are those that are attracted toward the Earth's geographic north pole.
- Like poles repel and unlike poles attract.
- Magnetic poles always occur in pairs of north and south—it is not possible to isolate north and south poles.

Glossary

north magnetic pole

the end or the side of a magnet that is attracted toward Earth's geographic north pole

south magnetic pole

the end or the side of a magnet that is attracted toward Earth's geographic south pole

8.3: Magnetism and Its Historical Discoveries

Learning Objectives

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

- Explain attraction and repulsion by magnets
- Describe the historical and contemporary applications of magnetism

Magnetism has been known since the time of the ancient Greeks, but it has always been a bit mysterious. You can see electricity in the flash of a lightning bolt, but when a compass needle points to magnetic north, you can't see any force causing it to rotate. People learned about magnetic properties gradually, over many years, before several physicists of the nineteenth century connected magnetism with electricity. In this section, we review the basic ideas of magnetism and describe how they fit into the picture of a magnetic field.

Brief History of Magnetism

Magnets are commonly found in everyday objects, such as toys, hangers, elevators, doorbells, and computer devices. Experimentation on these magnets shows that all magnets have two poles: One is labeled north (N) and the other is labeled south (S). Magnetic poles repel if they are alike (both N or both S), they attract if they are opposite (one N and the other S), and both poles of a magnet attract unmagnetized pieces of iron. An important point to note here is that you cannot isolate an individual magnetic pole. Every piece of a magnet, no matter how small, which contains a north pole must also contain a south pole.

Note

Visit this website for an interactive demonstration of magnetic north and south poles.

An example of a magnet is a **compass needle**. It is simply a thin bar magnet suspended at its center, so it is free to rotate in a horizontal plane. Earth itself also acts like a very large bar magnet, with its south-seeking pole near the geographic North Pole (Figure $8.3.1$). The north pole of a compass is attracted toward Earth's geographic North Pole because the magnetic pole that is near the geographic North Pole is actually a south magnetic pole. Confusion arises because the geographic term "North Pole" has come to be used (incorrectly) for the magnetic pole that is near the North Pole. Thus, "**north magnetic pole**" is actually a misnomer—it should be called the **south magnetic pole**. [Note that the orientation of Earth's magnetic field is not permanent but changes ("flips") after long time intervals. Eventually, Earth's north magnetic pole may be located near its geographic North Pole.]

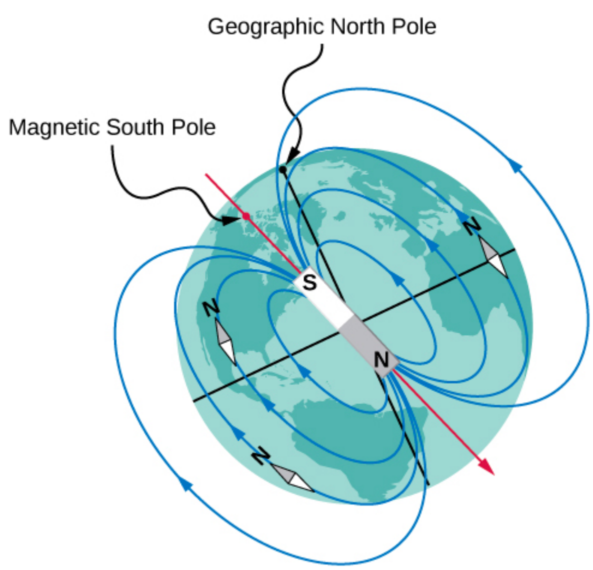


Figure $8.3.1$: The north pole of a compass needle points toward the south pole of a magnet, which is how today's magnetic field is oriented from inside Earth. It also points toward Earth's geographic North Pole because the geographic North Pole is near the magnetic south pole.

Back in 1819, the Danish physicist Hans **Oersted** was performing a lecture demonstration for some students and noticed that a compass needle moved whenever current flowed in a nearby wire. Further investigation of this phenomenon convinced Oersted that an electric current could somehow cause a magnetic force. He reported this finding to an 1820 meeting of the French Academy of Science.

Soon after this report, Oersted's investigations were repeated and expanded upon by other scientists. Among those whose work was especially important were Jean-Baptiste **Biot** and Felix **Savart**, who investigated the forces exerted on magnets by currents; André Marie **Ampère**, who studied the forces exerted by one current on another; François **Arago**, who found that iron could be magnetized by a current; and Humphry **Davy**, who discovered that a magnet exerts a force on a wire carrying an electric current. Within 10 years of Oersted's discovery, Michael **Faraday** found that the relative motion of a magnet and a metallic wire induced current in the wire. This finding showed not only that a current has a magnetic effect, but that a magnet can generate electric current. You will see later that the names of Biot, Savart, Ampère, and Faraday are linked to some of the fundamental laws of electromagnetism.

The evidence from these various experiments led Ampère to propose that electric current is the source of all magnetic phenomena. To explain permanent magnets, he suggested that matter contains microscopic current loops that are somehow aligned when a material is magnetized. Today, we know that permanent magnets are actually created by the alignment of spinning electrons, a situation quite similar to that proposed by Ampère. This model of permanent magnets was developed by Ampère almost a century before the atomic nature of matter was understood. (For a full quantum mechanical treatment of magnetic spins, see Quantum Mechanics and Atomic Structure.)

Contemporary Applications of Magnetism

Today, magnetism plays many important roles in our lives. Physicists' understanding of magnetism has enabled the development of technologies that affect both individuals and society. The electronic tablet in your purse or backpack, for example, wouldn't have been possible without the applications of magnetism and electricity on a small scale (Figure 8.3.2). Weak changes in a magnetic field in a thin film of iron and chromium were discovered to bring about much larger changes in resistance, called **giant magnetoresistance**. Information can then be recorded magnetically based on the direction in which the iron layer is magnetized. As a result of the discovery of giant magnetoresistance and its applications to digital storage, the 2007 Nobel Prize in Physics was awarded to Albert Fert from France and Peter Grunberg from Germany.

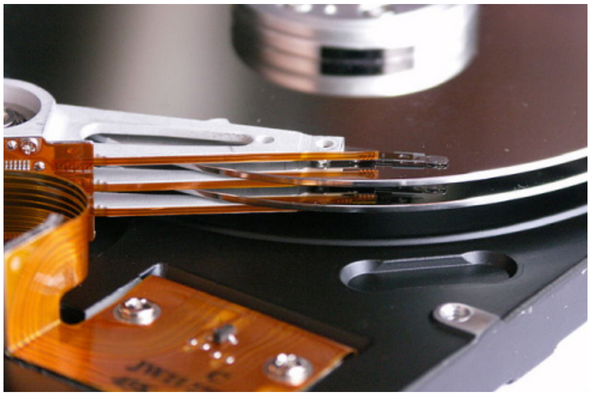


Figure 8.3.2: Engineering technology like computer storage would not be possible without a deep understanding of magnetism. (credit: Klaus Eifert)

All electric motors—with uses as diverse as powering refrigerators, starting cars, and moving elevators—contain magnets. Generators, whether producing hydroelectric power or running bicycle lights, use magnetic fields. Recycling facilities employ magnets to separate iron from other refuse. Research into using magnetic containment of fusion as a future energy source has been continuing for several years. Magnetic resonance imaging (MRI) has become an important diagnostic tool in the field of medicine, and the use of magnetism to explore brain activity is a subject of contemporary research and development. The list of applications also includes computer hard drives, tape recording, detection of inhaled asbestos, and levitation of high-speed trains. Magnetism is involved in the structure of atomic energy levels, as well as the motion of cosmic rays and charged particles trapped in the Van Allen belts around Earth. Once again, we see that all these disparate phenomena are linked by a small number of underlying physical principles.

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).

8.4: The Biot-Savart Law

Learning Objectives

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

- Explain how to derive a magnetic field from an arbitrary current in a line segment
- Calculate magnetic field from the Biot-Savart law in specific geometries, such as a current in a line and a current in a circular arc

We have seen that mass produces a gravitational field and also interacts with that field. Charge produces an electric field and also interacts with that field. Since moving charge (that is, current) interacts with a magnetic field, we might expect that it also creates that field—and it does.

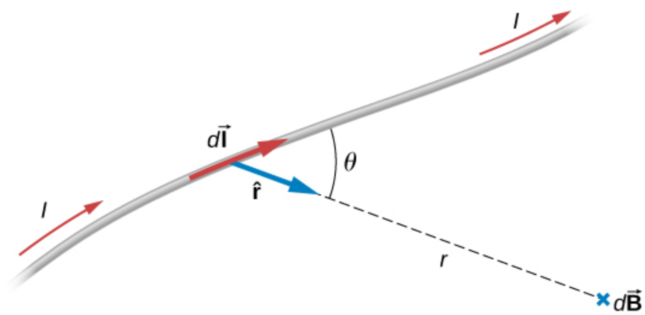


Figure 8.4.1: A current element $Id\vec{l}$ produces a magnetic field at point P given by the Biot-Savart law (Equation 8.4.4).

The equation used to calculate the magnetic field produced by a current is known as the Biot-Savart law. It is an empirical law named in honor of two scientists who investigated the interaction between a straight, current-carrying wire and a permanent magnet. This law enables us to calculate the magnitude and direction of the magnetic field produced by a current in a wire. The **Biot-Savart law** states that at any point P (Figure 8.4.1), the magnetic field $d\vec{B}$ due to an element $d\vec{l}$ of a current-carrying wire is given by

$$d\vec{B} = \frac{\mu_0}{4\pi}\frac{Id\vec{l}\times\hat{r}}{r^2}. \tag{8.4.1}$$

The constant μ_0 is known as the **permeability of free space** and is exactly

$$\mu_0 = 4\pi\times 10^{-7}\, T\cdot m/A \tag{8.4.2}$$

in the SI system. The infinitesimal wire segment $d\vec{l}$ is in the same direction as the current I (assumed positive), r is the distance from $d\vec{l}$ to P and $\hat{r}$ is a unit vector that points from $d\vec{l}$ to P, as shown in Figure 8.4.1. The direction of $d\vec{B}$ is determined by applying the right-hand rule to the vector product $d\vec{l}\times\hat{r}$. The magnitude of $d\vec{B}$ is

$$dB = \frac{\mu_0}{4\pi}\frac{I\,dl\,\sin\theta}{r^2} \tag{8.4.3}$$

where θ is the angle between $d\vec{l}$ and $\hat{r}$. Notice that if $\theta = 0$, then $d\vec{B} = \vec{0}$. The field produced by a current element $Id\vec{l}$ has no component parallel to $d\vec{l}$.

The magnetic field due to a finite length of current-carrying wire is found by integrating Equation 8.4.2 along the wire, giving us the usual form of the Biot-Savart law.

Biot-Savart law

The magnetic field $\vec{B}$ due to an element $d\vec{l}$ of a current-carrying wire is given by

$$\vec{B} = \frac{\mu_0}{4\pi}\int_{wire}\frac{I\,d\vec{l}\times\hat{r}}{r^2}. \tag{8.4.4}$$

Since this is a vector integral, contributions from different current elements may not point in the same direction. Consequently, the integral is often difficult to evaluate, even for fairly simple geometries. The following strategy may be helpful.

Problem-Solving Strategy: Solving Biot-Savart Problems

To solve Biot-Savart law problems, the following steps are helpful:

1. Identify that the Biot-Savart law is the chosen method to solve the given problem. If there is symmetry in the problem comparing $\vec{B}$ and $d\vec{l}$, Ampère's law may be the preferred method to solve the question.
2. Draw the current element length $d\vec{l}$ and the unit vector $\hat{r}$ noting that $d\vec{l}$ points in the direction of the current and $\hat{r}$ points from the current element toward the point where the field is desired.
3. Calculate the cross product $d\vec{l} \times \hat{r}$. The resultant vector gives the direction of the magnetic field according to the Biot-Savart law.
4. Use Equation 8.4.4 and substitute all given quantities into the expression to solve for the magnetic field. Note all variables that remain constant over the entire length of the wire may be factored out of the integration.
5. Use the right-hand rule to verify the direction of the magnetic field produced from the current or to write down the direction of the magnetic field if only the magnitude was solved for in the previous part.

Calculating Magnetic Fields of Short Current Segments

A short wire of length 1.0 cm carries a current of 2.0 A in the vertical direction (Figure 8.4.2). The rest of the wire is shielded so it does not add to the magnetic field produced by the wire. Calculate the magnetic field at point **P**, which is 1 meter from the wire in the **x**-direction.

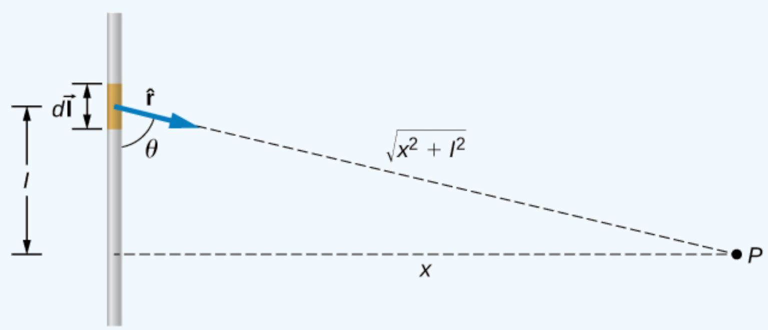


Figure 8.4.2: A small line segment carries a current I in the vertical direction. What is the magnetic field at a distance **x** from the segment?

Strategy

We can determine the magnetic field at point P using the Biot-Savart law. Since the current segment is much smaller than the distance **x**, we can drop the integral from the expression. The integration is converted back into a summation, but only for small dl, which we now write as Δl. Another way to think about it is that each of the radius values is nearly the same, no matter where the current element is on the line segment, if Δl is small compared to **x**. The angle θ is calculated using a tangent function. Using the numbers given, we can calculate the magnetic field at P.

Solution

The angle between $\Delta\vec{l}$ and $\hat{r}$ is calculated from trigonometry, knowing the distances **l** and **x** from the problem:

$$\theta = \tan^{-1}\left(\frac{1\,m}{0.01\,m}\right) = 89.4^o.$$

The magnetic field at point P is calculated by the Biot-Savart law (Equation 8.4.3):

$$\begin{aligned} B &= \frac{\mu_0}{4\pi} \frac{I \Delta l \sin \theta}{r^2} \\ &= (1 \times 10^{-7} T \cdot m/A) \left(\frac{2\,A(0.01\,m)\,sin\,(89.4^o)}{(1\,m)^2} \right) \\ &= 2.0 \times 10^{-9} T. \end{aligned}$$

From the right-hand rule and the Biot-Savart law, the field is directed into the page.

Significance

This approximation is only good if the length of the line segment is very small compared to the distance from the current element to the point. If not, the integral form of the Biot-Savart law must be used over the entire line segment to calculate the magnetic field.

Exercise 8.4.1

Using Example 8.4.1, at what distance would **P** have to be to measure a magnetic field half of the given answer?

Solution

1.41 meters

Example 8.4.1: Calculating Magnetic Field of a Circular Arc of Wire

A wire carries a current **I** in a circular arc with radius **R** swept through an arbitrary angle θ (Figure 8.4.3). Calculate the magnetic field at the center of this arc at point **P**.

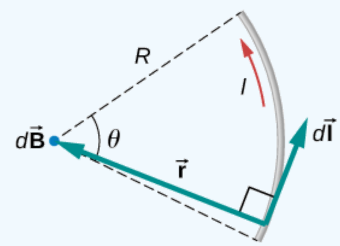


Figure 8.4.3: A wire segment carrying a current **I**. The path $d\vec{l}$ and radial direction $\hat{r}$ are indicated.

Strategy

We can determine the magnetic field at point **P** using the Biot-Savart law. The radial and path length directions are always at a right angle, so the cross product turns into multiplication. We also know that the distance along the path **dl** is related to the radius times the angle θ (in radians). Then we can pull all constants out of the integration and solve for the magnetic field.

Solution

The Biot-Savart law starts with the following equation:

$$\vec{B} = \frac{\mu_0}{4\pi} \int_{wire} \frac{I d\vec{l} \times \hat{r}}{r^2}.$$

As we integrate along the arc, all the contributions to the magnetic field are in the same direction (out of the page), so we can work with the magnitude of the field. The cross product turns into multiplication because the path dl and the radial direction are perpendicular. We can also substitute the arc length formula, $dl = r\,d\theta$:

$$B = \frac{\mu_0}{4\pi} \int_{wire} \frac{I r\,d\theta}{r^2}.$$

The current and radius can be pulled out of the integral because they are the same regardless of where we are on the path. This leaves only the integral over the angle,

$$B = \frac{\mu_0 I}{4\pi r} \int_{wire} d\theta.$$

The angle varies on the wire from 0 to θ; hence, the result is

$$B = \frac{\mu_0 I \theta}{4\pi r}.$$

Significance

The direction of the magnetic field at point P is determined by the right-hand rule, as shown in the previous chapter. If there are other wires in the diagram along with the arc, and you are asked to find the net magnetic field, find each contribution from a wire or arc and add the results by superposition of vectors. Make sure to pay attention to the direction of each contribution. Also note that in a symmetric situation, like a straight or circular wire, contributions from opposite sides of point P cancel each other.

Exercise 8.4.2

The wire loop forms a full circle of radius **R** and current **I**. What is the magnitude of the magnetic field at the center?

Solution

$\frac{\mu_0 I}{2R}$

8.5: Common Magnetic Field Models

Learning Objectives

- Explain the concept of continuous distribution of current elements and explain why it is used.
- List and apply magnetic field models for some common geometric distributions of current.

Magnetic Fields due to Current Distributions

By using the Biot-Savart Law, it is possible to calculate the magnetic field at a point in space for any distribution of current elements. In general, the mathematics of calculating the magnetic field can be challenging. In some special cases, we can construct the integrals, perform the integration analytically, and then write down closed-form formulas for the magnetic field. In other special cases of high symmetry, it is possible to use a technique involving Ampère's Law to make the calculations much easier. However, in general, it is often necessary to use computer simulations to calculate the magnetic field throughout the space surrounding an arbitrary current distribution.

In this chapter, we will summarize several models corresponding to the most common geometric distributions of current. In each model, the magnetic field can be written in a closed-formed expression for a specified region of space around the current distribution. Students who are interested in the details of the calculations should refer to the chapter Calculation of Magnetic Quantities from Currents. We will make use of these results in later chapters.

Magnetic Fields of Common Current Current Distributions

In the following results, it is assumed that the current is uniform throughout the current-carrying wire and that the current is steady in time. The width of the current-carrying wire is assumed to be small enough that the wire can be effectively treated as one-dimensional.

Magnetic Field due to a Thin Straight Wire

Figure $8.5.1$ shows a section of an infinitely long, straight wire that carries a current I. What is the magnetic field at a point P, located a distance R from the wire?

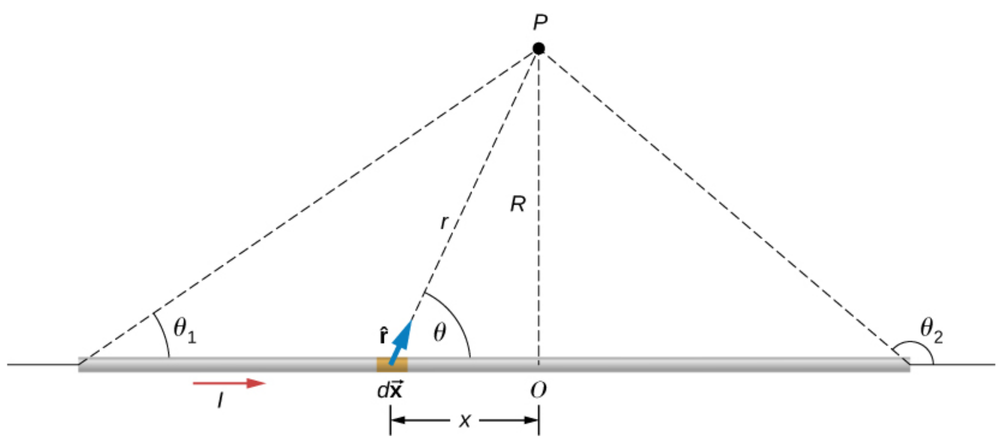


Figure $8.5.1$: A section of a thin, straight current-carrying wire. The independent variable θ has the limits θ_1 and θ_2.

Using the Biot-Savart to integrate along the length of the wire, the magnetic field strength at a distance R is given by

$$B=\frac{\mu_0 I}{2\pi R}.\quad \text{(infinite, thin, straight wire)} \tag{8.5.1}$$

The magnetic field lines of the infinite wire are circular and centered at the wire (Figure $8.5.2$), 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 form of the **right-hand rule** (Figure $8.5.2$). 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 $\vec{B}$.

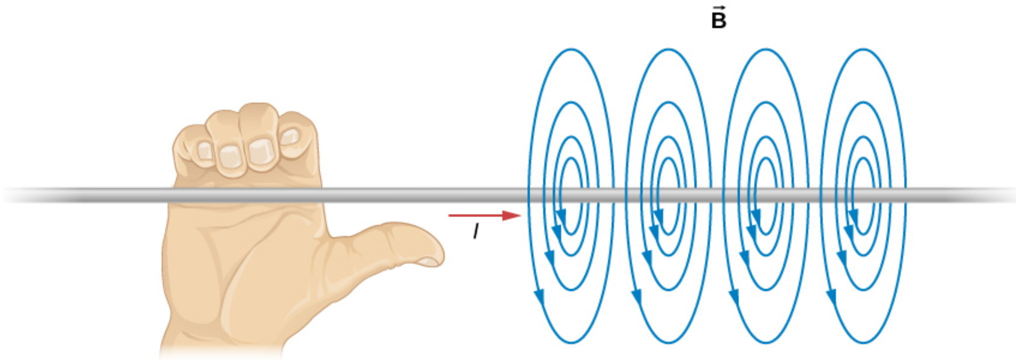


Figure 8.5.2:. Some magnetic field lines of an infinite wire. The direction of B can be found with a form of the right-hand rule.

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 $8.5.3a$. 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 $8.5.3b$.

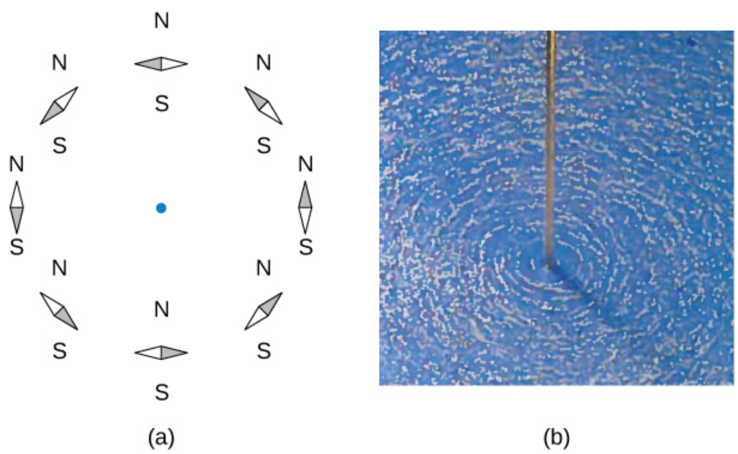


Figure 8.5.3: The shape of the magnetic field lines of a long wire can be seen using (a) small compass needles and (b) iron filings.

Magnetic Field due to a Current Loop

The circular loop of Figure $8.5.4$ has a radius R, carries a current I, and lies in the xz-plane. We will assume that the loop in an ideal circular loop. This will be a reasonable approximation for a real loop if the gap between the input and output wires is small compared to the radius of the loop. What is the magnetic field due to the current at an arbitrary point P along the axis of the loop?

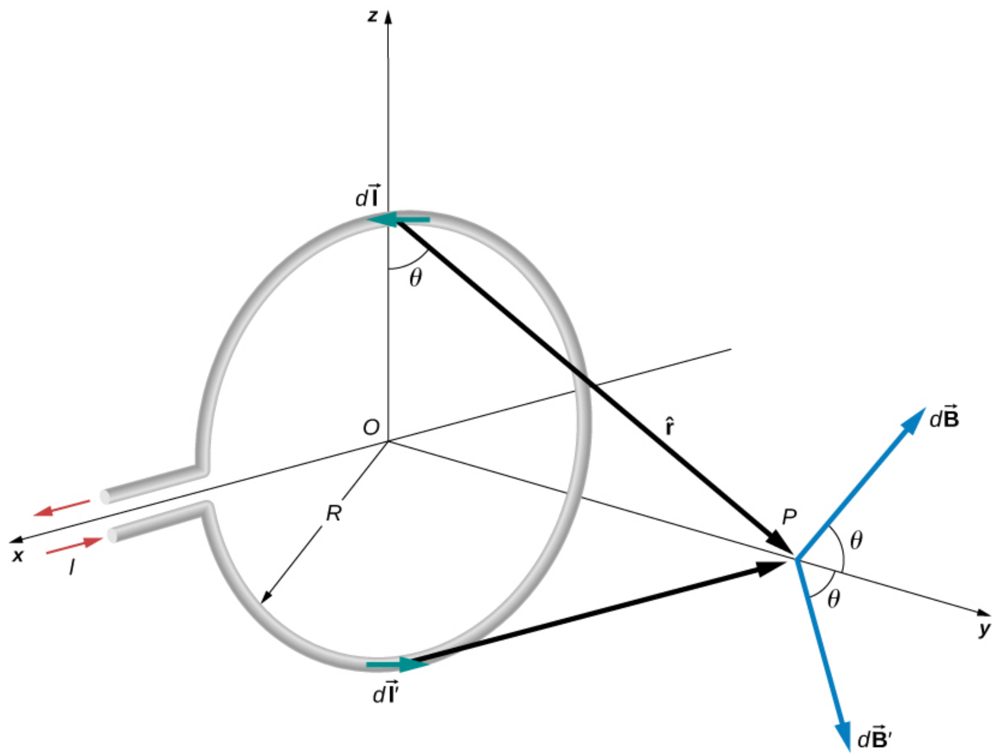


Figure 8.5.4: Determining the magnetic field at point P along the axis of a current-carrying loop of wire.

Using the Biot-Savart Law and integrating around the loop, the magnetic field along the y-axis at point P can shown to be

$$\vec{B} = \frac{\mu_0 R^2 I \hat{j}}{2(y^2 + R^2)^{3/2}}. \tag{8.5.2}$$

By setting $y = 0$ in Equation 8.5.2, we obtain the magnetic field at the center of the single loop:

$$\vec{B} = \frac{\mu_0 I}{2R}\hat{j} \text{ (single loop)}. \tag{8.5.3}$$

For a flat coil of N loops, this equation becomes

$$B = \frac{\mu_0 N I}{(2R)} \text{ (short coil of multiple loops)} \tag{8.5.4}$$

This latter equation is also approximately valid for short coils, where the length of the coil is small compared to the diameter of the coil.

If we consider $y >> R$ in Equation 8.5.2, the expression reduces to an expression known as the magnetic field from a dipole:

$$\vec{B} = \frac{\mu_0 R^2 I}{2y^3}. \tag{8.5.5}$$

The calculation of the magnetic field due to the circular current loop at points off-axis requires rather complex mathematics, so we will just look at the results. The magnetic field lines are shaped as shown in Figure 8.5.5. 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.

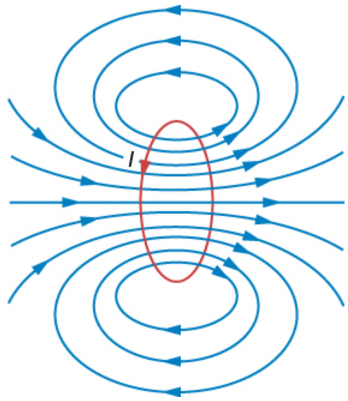


Figure 8.5.5: Sketch of the magnetic field lines of a circular current loop.

Magnetic Field due to a Solenoid

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 8.5.6 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.

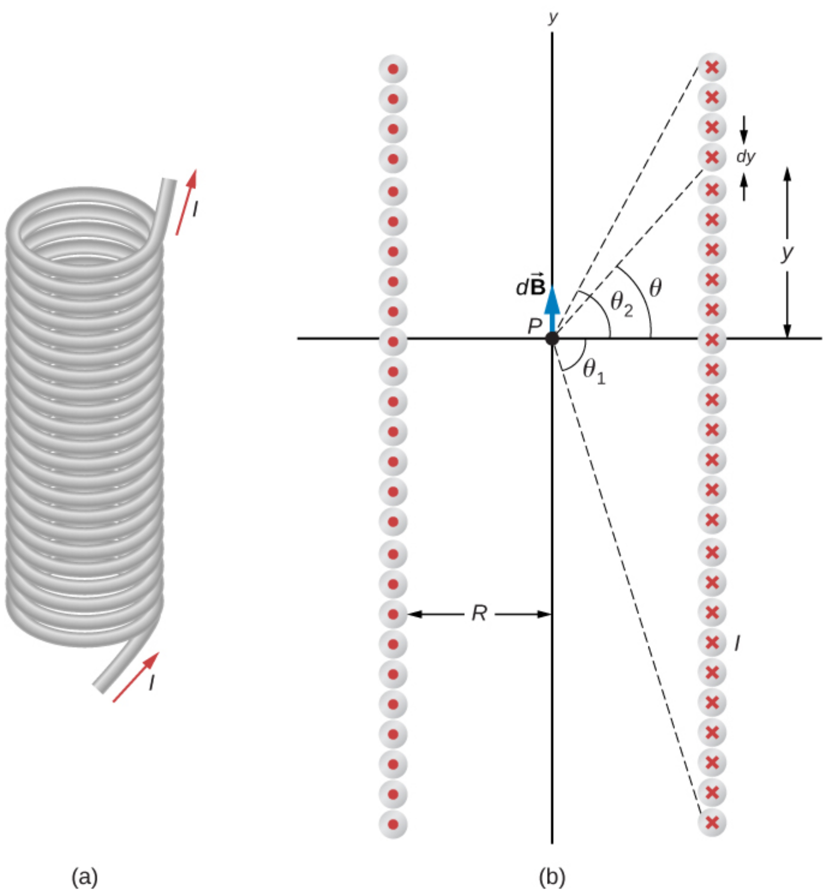


Figure 8.5.6: (a) A solenoid is a long wire wound in the shape of a helix. (b) The magnetic field at the point **P** on the axis of the solenoid is the net field due to all of the current loops.

Using the Biot-Savart Law and adding up the contributions of all the loops by integrating with respect to θ, we find that the magnetic field along the central axis of the finite solenoid is

$$\vec{B} = \frac{\mu_0 IN}{2L}(\sin\theta_2 - \sin\theta_1)\hat{j}. \quad \text{(finite solenoid)} \tag{8.5.6}$$

The infinitely long solenoid, for which $L \to \infty$ is of special interest and yields a particularly simple result. From a practical point of view, the infinite solenoid is one whose length is much larger than its radius $(L >> R)$. In this case, $\theta_1 = -\pi/2$ and $\theta_2 = \pi/2$. Then from Equation 8.5.6, the magnetic field along the central axis of an infinite solenoid is

$$\vec{B} = \frac{\mu_0 IN}{L}\hat{j} = \mu_0 nI\hat{j}, \ \text{(infinite solenoid)} \tag{8.5.7}$$

where $n = N/L$ 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.

Example 8.5.1: 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 8.5.7. Outside the solenoid, the magnetic field is zero.

Solution

The number of turns per unit length is

$$n = \frac{300 \text{ turns}}{0.140 \text{ m}} = 2.14 \times 10^3 \text{ turns/m}. \tag{8.5.8}$$

The magnetic field produced inside the solenoid is

$$B = \mu_0 nI = (4\pi \times 10^{-7} \text{ T} \cdot \text{m/A})(2.14 \times 10^3 \text{ turns/m})(0.410 \text{ A}) \tag{8.5.9}$$

$$B = 1.10 \times 10^{-3} \text{ T}. \tag{8.5.10}$$

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 8.5.1

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

Magnetic Field due to Toroid

A **toroid** is a donut-shaped coil closely wound with one continuous wire, as illustrated in part (a) of Figure 8.5.7. 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?

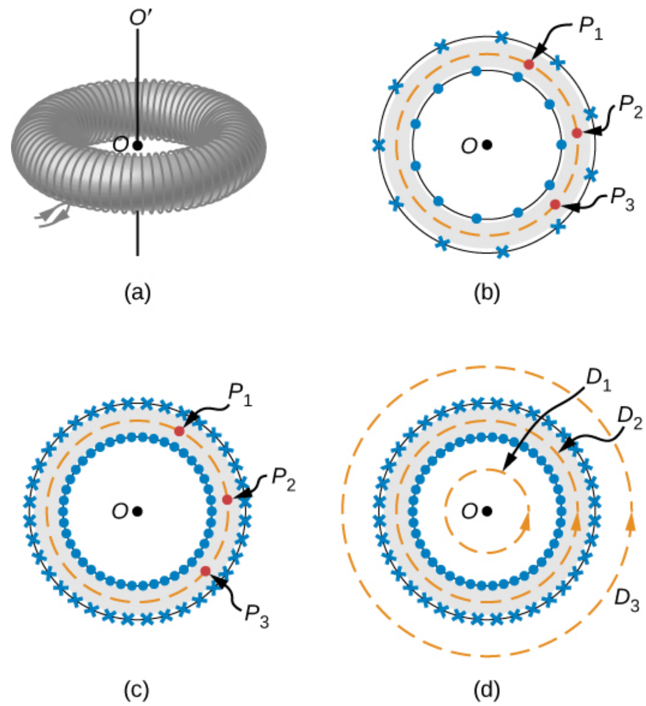


Figure 8.5.7: (a) A toroid is a coil wound into a donut-shaped object. (b) A loosely wound toroid does not have cylindrical symmetry. (c) In a tightly wound toroid, cylindrical symmetry is a very good approximation. (d) Several paths of integration for Ampère’s law.

Assuming cylindrical symmetry around the axis OO', Ampère’s law can be used to show that

$$B=\frac{\mu_0 NI}{2\pi r} \text{ (within the toroid)}. \tag{8.5.11}$$

Actually, this assumption of circular symmetry is not precisely correct, for as part (b) of Figure 8.5.4 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 8.5.4], and cylindrical symmetry is an accurate approximation.

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 OO' However, if the central radius R (the radius midway between the inner and outer radii of the toroid) is much larger than the cross-sectional diameter of the coils r, the variation is fairly small, and the magnitude of the magnetic field may be calculated by Equation 8.5.11 where $r=R$.

8.6: Magnetic Fields and Lines

Learning Objectives

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

- Define the magnetic field based on a moving charge experiencing a force
- Apply the right-hand rule to determine the direction of a magnetic force based on the motion of a charge in a magnetic field
- Sketch magnetic field lines to understand which way the magnetic field points and how strong it is in a region of space

We have outlined the properties of magnets, described how they behave, and listed some of the applications of magnetic properties. Even though there are no such things as isolated magnetic charges, we can still define the attraction and repulsion of magnets as based on a field. In this section, we define the magnetic field, determine its direction based on the right-hand rule, and discuss how to draw magnetic field lines.

Defining the Magnetic Field

A magnetic field is defined by the force that a charged particle experiences moving in this field, after we account for the gravitational and any additional electric forces possible on the charge. The magnitude of this force is proportional to the amount of charge **q**, the speed of the charged particle **v**, and the magnitude of the applied magnetic field. The direction of this force is perpendicular to both the direction of the moving charged particle and the direction of the applied magnetic field. Based on these observations, we define the magnetic field strength **B** based on the **magnetic force** $\vec{F}$ on a charge **q** moving at velocity as the cross product of the velocity and magnetic field, that is,

$$\vec{F} = q\vec{v} \times \vec{B}. \tag{8.6.1}$$

In fact, this is how we define the magnetic field $\vec{B}$ - in terms of the force on a charged particle moving in a magnetic field. The magnitude of the force is determined from the definition of the cross product as it relates to the magnitudes of each of the vectors. In other words, the magnitude of the force satisfies

$$F = qvB\sin\theta \tag{8.6.2}$$

where **θ** is the angle between the velocity and the magnetic field.

The SI unit for magnetic field strength B is called the tesla (T) after the eccentric, but brilliant inventor Nikola Tesla (1856–1943), where

$$1\,T = \frac{1\,N}{A\cdot m}. \tag{8.6.3}$$

A smaller unit, called the **gauss** (G) is sometimes used, where

$$1\,G = 10^{-4}\,T \tag{8.6.4}$$

The strongest permanent magnets have fields near 2 T; superconducting electromagnets may attain 10 T or more. Earth's magnetic field on its surface is only about $5\times 10^{-5}\,T$ or $0.5\,G$.

Problem-Solving Strategy: Direction of the Magnetic Field by the Right-Hand Rule

The direction of the magnetic force $\vec{F}$ is perpendicular to the plane formed by $\vec{v}$ and $\vec{B}$ as determined by the **right-hand rule-1** (or RHR-1), which is illustrated in Figure $8.6.1$.

1. Orient your right hand so that your fingers curl in the plane defined by the velocity and magnetic field vectors.
2. Using your right hand, sweep from the velocity toward the magnetic field with your fingers through the smallest angle possible.
3. The magnetic force is directed where your thumb is pointing.
4. If the charge was negative, reverse the direction found by these steps.

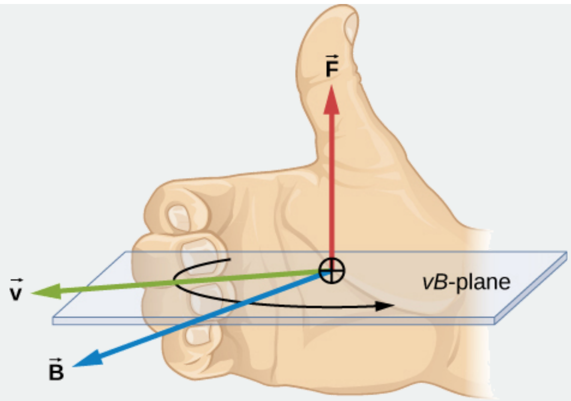


Figure 8.6.1: Magnetic fields exert forces on moving charges. The direction of the magnetic force on a moving charge is perpendicular to the plane formed by b$\vec{v}$ and $\vec{B}$ and follows the right-hand rule-1 (RHR-1) as shown. The magnitude of the force is proportional to q, v, B, and the sine of the angle between $\vec{v}$ and $\vec{B}$.

Note

Visit this website for additional practice with the direction of magnetic fields.

There is no magnetic force on static charges. However, there is a magnetic force on charges moving at an angle to a magnetic field. When charges are stationary, their electric fields do not affect magnets. However, when charges move, they produce magnetic fields that exert forces on other magnets. When there is relative motion, a connection between electric and magnetic forces emerges - each affects the other.

Example 8.6.1: An Alpha-Particle Moving in a Magnetic Field

An alpha-particle $(q = 3.2 \times 10^{-19} C)$ moves through a uniform magnetic field whose magnitude is 1.5 T. The field is directly parallel to the positive **z**-axis of the rectangular coordinate system of Figure 8.6.2. What is the magnetic force on the alpha-particle when it is moving (a) in the positive **x**-direction with a speed of $5.0 \times 10^4 m/s$? (b) in the negative **y**-direction with a speed of $5.0 \times 10^4 m/s$? (c) in the positive **z**-direction with a speed of $5.0 \times 10^4 m/s$? (d) with a velocity $\vec{v} = \left(2.0\hat{i} - 3.0\hat{j} + 1.0\hat{k}\right) \times 10^4 m/s$?

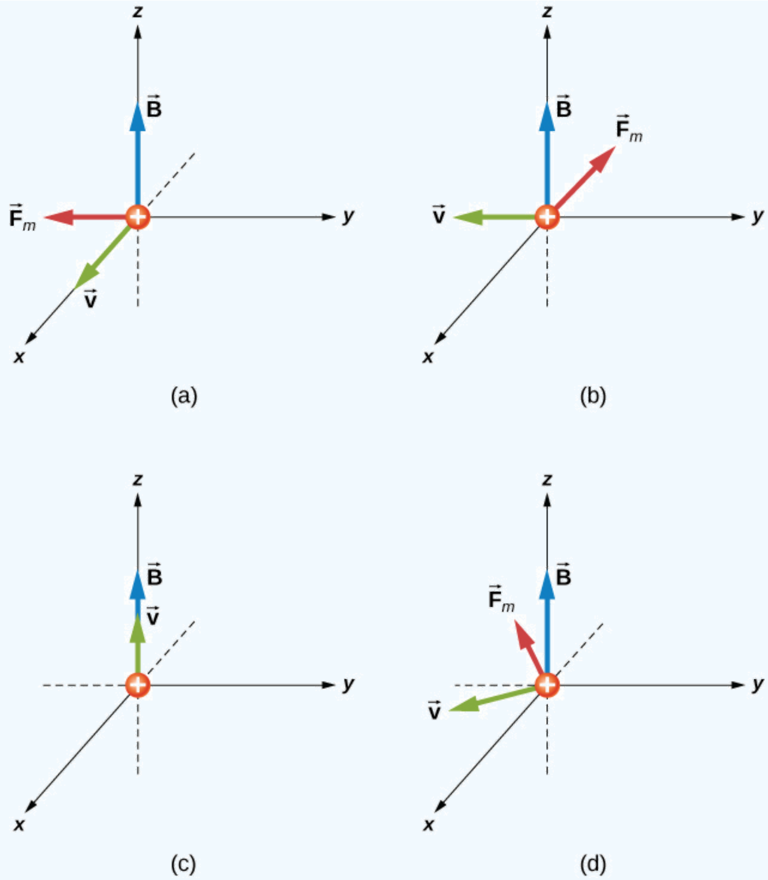


Figure 8.6.2: The magnetic forces on an alpha-particle moving in a uniform magnetic field. The field is the same in each drawing, but the velocity is different.

Strategy

We are given the charge, its velocity, and the magnetic field strength and direction. We can thus use the equation $\vec{F} = q\vec{v} \times \vec{B}$ or $F = qvBsin\,\theta$ to calculate the force. The direction of the force is determined by RHR-1.

Solution

1. First, to determine the direction, start with your fingers pointing in the positive **x**-direction. Sweep your fingers upward in the direction of magnetic field. Your thumb should point in the negative **y**-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:

$$\vec{F} = q\vec{v} \times \vec{B} = (3.2 \times 10^{-19} C)(5.0 \times 10^4 m/s\,\hat{i}) \times (1.5\,T\,\hat{k}) = -2.4 \times 10^{-14} N\,\hat{j} \tag{8.6.5}$$

2. First, to determine the directionality, start with your fingers pointing in the negative **y**-direction. Sweep your fingers upward in the direction of magnetic field as in the previous problem. Your thumb should be open in the negative **x**-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:

$$\vec{F} = q\vec{v} \times \vec{B} = (3.2 \times 10^{-19} C)(-5.0 \times 10^4 m/s\,\hat{i}) \times (1.5\,T\,\hat{k}) = -2.4 \times 10^{-14} N\,\hat{i} \tag{8.6.6}$$

   An alternative approach is to use Equation 8.6.2 to find the magnitude of the force. This applies for both parts (a) and (b). Since the velocity is perpendicular to the magnetic field, the angle between them is 90 degrees. Therefore, the magnitude of the force is:

$$F = qvB\sin\,\theta = (3.2 \times 10^{-19} C)(5.0 \times 10^4 m/s)(1.5\,T)sin(90^o)) = 2.4 \times 10^{-14} N. \tag{8.6.7}$$

3. Since the velocity and magnetic field are parallel to each other, there is no orientation of your hand that will result in a force direction. Therefore, the force on this moving charge is zero. This is confirmed by the cross product. When you cross two vectors pointing in the same direction, the result is equal to zero.
4. First, to determine the direction, your fingers could point in any orientation; however, you must sweep your fingers upward in the direction of the magnetic field. As you rotate your hand, notice that the thumb can point in any **x**- or **y**-direction possible, but not in the **z**-direction. This should match the mathematical answer. To calculate the force, we use the given charge, velocity, and magnetic field and the definition of the magnetic force in cross-product form to calculate:

$$\vec{F} = q\vec{v} \times \vec{B} = (3.2 \times 10^{-19} C)((2.0\hat{i} - 3.0\hat{j} + 1.0\hat{k}) \times 10^4 m/s) \times (1.5\, T\hat{k}) \tag{8.6.8}$$

$$(-14.4\hat{i} - 9.6\hat{j}) \times 10^{-15} N. \tag{8.6.9}$$

This solution can be rewritten in terms of a magnitude and angle in the **xy**-plane:

$$|\vec{F}| = \sqrt{F_x^2 + F_y^2} = \sqrt{(-14.4)^2 + (-9.6)^2} \times 10^{-15} N = 1.7 \times 10^{-14} N \tag{8.6.10}$$

$$\theta = tan^{-1}\left(\frac{F_y}{F_x}\right) = tan^{-1}\left(\frac{-9.6 \times 10^{-15} N}{-14.4 \times 10^{-15} N}\right) = 34^o. \tag{8.6.11}$$

The magnitude of the force can also be calculated using Equation 8.6.2. The velocity in this question, however, has three components. The **z**-component of the velocity can be neglected, because it is parallel to the magnetic field and therefore generates no force. The magnitude of the velocity is calculated from the **x**- and **y**-components. The angle between the velocity in the **xy**-plane and the magnetic field in the **z**-plane is 90 degrees. Therefore, the force is calculated to be:

$$|\vec{v}| = \sqrt{(2)^2 + (-3)^2} \times 10^4 \frac{m}{s} = 3.6 \times 10^4 \frac{m}{s} \tag{8.6.12}$$

$$F = qv\,Bsin\,\theta = (3.2 \times 10^{-19} C)(3.6 \times 10^4 m/s)(1.5\,T)sin(90^o) = 1.7 \times 10^{-14} N \tag{8.6.13}$$

This is the same magnitude of force calculated by unit vectors.

Significance

The cross product in this formula results in a third vector that must be perpendicular to the other two. Other physical quantities, such as angular momentum, also have three vectors that are related by the cross product. Note that typical force values in magnetic force problems are much larger than the gravitational force. Therefore, for an isolated charge, the magnetic force is the dominant force governing the charge's motion.

? Exercise 8.6.1

Repeat the previous problem with the magnetic field in the **x**-direction rather than in the **z**-direction. Check your answers with RHR-1.

Answer a

0 N

Answer b

$2.4 \times 10^{-14} \hat{k} N$

Answer c

$2.4 \times 10^{-14} \hat{j} N$

Answer d

$7.2\hat{j} + 2.2\hat{k}) \times 10^{-15} N$

Representing Magnetic Fields

The representation of magnetic fields by **magnetic field lines** is very useful in visualizing the strength and direction of the magnetic field. As shown in Figure $8.6.3$, each of these lines forms a closed loop, even if not shown by the constraints of the space available for the figure. The field lines emerge from the north pole (N), loop around to the south pole (S), and continue through the bar magnet back to the north pole.

Magnetic field lines have several hard-and-fast rules:

1. The direction of the magnetic field is tangent to the field line at any point in space. A small compass will point in the direction of the field line.
2. The strength of the field is proportional to the closeness of the lines. It is exactly proportional to the number of lines per unit area perpendicular to the lines (called the areal density).
3. Magnetic field lines can never cross, meaning that the field is unique at any point in space.
4. Magnetic field lines are continuous, forming closed loops without a beginning or end. They are directed from the north pole to the south pole.

The last property is related to the fact that the north and south poles cannot be separated. It is a distinct difference from electric field lines, which generally begin on positive charges and end on negative charges or at infinity. If isolated magnetic charges (referred to as **magnetic monopoles**) existed, then magnetic field lines would begin and end on them.

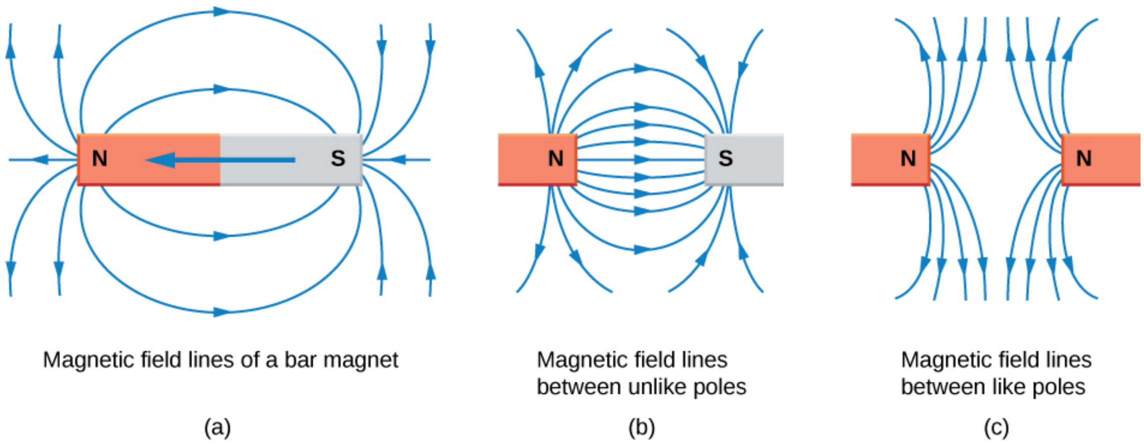


Figure $8.6.3$: Magnetic field lines are defined to have the direction in which a small compass points when placed at a location in the field. The strength of the field is proportional to the closeness (or density) of the lines. If the interior of the magnet could be probed, the field lines would be found to form continuous, closed loops. To fit in a reasonable space, some of these drawings may not show the closing of the loops; however, if enough space were provided, the loops would be closed.

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).

8.7: Motion of a Charged Particle in a Magnetic Field

Learning Objectives

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

- Explain how a charged particle in an external magnetic field undergoes circular motion
- Describe how to determine the radius of the circular motion of a charged particle in a magnetic field

A charged particle experiences a force when moving through a magnetic field. What happens if this field is uniform over the motion of the charged particle? What path does the particle follow? In this section, we discuss the circular motion of the charged particle as well as other motion that results from a charged particle entering a magnetic field.

The simplest case occurs when a charged particle moves perpendicular to a uniform **B**-field (Figure $8.7.1$). If the field is in a vacuum, the magnetic field is the dominant factor determining the motion. Since the magnetic force is perpendicular to the direction of travel, a charged particle follows a curved path in a magnetic field. The particle continues to follow this curved path until it forms a complete circle. Another way to look at this is that the magnetic force is always perpendicular to velocity, so that it does no work on the charged particle. The particle's kinetic energy and speed thus remain constant. The direction of motion is affected but not the speed.

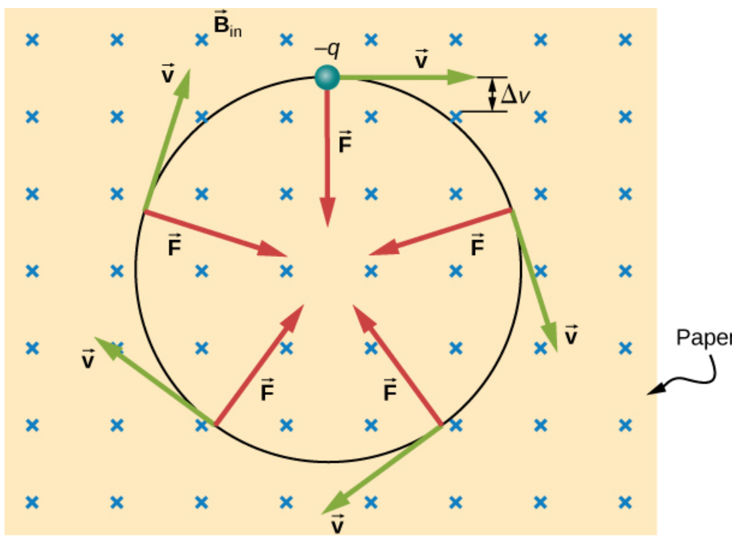


Figure $8.7.1$: A negatively charged particle moves in the plane of the paper in a region where the magnetic field is perpendicular to the paper (represented by the small X's - like the tails of arrows). The magnetic force is perpendicular to the velocity, so velocity changes in direction but not magnitude. The result is uniform circular motion. (Note that because the charge is negative, the force is opposite in direction to the prediction of the right-hand rule.)

In this situation, the magnetic force supplies the centripetal force $F_C = \frac{mv^2}{r}$. Noting that the velocity is perpendicular to the magnetic field, the magnitude of the magnetic force is reduced to $F = qvB$. Because the magnetic force **F** supplies the centripetal force F_C, we have

$$qvB = \frac{mv^2}{r}.$$

Solving for **r** yields

$$r = \frac{mv}{qB}. \tag{8.7.1}$$

Here, **r** is the radius of curvature of the path of a charged particle with mass **m** and charge **q**, moving at a speed **v** that is perpendicular to a magnetic field of strength **B**. The time for the charged particle to go around the circular path is defined as the period, which is the same as the distance traveled (the circumference) divided by the speed. Based on this and Equation, we can derive the period of motion as

$$T=\frac{2\pi r}{v}=\frac{2\pi}{v}\frac{mv}{qB}=\frac{2\pi m}{qB}. \tag{8.7.2}$$

If the velocity is not perpendicular to the magnetic field, then we can compare each component of the velocity separately with the magnetic field. The component of the velocity perpendicular to the magnetic field produces a magnetic force perpendicular to both this velocity and the field:

$$v_{perp}=v\sin\theta \tag{8.7.3}$$

$$v_{para}=v\cos\theta. \tag{8.7.4}$$

where θ is the angle between **v** and **B**. The component parallel to the magnetic field creates constant motion along the same direction as the magnetic field, also shown in Equation. The parallel motion determines the **pitch p** of the helix, which is the distance between adjacent turns. This distance equals the parallel component of the velocity times the period:

$$p=v_{para}T. \tag{8.7.5}$$

The result is a **helical motion**, as shown in the following figure.

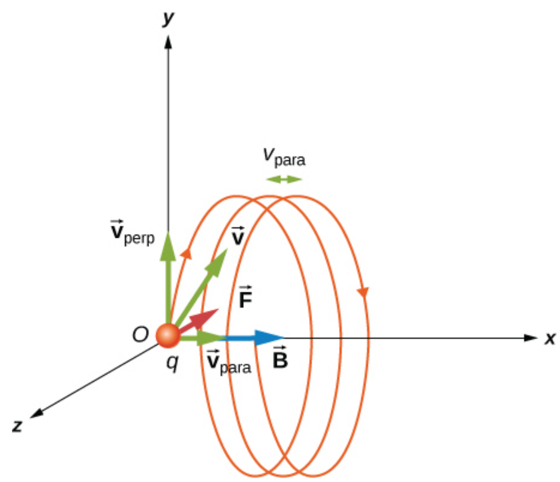


Figure 8.7.2: A charged particle moving with a velocity not in the same direction as the magnetic field. The velocity component perpendicular to the magnetic field creates circular motion, whereas the component of the velocity parallel to the field moves the particle along a straight line. The pitch is the horizontal distance between two consecutive circles. The resulting motion is helical.

While the charged particle travels in a helical path, it may enter a region where the magnetic field is not uniform. In particular, suppose a particle travels from a region of strong magnetic field to a region of weaker field, then back to a region of stronger field. The particle may reflect back before entering the stronger magnetic field region. This is similar to a wave on a string traveling from a very light, thin string to a hard wall and reflecting backward. If the reflection happens at both ends, the particle is trapped in a so-called magnetic bottle.

Trapped particles in magnetic fields are found in the Van Allen radiation belts around Earth, which are part of Earth's magnetic field. These belts were discovered by James Van Allen while trying to measure the flux of cosmic rays on Earth (high-energy particles that come from outside the solar system) to see whether this was similar to the flux measured on Earth. Van Allen found that due to the contribution of particles trapped in Earth's magnetic field, the flux was much higher on Earth than in outer space. Aurorae, like the famous aurora borealis (northern lights) in the Northern Hemisphere (Figure 8.7.3), are beautiful displays of light emitted as ions recombine with electrons entering the atmosphere as they spiral along magnetic field lines. (The ions are primarily oxygen and nitrogen atoms that are initially ionized by collisions with energetic particles in Earth's atmosphere.) Aurorae have also been observed on other planets, such as Jupiter and Saturn.

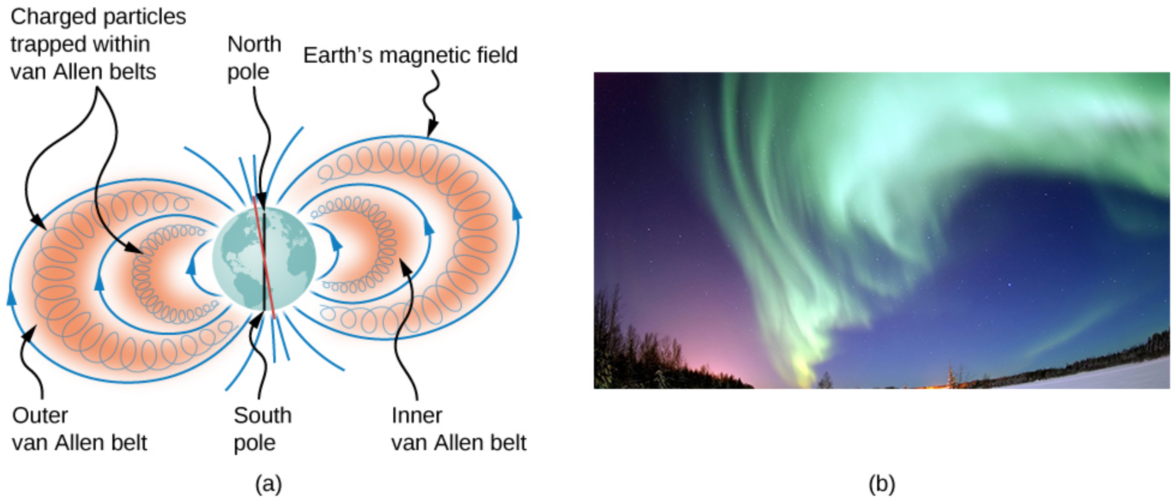


Figure 8.7.3: (a) The Van Allen radiation belts around Earth trap ions produced by cosmic rays striking Earth's atmosphere. (b) The magnificent spectacle of the aurora borealis, or northern lights, glows in the northern sky above Bear Lake near Eielson Air Force Base, Alaska. Shaped by Earth's magnetic field, this light is produced by glowing molecules and ions of oxygen and nitrogen. (credit b: modification of work by USAF Senior Airman Joshua Strang)

✓ Example 8.7.1: Beam Deflector

A research group is investigating short-lived radioactive isotopes. They need to design a way to transport alpha-particles (helium nuclei) from where they are made to a place where they will collide with another material to form an isotope. The beam of alpha-particles $(m = 6.64 \times 10^{-27} kg,\ q = 3.2 \times 10^{-19} C)$ bends through a 90-degree region with a uniform magnetic field of 0.050 T (Figure 8.7.4). (a) In what direction should the magnetic field be applied? (b) How much time does it take the alpha-particles to traverse the uniform magnetic field region?

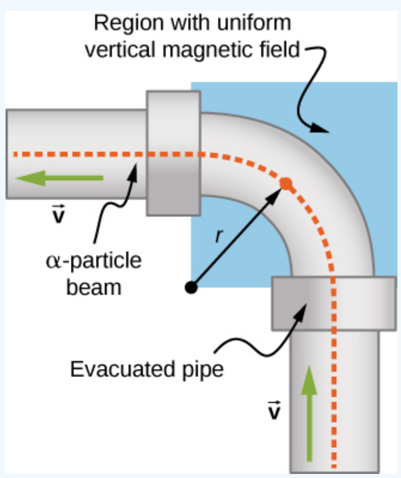


Figure 8.7.4: Top view of the beam deflector setup.

Strategy

1. The direction of the magnetic field is shown by the RHR-1. Your fingers point in the direction of **v**, and your thumb needs to point in the direction of the force, to the left. Therefore, since the alpha-particles are positively charged, the magnetic field must point down.
2. The period of the alpha-particle going around the circle is

$$T = \frac{2\pi m}{qB}.$$

Because the particle is only going around a quarter of a circle, we can take 0.25 times the period to find the time it takes to go around this path.

Solution

1. Let's start by focusing on the alpha-particle entering the field near the bottom of the picture. First, point your thumb up the page. In order for your palm to open to the left where the centripetal force (and hence the magnetic force) points, your fingers need to change orientation until they point into the page. This is the direction of the applied magnetic field.
2. The period of the charged particle going around a circle is calculated by using the given mass, charge, and magnetic field in the problem. This works out to be

$$T=\frac{2\pi m}{qB}=\frac{2\pi(6.64\times 10^{-27}kg)}{(3.2\times 10^{-19}C)(0.050\,T)}=2.6\times 10^{-6}s.$$

However, for the given problem, the alpha-particle goes around a quarter of the circle, so the time it takes would be

$$t=0.25\times 2.61\times 10^{-6}s=6.5\times 10^{-7}s.$$

Significance

This time may be quick enough to get to the material we would like to bombard, depending on how short-lived the radioactive isotope is and continues to emit alpha-particles. If we could increase the magnetic field applied in the region, this would shorten the time even more. The path the particles need to take could be shortened, but this may not be economical given the experimental setup.

? Exercise 8.7.1

A uniform magnetic field of magnitude 1.5 T is directed horizontally from west to east. (a) What is the magnetic force on a proton at the instant when it is moving vertically downward in the field with a speed of $4\times 10^7\ m/s$? (b) Compare this force with the weight **w** of a proton.

Solution

a. $9.6\times 10^{-12}N$ toward the south;

b. $\frac{w}{F_m}=1.7\times 10^{-15}$

✓ Example 8.7.2: Helical Motion in a Magnetic Field

A proton enters a uniform magnetic field of $1.0\times 10^{-4}T$ with a speed of $5\times 10^5\ m/s$. At what angle must the magnetic field be from the velocity so that the pitch of the resulting helical motion is equal to the radius of the helix?

Strategy

The pitch of the motion relates to the parallel velocity times the period of the circular motion, whereas the radius relates to the perpendicular velocity component. After setting the radius and the pitch equal to each other, solve for the angle between the magnetic field and velocity or θ.

Solution

The pitch is given by Equation 8.7.5, the period is given by Equation 8.7.2, and the radius of circular motion is given by Equation 8.7.1. Note that the velocity in the radius equation is related to only the perpendicular velocity, which is where the circular motion occurs. Therefore, we substitute the sine component of the overall velocity into the radius equation to equate the pitch and radius

$$p=r$$

$$v_{\parallel}T=\frac{mv}{qB}$$

$$v\cos\theta\frac{2\pi m}{qB}=\frac{mv\sin\theta}{qB}$$

$$2\pi=\tan\theta$$

$$\theta=81.0^o.$$

Significance

If this angle were 0^o, only parallel velocity would occur and the helix would not form, because there would be no circular motion in the perpendicular plane. If this angle were 90^o only circular motion would occur and there would be no movement of the circles perpendicular to the motion. That is what creates the helical motion.

8.8: Magnetic Force on a Current-Carrying Conductor

Learning Objectives

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

- Determine the direction in which a current-carrying wire experiences a force in an external magnetic field
- Calculate the force on a current-carrying wire in an external magnetic field

Moving charges experience a force in a magnetic field. If these moving charges are in a wire—that is, if the wire is carrying a current—the wire should also experience a force. However, before we discuss the force exerted on a current by a magnetic field, we first examine the magnetic field generated by an electric current. We are studying two separate effects here that interact closely: A current-carrying wire generates a magnetic field and the magnetic field exerts a force on the current-carrying wire.

Magnetic Fields Produced by Electrical Currents

When discussing historical discoveries in magnetism, we mentioned Oersted's finding that a wire carrying an electrical current caused a nearby compass to deflect. A connection was established that electrical currents produce magnetic fields. (This connection between electricity and magnetism is discussed in more detail in Sources of Magnetic Fields.)

The compass needle near the wire experiences a force that aligns the needle tangent to a circle around the wire. Therefore, a current-carrying wire produces circular loops of magnetic field. To determine the direction of the magnetic field generated from a wire, we use a second right-hand rule. In RHR-2, your thumb points in the direction of the current while your fingers wrap around the wire, pointing in the direction of the magnetic field produced (Figure $8.8.1$). If the magnetic field were coming at you or out of the page, we represent this with a dot. If the magnetic field were going into the page, we represent this with an

These symbols come from considering a vector arrow: An arrow pointed toward you, from your perspective, would look like a dot or the tip of an arrow. An arrow pointed away from you, from your perspective, would look like a cross or an A composite sketch of the magnetic circles is shown in Figure $8.8.1$, where the field strength is shown to decrease as you get farther from the wire by loops that are farther separated.

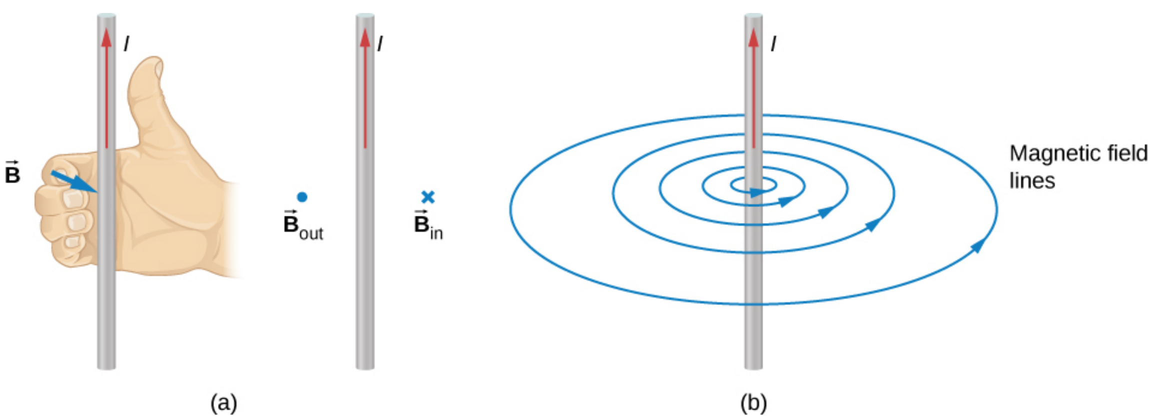


Figure $8.8.1$: (a) When the wire is in the plane of the paper, the field is perpendicular to the paper. Note the symbols used for the field pointing inward (like the tail of an arrow) and the field pointing outward (like the tip of an arrow). (b) A long and straight wire creates a field with magnetic field lines forming circular loops.

Calculating the Magnetic Force

Electric current is an ordered movement of charge. A current-carrying wire in a magnetic field must therefore experience a force due to the field. To investigate this force, let's consider the infinitesimal section of wire as shown in Figure $8.8.3$. The length and cross-sectional area of the section are **dl** and **A**, respectively, so its volume is $V = A\cdot dl$. The wire is formed from material that contains **n** charge carriers per unit volume, so the number of charge carriers in the section is $nA\cdot dl$. If the charge carriers move with drift velocity $\vec{v}_d$ the current **I** in the wire is (from Current and Resistance)

$$I = neAv_d. \tag{8.8.1}$$

The magnetic force on any single charge carrier is $e\vec{v}_d \times \vec{B}$, so the total magnetic force $d\vec{F}$ on the $nA \cdot dl$ charge carriers in the section of wire is

$$d\vec{F} = (nA \cdot dl)e\vec{v}_d \times \vec{B}. \tag{8.8.2}$$

We can define **dl** to be a vector of length **dl** pointing along $\vec{v}_d$, which allows us to rewrite this equation as

$$d\vec{F} = neAv_d d\vec{l} \times \vec{B}, \tag{8.8.3}$$

or

$$d\vec{F} = Id\vec{l} \times \vec{B}. \tag{8.8.4}$$

This is the magnetic force on the section of wire. Note that it is actually the net force exerted by the field on the charge carriers themselves. The direction of this force is given by RHR-1, where you point your fingers in the direction of the current and curl them toward the field. Your thumb then points in the direction of the force.

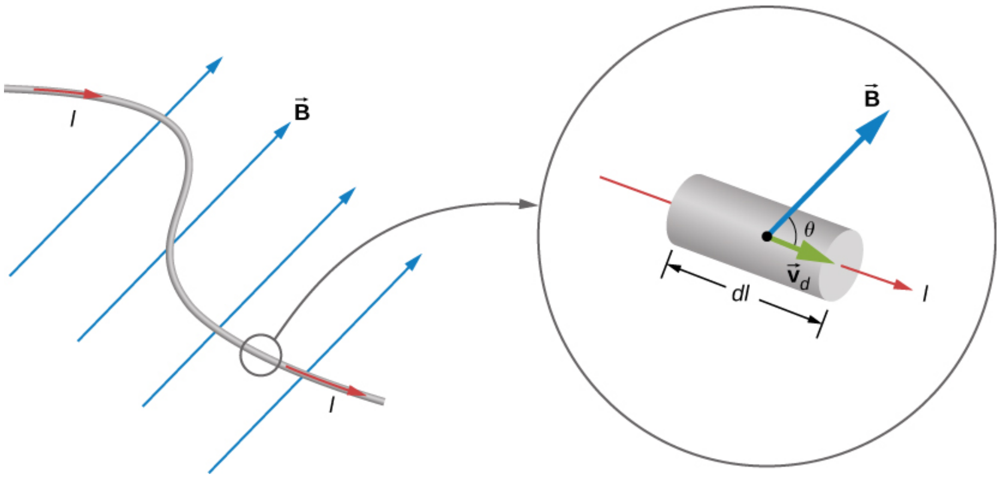


Figure 8.8.2: An infinitesimal section of current-carrying wire in a magnetic field.

To determine the magnetic force $\vec{F}$ on a wire of arbitrary length and shape, we must integrate Equation 8.8.4 over the entire wire. If the wire section happens to be straight and **B** is uniform, the equation differentials become absolute quantities, giving us

$$\vec{F} = I\vec{l} \times \vec{B}. \tag{8.8.5}$$

This is the force on a straight, current-carrying wire in a uniform magnetic field.

✓ Example 8.8.1: Balancing the Gravitational and Magnetic Forces on a Current-Carrying Wire

A wire of length 50 cm and mass 10 g is suspended in a horizontal plane by a pair of flexible leads (Figure 8.8.3). The wire is then subjected to a constant magnetic field of magnitude 0.50 T, which is directed as shown. What are the magnitude and direction of the current in the wire needed to remove the tension in the supporting leads?

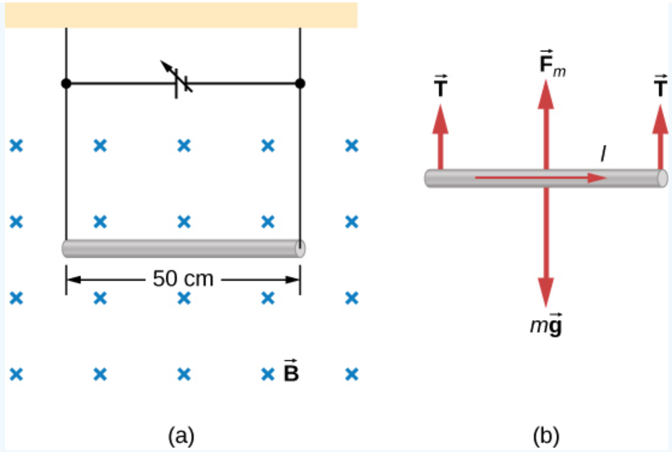


Figure 8.8.3: (a) A wire suspended in a magnetic field. (b) The free-body diagram for the wire.

Strategy

From the free-body diagram in the figure, the tensions in the supporting leads go to zero when the gravitational and magnetic forces balance each other. Using the RHR-1, we find that the magnetic force points up. We can then determine the current **I** by equating the two forces.

Solution

Equate the two forces of weight and magnetic force on the wire:

$$mg = IlB. \tag{8.8.6}$$

Thus,

$$I = \frac{mg}{lB} = \frac{(0.010\,kg}{9.8\,m/s^2)}(0.50\,m)(0.50\,T) = 0.39\,A. \tag{8.8.7}$$

Significance

This large magnetic field creates a significant force on a length of wire to counteract the weight of the wire.

Example 8.8.2: Calculating Magnetic Force on a Current-Carrying Wire

A long, rigid wire lying along the **y**-axis carries a 5.0-A current flowing in the positive **y**-direction. (a) If a constant magnetic field of magnitude 0.30 T is directed along the positive **x**-axis, what is the magnetic force per unit length on the wire? (b) If a constant magnetic field of 0.30 T is directed 30 degrees from the +**x**-axis towards the +**y**-axis, what is the magnetic force per unit length on the wire?

Strategy

The magnetic force on a current-carrying wire in a magnetic field is given by $\vec{F} = I\vec{l} \times \vec{B}$. For part a, since the current and magnetic field are perpendicular in this problem, we can simplify the formula to give us the magnitude and find the direction through the RHR-1. The angle **θ** is 90 degrees, which means $sin\,\theta = 1$. Also, the length can be divided over to the left-hand side to find the force per unit length. For part b, the current times length is written in unit vector notation, as well as the magnetic field. After the cross product is taken, the directionality is evident by the resulting unit vector.

Solution

1. We start with the general formula for the magnetic force on a wire. We are looking for the force per unit length, so we divide by the length to bring it to the left-hand side. We also set $sin\,\theta$. The solution therefore is

$$F = IlB\,sin\,\theta \tag{8.8.8}$$

$$\frac{F}{l} = (5.0\,A)(0.30\,T) \tag{8.8.9}$$

$$\frac{F}{l} = 1.5\,N/m. \tag{8.8.10}$$

Directionality: Point your fingers in the positive **y**-direction and curl your fingers in the positive **x**-direction. Your thumb will point in the $-\vec{k}$ direction. Therefore, with directionality, the solution is

$$\frac{\vec{F}}{l} = -1.5\vec{k}\,N/m. \tag{8.8.11}$$

2. The current times length and the magnetic field are written in unit vector notation. Then, we take the cross product to find the force:

$$\vec{F} = I\vec{l} \times \vec{B} = (5.0A)l\hat{j} \times (0.30T\,cos(30^o)\hat{i} \tag{8.8.12}$$

$$\vec{F}/l = -1.30\hat{k}\,N/m. \tag{8.8.13}$$

Significance

This large magnetic field creates a significant force on a small length of wire. As the angle of the magnetic field becomes more closely aligned to the current in the wire, there is less of a force on it, as seen from comparing parts a and b.

? Exercise 8.8.1

A straight, flexible length of copper wire is immersed in a magnetic field that is directed into the page. (a) If the wire's current runs in the +**x**-direction, which way will the wire bend? (b) Which way will the wire bend if the current runs in the –**x**-direction?

Solution

a. bends upward; b. bends downward

✓ Example 8.8.3: Force on a Circular Wire

A circular current loop of radius **R** carrying a current **I** is placed in the **xy**-plane. A constant uniform magnetic field cuts through the loop parallel to the **y**-axis (Figure 8.8.4). Find the magnetic force on the upper half of the loop, the lower half of the loop, and the total force on the loop.

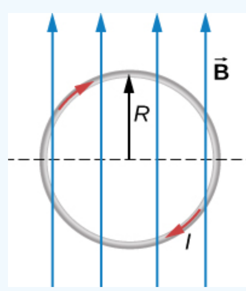


Figure 8.8.4: A loop of wire carrying a current in a magnetic field.

Strategy

The magnetic force on the upper loop should be written in terms of the differential force acting on each segment of the loop. If we integrate over each differential piece, we solve for the overall force on that section of the loop. The force on the lower loop is found in a similar manner, and the total force is the addition of these two forces.

Solution

A differential force on an arbitrary piece of wire located on the upper ring is:

$$dF = IB\,sin\,\theta\,dl, \tag{8.8.14}$$

where θ is the angle between the magnetic field direction (+**y**) and the segment of wire. A differential segment is located at the same radius, so using an arc-length formula, we have:

$$dl = Rd\theta \tag{8.8.15}$$

$$dF = IBR\, sin\,\theta\, d\theta. \tag{8.8.16}$$

In order to find the force on a segment, we integrate over the upper half of the circle, from 0 to π. This results in:

$$F = IBR\int_0^\pi sin\,\theta\, d\theta = IBR(-cos\pi + cos0) = 2IBR. \tag{8.8.17}$$

The lower half of the loop is integrated from π to zero, giving us:

$$F = IBR\int_\pi^0 sin\,\theta\, d\theta = IBR(-cos0 + cos\pi) = -2IBR. \tag{8.8.18}$$

The net force is the sum of these forces, which is zero.

Significance

The total force on any closed loop in a uniform magnetic field is zero. Even though each piece of the loop has a force acting on it, the net force on the system is zero. (Note that there is a net torque on the loop, which we consider in the next section.)

8.9: Magnetism in Matter

Learning Objectives

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

- Classify magnetic materials as paramagnetic, diamagnetic, or ferromagnetic, based on their response to a magnetic field
- Sketch how magnetic dipoles align with the magnetic field in each type of substance
- Define hysteresis and magnetic susceptibility, which determines the type of magnetic material

Why are certain materials magnetic and others not? And why do certain substances become magnetized by a field, whereas others are unaffected? To answer such questions, we need an understanding of magnetism on a microscopic level.

Within an atom, every electron travels in an orbit and spins on an internal axis. Both types of motion produce current loops and therefore magnetic dipoles. For a particular atom, the net magnetic dipole moment is the vector sum of the magnetic dipole moments. Values of μ for several types of atoms are given in Table $8.9.1$. Notice that some atoms have a zero net dipole moment and that the magnitudes of the nonvanishing moments are typically $10^{23}\ A \cdot m^2$.

Table $8.9.1$: Magnetic Moments of Some Atoms

Atom	Magnetic Moment $(10^{-24}\ A \cdot m^2)$
H	9.27
He	0
Li	9.27
O	13.9
Na	9.27
S	13.9

A handful of matter has approximately 10^{26} atoms and ions, each with its magnetic dipole moment. If no external magnetic field is present, the magnetic dipoles are randomly oriented—as many are pointed up as down, as many are pointed east as west, and so on. Consequently, the net magnetic dipole moment of the sample is zero. However, if the sample is placed in a magnetic field, these dipoles tend to align with the field, and this alignment determines how the sample responds to the field. On the basis of this response, a material is said to be either paramagnetic, ferromagnetic, or diamagnetic.

In a **paramagnetic material**, only a small fraction (roughly one-third) of the magnetic dipoles are aligned with the applied field. Since each dipole produces its own magnetic field, this alignment contributes an extra magnetic field, which enhances the applied field. When a **ferromagnetic material** is placed in a magnetic field, its magnetic dipoles also become aligned; furthermore, they become locked together so that a permanent magnetization results, even when the field is turned off or reversed. This permanent magnetization happens in ferromagnetic materials but not paramagnetic materials. **Diamagnetic materials** are composed of atoms that have no net magnetic dipole moment. However, when a diamagnetic material is placed in a magnetic field, a magnetic dipole moment is directed opposite to the applied field and therefore produces a magnetic field that opposes the applied field. We now consider each type of material in greater detail.

Paramagnetic Materials

For simplicity, we assume our sample is a long, cylindrical piece that completely fills the interior of a long, tightly wound solenoid. When there is no current in the solenoid, the magnetic dipoles in the sample are randomly oriented and produce no net magnetic field. With a solenoid current, the magnetic field due to the solenoid exerts a torque on the dipoles that tends to align them with the field. In competition with the aligning torque are thermal collisions that tend to randomize the orientations of the dipoles. The relative importance of these two competing processes can be estimated by comparing the energies involved. The energy difference between a magnetic dipole aligned with and against a magnetic field is $U_B = 2\mu B$. If $\mu = 9.3 \times 10^{-24} A \cdot m^2$ (the value of atomic hydrogen) and **B** = 1.0 T, then

$$U_B = 1.9 \times 10^{-23} J. \tag{8.9.1}$$

At a room temperature of $27^o C$ the thermal energy per atom is

$$U_T \approx kT = (1.38 \times 10^{-23} J/K)(300\ K) = 4.1 \times 10^{-21} J, \tag{8.9.2}$$

which is about 220 times greater than U_B. Clearly, energy exchanges in thermal collisions can seriously interfere with the alignment of the magnetic dipoles. As a result, only a small fraction of the dipoles is aligned at any instant.

The four sketches of Figure $8.9.1$ furnish a simple model of this alignment process. In part (a), before the field of the solenoid (not shown) containing the paramagnetic sample is applied, the magnetic dipoles are randomly oriented and there is no net magnetic dipole moment associated with the material. With the introduction of the field, a partial alignment of the dipoles takes place, as depicted in part (b). The component of the net magnetic dipole moment that is perpendicular to the field vanishes. We may then represent the sample by part (c), which shows a collection of magnetic dipoles completely aligned with the field. By treating these dipoles as current loops, we can picture the dipole alignment as equivalent to a current around the surface of the material, as in part (d). This fictitious surface current produces its own magnetic field, which enhances the field of the solenoid.

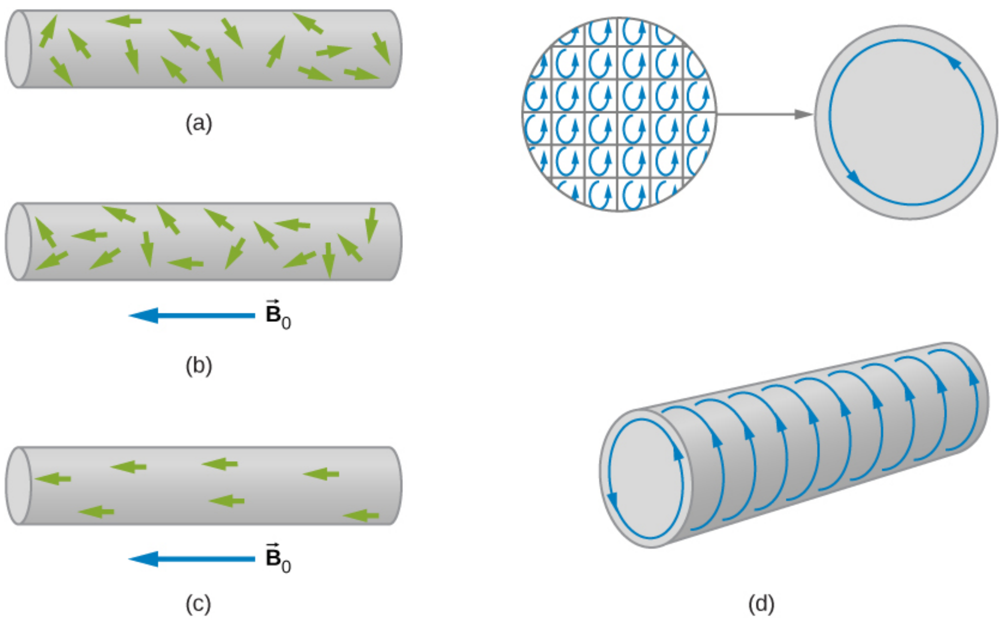


Figure $8.9.1$: The alignment process in a paramagnetic material filling a solenoid (not shown). (a) Without an applied field, the magnetic dipoles are randomly oriented. (b) With a field, partial alignment occurs. (c) An equivalent representation of part (b). (d) The internal currents cancel, leaving an effective surface current that produces a magnetic field similar to that of a finite solenoid.

We can express the total magnetic field $\vec{B}$ in the material as

$$\vec{B} = \vec{B}_0 + \vec{B}_m, \tag{8.9.3}$$

where $\vec{B}_0$ is the field due to the current I_0 in the solenoid and $\vec{B}_m$ is the field due to the surface current I_m around the sample. Now $\vec{B}_m$ is usually proportional to $\vec{B}_0$ a fact we express by

$$\vec{B}_m = \chi \vec{B}_0, \tag{8.9.4}$$

where χ is a dimensionless quantity called the magnetic susceptibility. Values of χ for some paramagnetic materials are given in Table $8.9.2$. Since the alignment of magnetic dipoles is so weak, χ is very small for paramagnetic materials. By combining Equation $8.9.3$ and Equation $8.9.4$, we obtain:

$$\vec{B} = \vec{B}_0 + \chi \vec{B}_0 = (1 + \chi)\vec{B}_0. \tag{8.9.5}$$

For a sample within an infinite solenoid, this becomes

$$B = (1 + \chi)\mu_0 nI. \tag{8.9.6}$$

This expression tells us that the insertion of a paramagnetic material into a solenoid increases the field by a factor of $(1 + \chi)$. However, since χ is so small, the field isn't enhanced very much.

The quantity

$$\mu = (1 + \chi)\mu_0. \tag{8.9.7}$$

is called the **magnetic permeability of a material**. In terms of μ, Equation 8.9.6 can be written as

$$B = \mu n I \tag{8.9.8}$$

for the filled solenoid.

Table 8.9.2: Magnetic Susceptibilities*Note: Unless otherwise specified, values given are for room temperature.

Paramagnetic Materials	χ	Diamagnetic Materials	χ
Aluminum	2.2×10^{-5}	Bismuth	-1.7×10^{-5}
Calcium	1.4×10^{-5}	Carbon (diamond)	-2.2×10^{-5}
Chromium	3.1×10^{-4}	Copper	-9.7×10^{-6}
Magnesium	1.2×10^{-5}	Lead	-1.8×10^{-5}
Oxygen gas (1 atm)	1.8×10^{-6}	Mercury	-2.8×10^{-5}
Oxygen liquid (90 K)	3.5×10^{-3}	Hydrogen gas (1 atm)	-2.2×10^{-9}
Tungsten	6.8×10^{-5}	Nitrogen gas (1 atm)	-6.7×10^{-9}
Air (1 atm)	3.6×10^{-7}	Water	-9.1×10^{-6}

Diamagnetic Materials

A magnetic field always induces a magnetic dipole in an atom. This induced dipole points opposite to the applied field, so its magnetic field is also directed opposite to the applied field. In paramagnetic and ferromagnetic materials, the induced magnetic dipole is masked by much stronger permanent magnetic dipoles of the atoms. However, in diamagnetic materials, whose atoms have no permanent magnetic dipole moments, the effect of the induced dipole is observable.

We can now describe the magnetic effects of diamagnetic materials with the same model developed for paramagnetic materials. In this case, however, the fictitious surface current flows opposite to the solenoid current, and the magnetic susceptibility χ is negative. Values of χ for some diamagnetic materials are also given in Table 8.9.2.

> Water is a common diamagnetic material. Animals are mostly composed of water. Experiments have been performed on frogs and mice in diverging magnetic fields. The water molecules are repelled from the applied magnetic field against gravity until the animal reaches an equilibrium. The result is that the animal is levitated by the magnetic field.

Ferromagnetic Materials

Common magnets are made of a ferromagnetic material such as iron or one of its alloys. Experiments reveal that a ferromagnetic material consists of tiny regions known as **magnetic domains**. Their volumes typically range from 10^{-12} to $10^{-8} m^3$, and they contain about 10^{17} to 10^{21} atoms. Within a domain, the magnetic dipoles are rigidly aligned in the same direction by coupling among the atoms. This coupling, which is due to quantum mechanical effects, is so strong that even thermal agitation at room temperature cannot break it. The result is that each domain has a net dipole moment. Some materials have weaker coupling and are ferromagnetic only at lower temperatures.

If the domains in a ferromagnetic sample are randomly oriented, as shown in Figure 8.9.1a, the sample has no net magnetic dipole moment and is said to be unmagnetized. Suppose that we fill the volume of a solenoid with an unmagnetized ferromagnetic sample. When the magnetic field $\vec{B}_0$ of the solenoid is turned on, the dipole moments of the domains rotate so that they align somewhat with the field, as depicted in Figure 8.9.1b. In addition, the aligned domains tend to increase in size at the expense of unaligned ones. The net effect of these two processes is the creation of a net magnetic dipole moment for the ferromagnet that is directed along the applied magnetic field. This net magnetic dipole moment is much larger than that of a paramagnetic sample, and the domains, with their large numbers of atoms, do not become misaligned by thermal agitation. Consequently, the field due to the alignment of the domains is quite large.

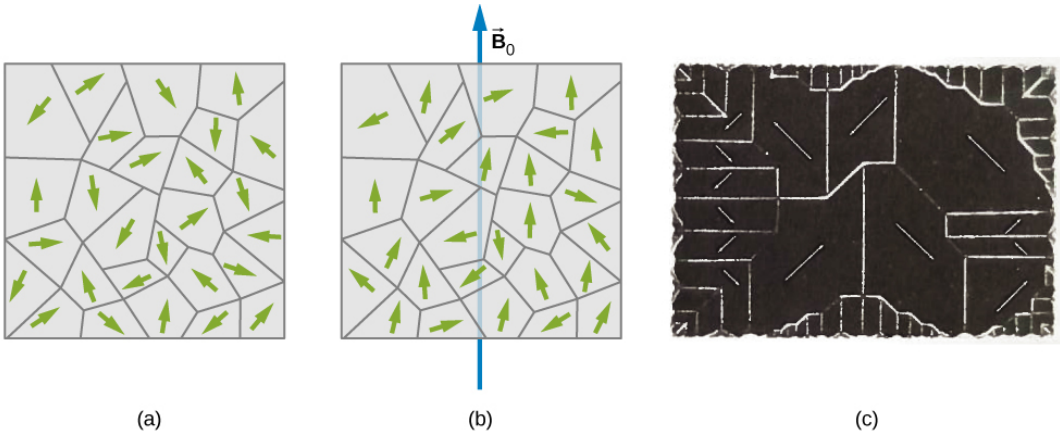


Figure 8.9.2: (a) Domains are randomly oriented in an unmagnetized ferromagnetic sample such as iron. The arrows represent the orientations of the magnetic dipoles within the domains. (b) In an applied magnetic field, the domains align somewhat with the field. (c) The domains of a single crystal of nickel. The white lines show the boundaries of the domains. These lines are produced by iron oxide powder sprinkled on the crystal.

Besides iron, only four elements contain the magnetic domains needed to exhibit ferromagnetic behavior: cobalt, nickel, gadolinium, and dysprosium. Many alloys of these elements are also ferromagnetic. Ferromagnetic materials can be described using Equation 8.9.5 through Equation 8.9.8, the paramagnetic equations. However, the value of χ for ferromagnetic material is usually on the order of 10^3 to 10^4, and it also depends on the history of the magnetic field to which the material has been subject. A typical plot of **B** (the total field in the material) versus B_0 (the applied field) for an initially unmagnetized piece of iron is shown in Figure 8.9.2c. Some sample numbers are (1) for $B_0 = 1.0 \times 10^{-4} T$, $B = 0.60\, T$, and $\chi = \left({}^{0.60}/_{1.0\times10^{-4}}\right) - 1 \approx 6.0 \times 10^3$; for (2) for $B_0 = 6.0 \times 10^{-4} T$, $B = 1.5\, T$, and $\chi = \left({}^{1.5}/_{6.0\times10^{-4}}\right) - 1 \approx 2.5 \times 10^3$.

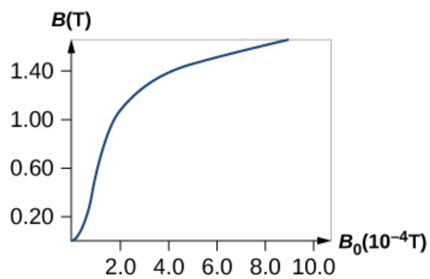


Figure 8.9.3: (a) The magnetic field B in annealed iron as a function of the applied field B_0.

When B_0 is varied over a range of positive and negative values, **B** is found to behave as shown in Figure 8.9.3. Note that the same B_0 (corresponding to the same current in the solenoid) can produce different values of **B** in the material. The magnetic field **B** produced in a ferromagnetic material by an applied field B_0 depends on the magnetic history of the material. This effect is called **hysteresis**, and the curve of Figure 8.9.4 is called a hysteresis loop. Notice that **B** does not disappear when $B_0 = 0$ (i.e., when the current in the solenoid is turned off). The iron stays magnetized, which means that it has become a permanent magnet.

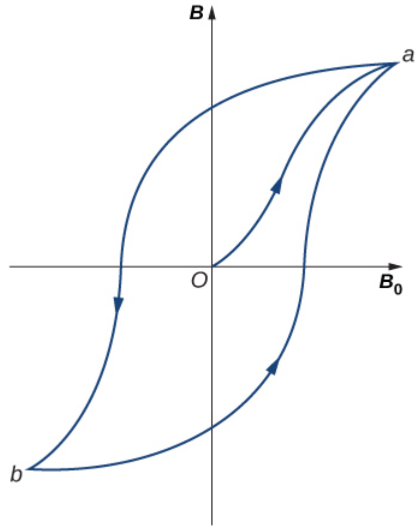


Figure 8.9.4: A typical hysteresis loop for a ferromagnet. When the material is first magnetized, it follows a curve from 0 to **a**. When B_0 is reversed, it takes the path shown from **a** to **b**. If B_0 is reversed again, the material follows the curve from **b** to **a**.

Like the paramagnetic sample of Figure 8.9.2, the partial alignment of the domains in a ferromagnet is equivalent to a current flowing around the surface. A bar magnet can therefore be pictured as a tightly wound solenoid with a large current circulating through its coils (the surface current). You can see in Figure 8.9.5 that this model fits quite well. The fields of the bar magnet and the finite solenoid are strikingly similar. The figure also shows how the poles of the bar magnet are identified. To form closed loops, the field lines outside the magnet leave the north (N) pole and enter the south (S) pole, whereas inside the magnet, they leave S and enter N.

Figure 8.9.5: Comparison of the magnetic fields of a finite solenoid and a bar magnet.

Ferromagnetic materials are found in computer hard disk drives and permanent data storage devices (Figure 8.9.6). A material used in your hard disk drives is called a spin valve, which has alternating layers of ferromagnetic (aligning with the external magnetic field) and antiferromagnetic (each atom is aligned opposite to the next) metals. It was observed that a significant change in resistance was discovered based on whether an applied magnetic field was on the spin valve or not. This large change in resistance creates a quick and consistent way for recording or reading information by an applied current.

Figure 8.9.6: The inside of a hard disk drive. The silver disk contains the information, whereas the thin stylus on top of the disk reads and writes information to the disk.

Example 8.9.1: Iron Core in a Coil

A long coil is tightly wound around an iron cylinder whose magnetization curve is shown in Figure 8.9.3. (a) If $n = 20$ turns per centimeter, what is the applied field B_0 when $I_0 = 0.20\ A$? (b) What is the net magnetic field for this same current? (c) What is the magnetic susceptibility in this case?

Strategy

(a) The magnetic field of a solenoid is calculated using $\vec{B} = \mu_0 n I \hat{j}$. (b) The graph is read to determine the net magnetic field for this same current. (c) The magnetic susceptibility is calculated using Equation 8.9.6.

Solution

1. The applied field B_0 of the coil is

$$B_0 = \mu_0 n I_0 = (4\pi \times 10^{-7} T \cdot m/A)(2000/m)(0.20\ A) \tag{8.9.9}$$

$$B_0 = 5.0 \times 10^{-4} T. \tag{8.9.10}$$

2. From inspection of the magnetization curve of Figure 8.9.3, we see that, for this value of B_0, $B = 1.4\ T$. Notice that the internal field of the aligned atoms is much larger than the externally applied field.
3. The magnetic susceptibility is calculated to be

$$\chi = \frac{B}{B_0} - 1 = \frac{1.4\ T}{5.0 \times 10^{-4} T} - 1 = 2.8 \times 10^3. \tag{8.9.11}$$

Significance

Ferromagnetic materials have susceptibilities in the range of 10^3 which compares well to our results here. Paramagnetic materials have fractional susceptibilities, so their applied field of the coil is much greater than the magnetic field generated by the material.

Exercise 8.9.1

Repeat the calculations from the previous example for $I_0 = 0.040\ A$.

Answer

a. $1.0 \times 10^{-4} T$; b. 0.60 T; c. 6.0×10^3

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).

8.10: The Magnetic Field (Summary)

Key Terms

cosmic rays	comprised of particles that originate mainly from outside the solar system and reach Earth
gauss	G, unit of the magnetic field strength; $1G = 10^{-4}T$
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 **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
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	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: $1T = 1N/A-m$

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

Force on a charge in a magnetic field	$\vec{F} = q\vec{v} \times \vec{B}$
Magnitude of magnetic force	$F = qvBsin\theta$
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	$\vec{F} = I\vec{l} \times \vec{B}$

Permeability of free space	$\mu_0 = 4\pi \times 10^{-7} T \cdot m/A$
Contribution to magnetic field from a current element	$dB = \frac{\mu_0}{4\pi}\frac{Idlsin\theta}{r^2}$
Biot–Savart law	$\vec{B} = \frac{\mu_0}{4\pi}\int_{wire}\frac{Id\vec{l} \times \hat{r}}{r^2}$
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)
Magnetic field strength inside a solenoid	$B = \mu_0 nI$
Magnetic field strength inside a toroid	$B = \frac{\mu_o NI}{2\pi r}$
Magnetic permeability	$\mu = (1+\chi)\mu_0$
Magnetic field of a solenoid filled with paramagnetic material	$B = \mu nI$

Summary

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.

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.

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.

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.
- 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.

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.

Common Magnetic Field Models

- 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 $\mu_0 = 4\pi \times 10^{-7} T \cdot m/s$ 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.
- 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.
- The magnetic field strength inside a solenoid is $B = \mu_0 nI$ (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 $B = \frac{\mu_o NI}{2\pi r}$ (within the toroid) where N is the number of windings. The field inside a toroid is not uniform and varies with the distance as $1/r$.

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.

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).

8.11: 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 accidently 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?

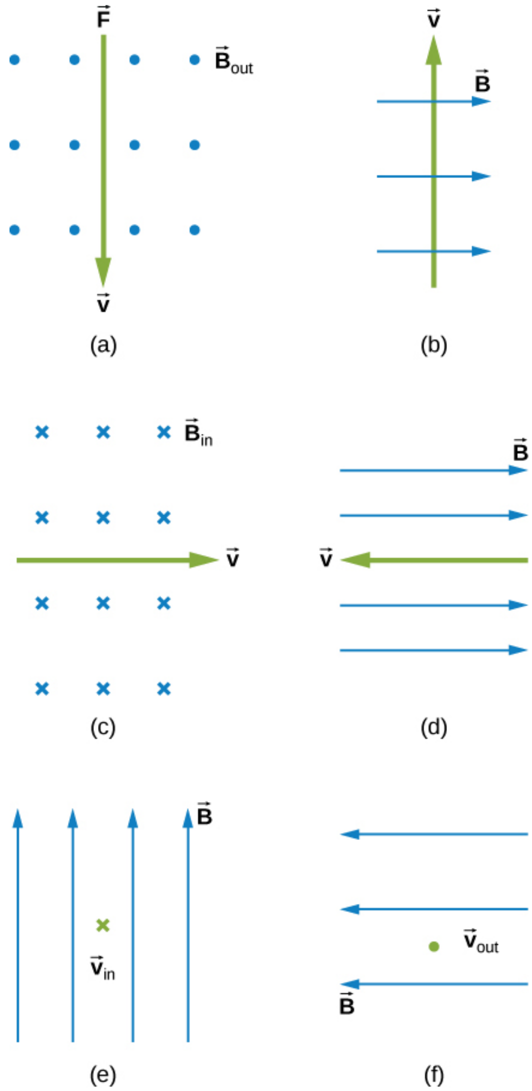
11.8 Applications of Magnetic Forces and Fields

14. Describe the primary function of the electric field and the magnetic field in a cyclotron.

Problems

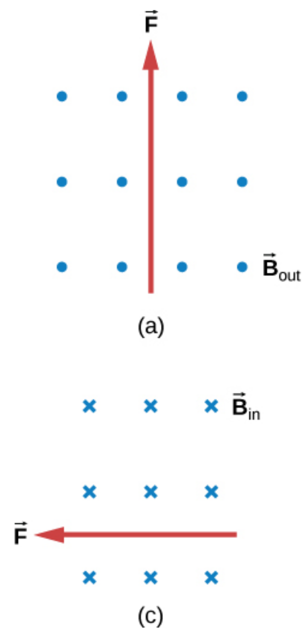
11.3 Magnetic Fields and Lines

15. What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases?



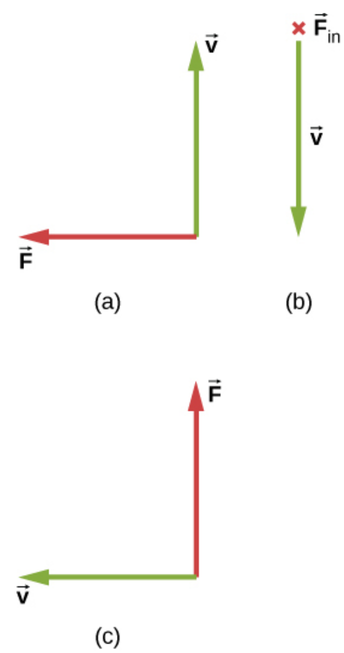
16. Repeat previous exercise for a negative charge.

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-μC charge and flies due west at a speed of 660. m/s over Earth's south magnetic pole, where the $8.00 \times 10^{-5} - T$ 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 $5.00 \times 10^{7} m/s$ experiences a magnetic force of $1.70 \times 10^{-16} N$. What is the strength of the magnetic field if there is a 45º 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 $4.00 \times 10^3 m/s$ in a 1.25-T magnetic field experiences a magnetic force of $1.40 \times 10^{-16} N$. 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 $1.00 \times 10^{-12} N$. 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 $7.5 \times 10^6 m/s$ perpendicular to Earth's magnetic field at an altitude where the field strength is $1.0 \times 10^{-5} T$. 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 $5.0 \times 10^7 m/s$ 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 $2.66 \times 10^{-26} kg$ travels at $5.0 \times 10^6 m/s$ 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 $6.0 \times 10^7 m/s$, in a direction perpendicular to Earth's field, which has a strength of $5.0 \times 10^{-5} T$. (a) What strength electric field must be applied perpendicular to the Earth's field to make the electron moves 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 $6.00 \times 10^{-7} m/s$? (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 $(m = 6.64 \times 10^{-27} kg, q = 3.2 \times 10^{-19} C)$ 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?

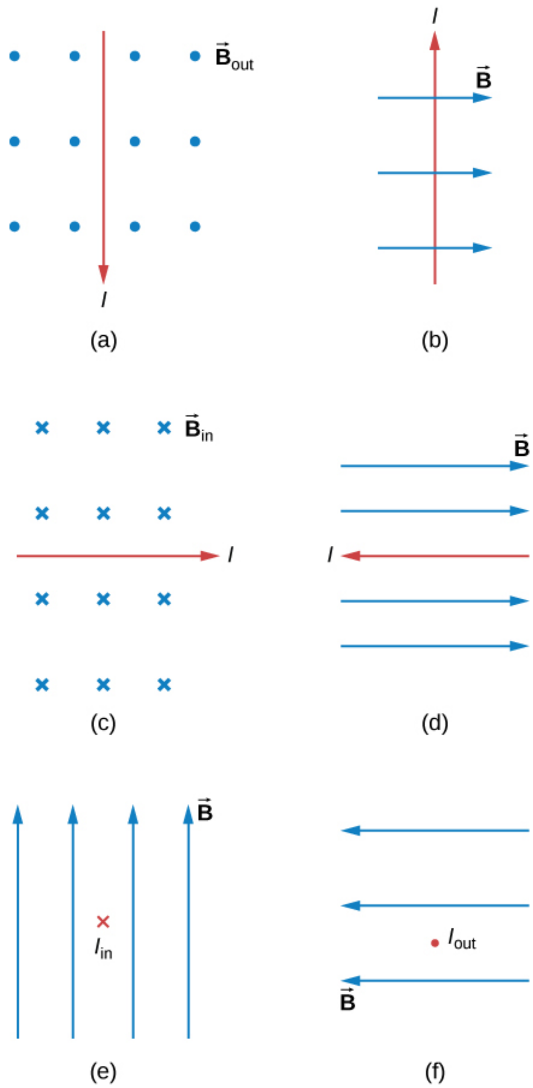
(b) What is the kinetic energy in electron-volts?

(c) Through what potential difference must the particle be accelerated in order to give it this kinetic energy?

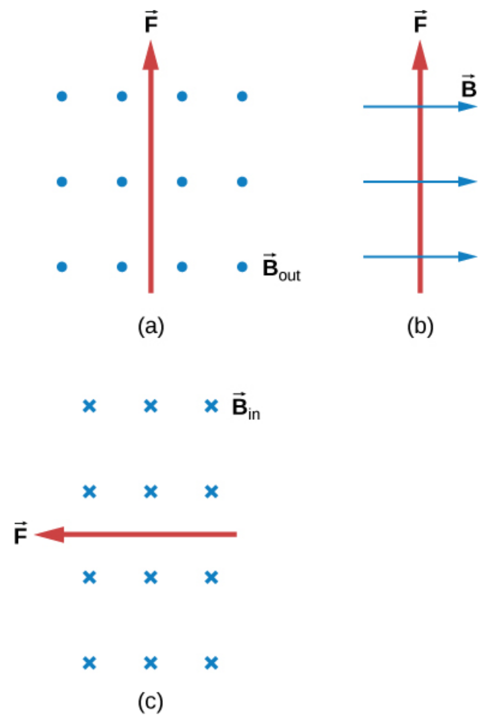
32. A particle of charge **q** and mass **m** is accelerated from rest through a potential difference **V**, after which it encounters a uniform magnetic field **B**. If the particle moves in a plane perpendicular to **B**, what is the radius of its circular orbit?

11.5 Magnetic Force on a Current-Carrying Conductor

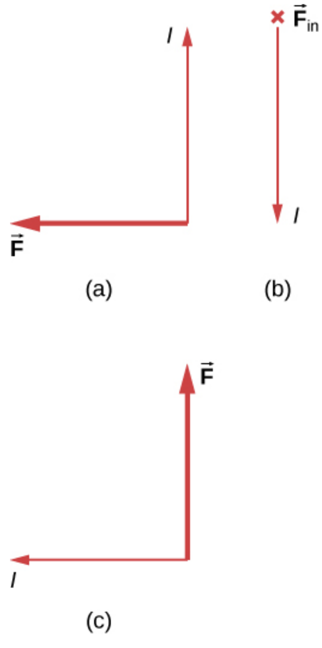
33. What is the direction of the magnetic force on the current in each of the six cases?



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 **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 30.0º 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 10.9º?

41. Find the current through a loop needed to create a maximum torque of **9.0N·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 **τ = NIAB** sin **θ**, 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?

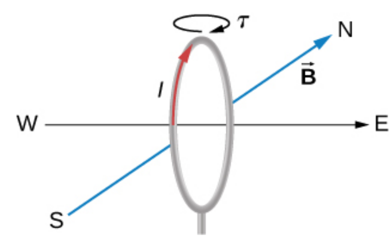
(b) 50.0% of maximum?

(c) 10.0% of maximum?

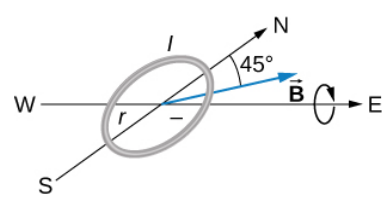
45. A proton has a magnetic field due to its spin. The field is similar to that created by a circular current loop $0.65 \times 10^{-15} m$ in radius with a current of $1.05 \times 10^4 A$. Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

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

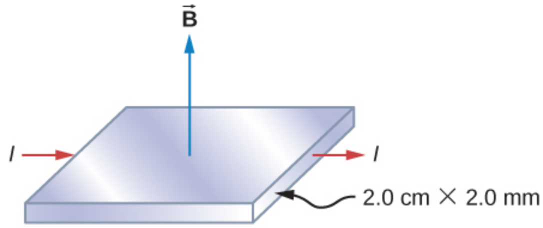
48. A strip of copper is placed in a uniform magnetic field of magnitude 2.5 T. The Hall electric field is measured to be $1.5 \times 10^{-3} V/m$.

(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 $_sO^{16}$ 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 **1.5μV** 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 **2μV** 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 $B_o = 0.750T$. 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 $3.0 \times 10^5 m/s$ 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

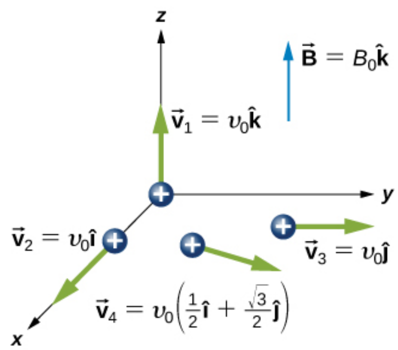
58. Calculate the magnetic force on a hypothetical particle of charge $1.0\times10^{-19}C$ moving with a velocity of $6.0\times10^{4}\hat{i}m/s$ in a magnetic field of $1.2\hat{k}T$.

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 $(3.0\hat{i}+4.0\hat{j})\times10^{6}m/s$. What is the magnetic force on the electron?

61. The mass and charge of a water droplet are $1.0\times10^{-4}g$ and $2.0\times10^{-8}C$, respectively. If the droplet is given an initial horizontal velocity of $5.0\times10^{5}\hat{i}m/s$, 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 $(m=6.64\times10^{-27}kg, q=3.2\times10^{-19}C)$ moving with a velocity $\vec{v}=(2.0\hat{i}-4.0\hat{k})\times10^{6}m/s$ enters a region where $\vec{E}=(5.0\hat{i}-2.0\hat{j})\times10^{4}V/m$ and $\vec{B}=(1.0\hat{i}+4.0\hat{k})\times10^{-2}T$. What is the initial force on it?

65. An electron moving with a velocity $\vec{v}=(4.0\hat{i}\,\hat{}+3.0\hat{j}\,\hat{}+2.0\hat{k})\times10^{6}m/s$ enters a region where there is a uniform electric field and a uniform magnetic field. The magnetic field is given by $\vec{B}=(1.0\hat{i}-2.0\hat{j}+4.0\hat{k})\times10^{-2}T$. 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 $1.8\times10^{-5}T$ north and a vertical component of $5.0\times10^{-5}T$ 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 x 10^{-5} T north and a vertical component of 5.0 x 10^{-5} T down?

68. What magnetic field is required in order to confine a proton moving with a speed of $4.0\times10^{6}m/s$ 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 $2.0\times10^{-3}s$ to complete eight revolutions in a uniform magnetic field of magnitude $2.0\times10^{-2}T$. What is the mass of the ion?

72. A particle moving downward at a speed of $6.0\times10^{6}m/s$ enters a uniform magnetic field that is horizontal and directed from east to west.

(a) If the particle is deflected initially to the north in a circular arc, is its charge positive or negative?

(b) If B = 0.25 T and the charge-to-mass ratio (q/m) of the particle is $4.0 \times 10^7 C/kg$, what is the radius of the path?

(c) What is the speed of the particle after it has moved in the field for $1.0 \times 10^{-5} s$? 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 $7.5 \times 10^{-2} T$ completes 10 revolutions in $3.47 \times 10^{-4} s$. 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 $\vec{v} = (4\hat{j} + 3\hat{k}) \times 10^6 m/s$ and travels in a helical path with a radius of 5.0 cm.

(a) What is the value of B?

(b) What is the time required for one trip around the helix?

(c) Where is the proton $5.0 \times 10^{-7} s$ after entering the field?

77. An electron moving at $5.0 \times 10^6 m/s$ enters a magnetic field that makes a 75^o angle with the x-axis of magnitude 0.20 T. Calculate the

(a) pitch and

(b) 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 60º with Earth's $5.5 \times 10^{-5} T$ field. What is the current when the wire experiences a force of $7.0 \times 10^{-3} N$?

(b) 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?

(b) What is the force on the wire if it is rotated to make an angle of 90º 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 $(3.0\hat{i} \times 4.0\hat{k}) \times 10^{-3} T$. 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 $8.0 \times 10^{-3} T$. Calculate the magnitude of the force on the section if the angle between the field and the direction of the current is

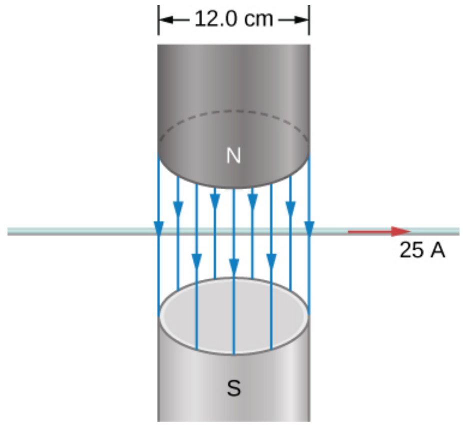
(a) 45°;

(b) 90°;

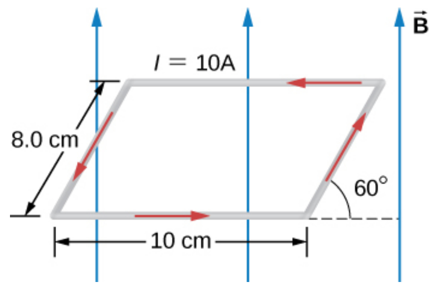
(c) 0°; or

(d) 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 A** and **B = 1.5 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 Am2?

(b) What is the maximum torque that the coil will experience in a uniform field of strength $5.0 \times 10^{-2} T$?

(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 **μV** 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 8.70×105m/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?

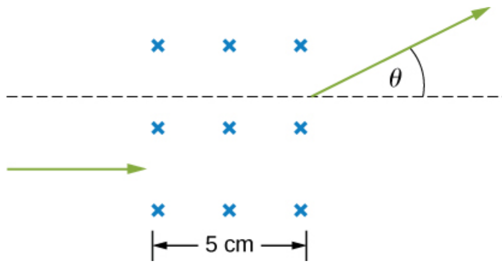
(c) Which premise is responsible?

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.

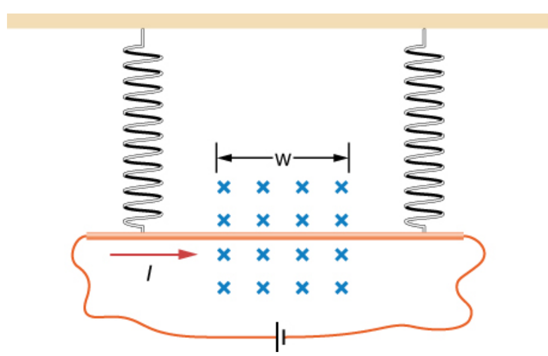
103. A proton of speed $v = 6 \times 10^5 m/s$ enters a region of uniform magnetic field of **B = 0.5 T** at an angle of $q = 30^\circ$ to the magnetic field. In the region of magnetic field proton describes a helical path with radius R and pitch p (distance between loops). Find **R** and **p**.

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 $4 \times 10^6 m/s$ 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.)

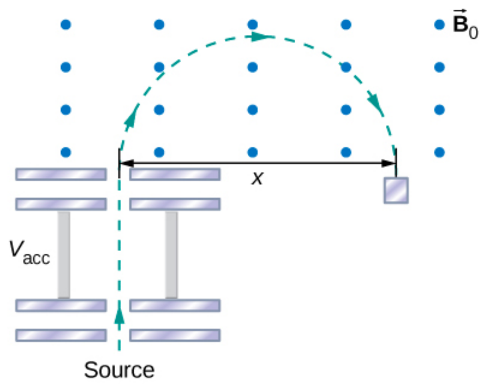


105. In a region a non-uniform magnetic field exists such that $B_x = 0, B_y = 0,$ and $B_z = ax$, where a is a constant. At some time **t**, a wire of length **L** is carrying a current **I** is located along the **x-axis** from origin to **x = L**. Find the magnetic force on the wire at this instant in time.

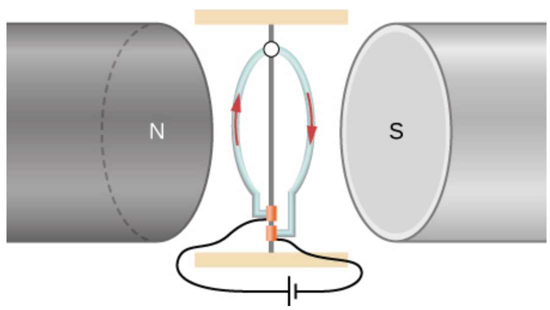
106. A copper rod of mass **m** and length **L** is hung from the ceiling using two springs of spring constant **k**. A uniform magnetic field of magnitude B_0 pointing perpendicular to the rod and spring (coming out of the page in the figure) exists in a region of space covering a length **w** of the 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 **B** in millitesla. Calculate the magnetic force on the segment of wire between **x = 2.0 m** and **x = 4.0 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 $\vec{B} = (2.0\hat{i} + 6.0\hat{j} + 8.0\hat{k}) \times 10^{-3} T$. 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?

8.12: 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 $\vec{B}=0$ inside a long, hollow copper pipe that is carrying an electric current parallel to the axis. Is $\vec{B}=0$ outside the pipe?

12.8 Magnetism in Matter

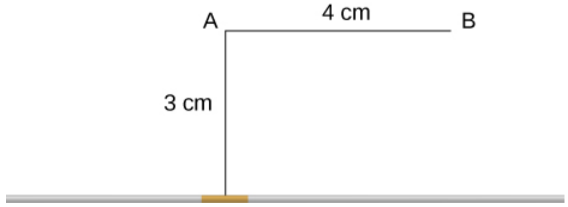
14. A diamagnetic material is brought close to a permanent magnet. What happens to the material?

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.

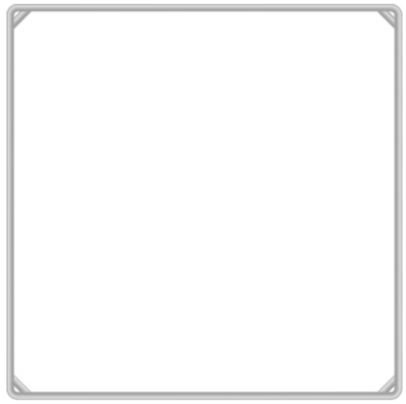
Problems

12.2 The Biot-Savart Law

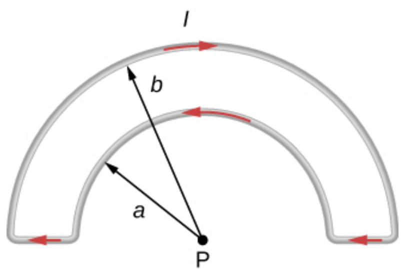
16. A 10-A current flows through the wire shown. What is the magnitude of the magnetic field due to a 0.5-mm segment of wire as measured at (a) point A and (b) point B?



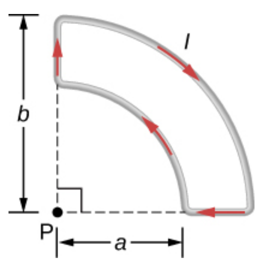
17. Ten amps flow through a square loop where each side is 20 cm in length. At each corner of the loop is a 0.01-cm segment that connects the longer wires as shown. Calculate the magnitude of the magnetic field at the center of the loop.



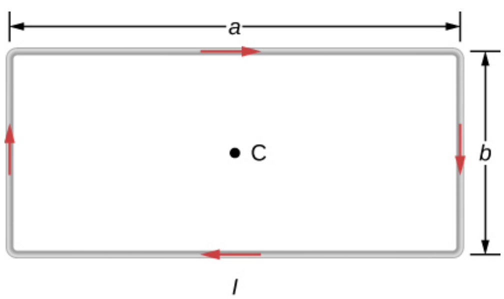
18. What is the magnetic field at P due to the current I in the wire shown?



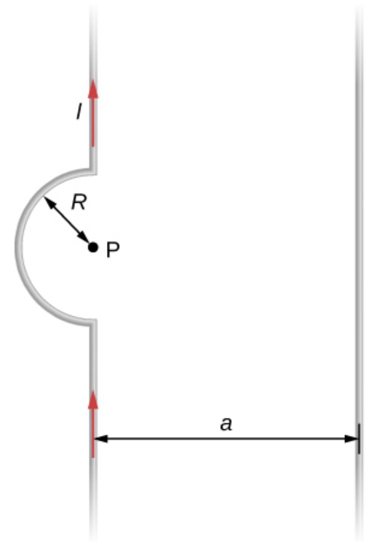
19. The accompanying figure shows a current loop consisting of two concentric circular arcs and two perpendicular radial lines. Determine the magnetic field at point P.



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.

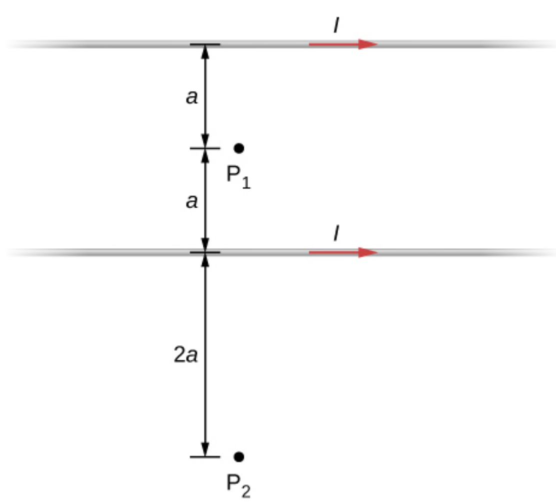
23. The magnitude of the magnetic field 50 cm from a long, thin, straight wire is **8.0μT**. What is the current through the long wire?

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 $4.0 \times 10^{-5} T$ 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 $I_1 = 10A$ and $I_2 = 20A$, 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.

(a) If each wire carries a current of 50 A in the same direction, what is the magnetic force per meter exerted on each wire?

(b) Does the force pull the wires together or push them apart?

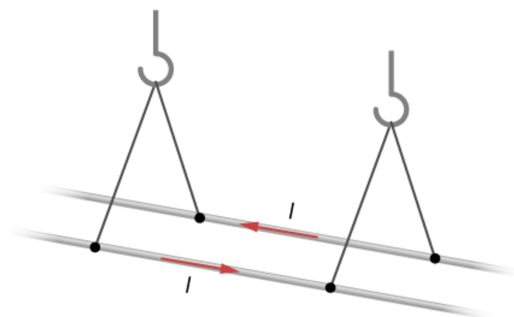
(c) What happens if the currents flow in opposite directions?

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.

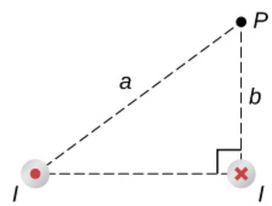
(a) If the two currents flow in opposite directions, what is the magnitude and direction of the force per unit length of one wire on the other?

(b) What is the magnitude and direction of the force per unit length if the currents flow in the same direction?

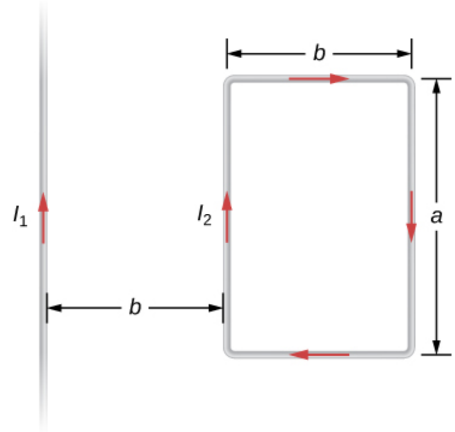
32. Two long, parallel wires are hung by cords of length 5.0 cm, as shown in the accompanying figure. Each wire has a mass per unit length of 30 g/m, and they carry the same current in opposite directions. What is the current if the cords hang at **6.0°** with respect to the vertical?



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.

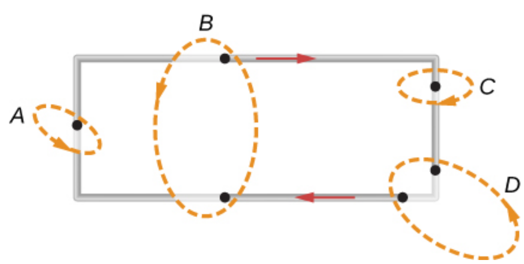
38. A circular loop of radius **R** carries a current **I**. At what distance along the axis of the loop is the magnetic field one-half its value at the center of the loop?

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?

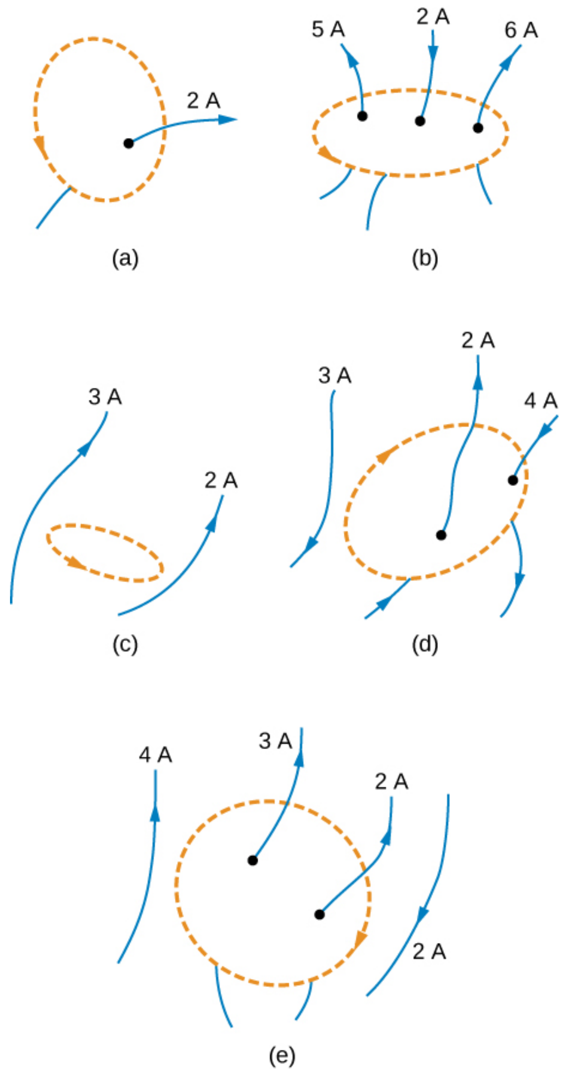
40. For the coils in the preceding problem, what is the magnetic field at the center of either coil?

12.6 Ampère's Law

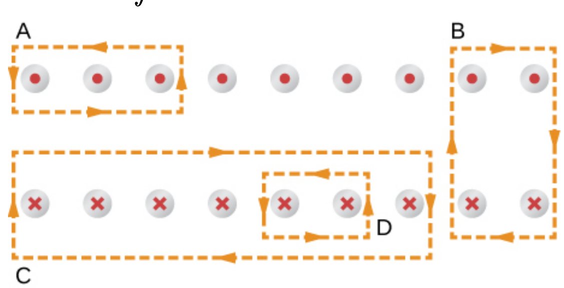
41. A current **I** flows around the rectangular loop shown in the accompanying figure. Evaluate $\oint \vec{B} \cdot \vec{dl}$ for the paths **A, B,C**, and **D.**



42. Evaluate $\oint \vec{B} \cdot \vec{dl}$ for each of the cases shown in the accompanying figure.



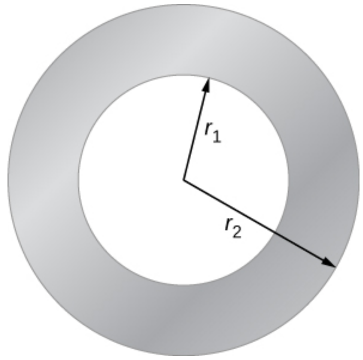
43. The coil whose lengthwise cross section is shown in the accompanying figure carries a current I and has N evenly spaced turns distributed along the length l. Evaluate $\oint \vec{B} \cdot \overrightarrow{dl}$ for the paths indicated.



44. A superconducting wire of diameter 0.25 cm carries a current of 1000 A. What is the magnetic field just outside the wire?

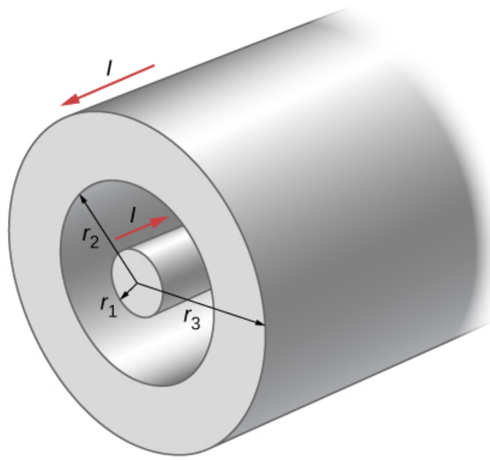
45. A long, straight wire of radius R carries a current I that is distributed uniformly over the cross-section of the wire. At what distance from the axis of the wire is the magnitude of the magnetic field a maximum?

46. The accompanying figure shows a cross-section of a long, hollow, cylindrical conductor of inner radius $r_1 = 3.0 cm$ and outer radius $r_2 = 5.0 cm$. A 50-A current distributed uniformly over the cross-section flows into the page. Calculate the magnetic field at $r = 2.0 cm$, $r = 4.0 cm$, and $r = 6.0 cm$.



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?

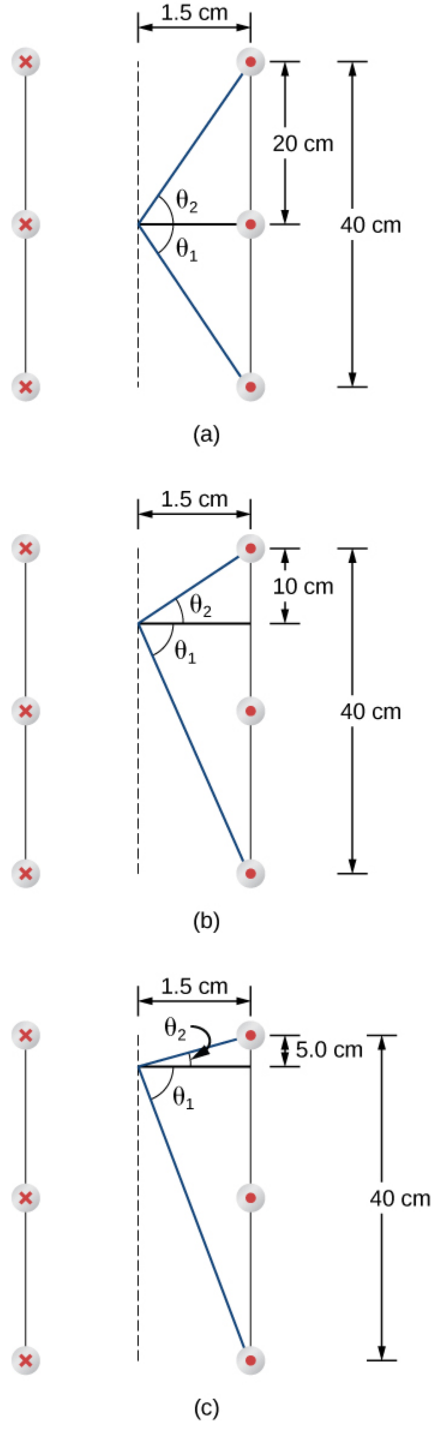
52. A solenoid is 40 cm long, has a diameter of 3.0 cm, and is wound with 500 turns. If the current through the windings is 4.0 A, what is the magnetic field at a point on the axis of the solenoid that is

(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.



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 $B=\mu_0 nI$ 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 $2.0\times10^5 m/s$, 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 **8.9cm**?

57. A toroid with a square cross section 3.0 cm × 3.0 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 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 $2.1 \times 10^{-23} A \cdot m^2$.

(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?

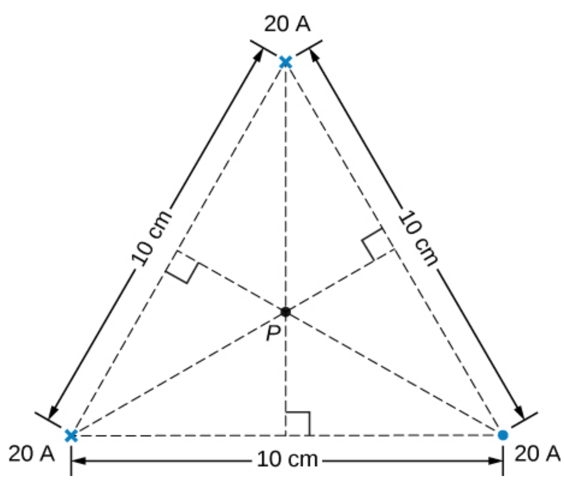
62. Suppose you wish to produce a 1.2-T magnetic field in a toroid with an iron core for which $\chi = 4.0 \times 10^3$. The toroid has a mean radius of 15 cm and is wound with 500 turns. What current is required?

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?

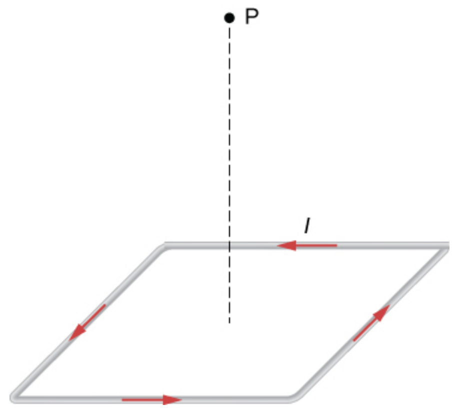
64. A solenoid with an iron core is 25 cm long and is wrapped with 100 turns of wire. When the current through the solenoid is 10 A, the magnetic field inside it is 2.0 T. For this current, what is the permeability of the iron? If the current is turned off and then restored to 10 A, will the magnetic field necessarily return to 2.0 T?

Additional Problems

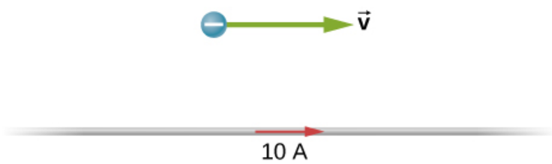
65. Three long, straight, parallel wires, all carrying 20 A, are positioned as shown in the accompanying figure. What is the magnitude of the magnetic field at the point **P**?



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)?



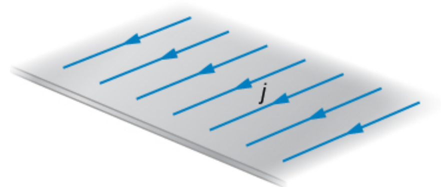
67. The accompanying figure shows a long, straight wire carrying a current of 10 A. What is the magnetic force on an electron at the instant it is 20 cm from the wire, traveling parallel to the wire with a speed of $2.0 \times 10^5 m/s$? Describe qualitatively the subsequent motion of the electron.



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.

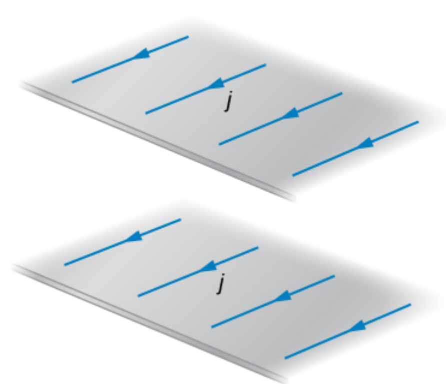
(a) Use the Biot-Savart law to show that $B = \mu_0 J/2$ on either side of the sheet. What is the direction of $\vec{B}$ on each side?

(b) Now use Ampère's law to calculate the field.

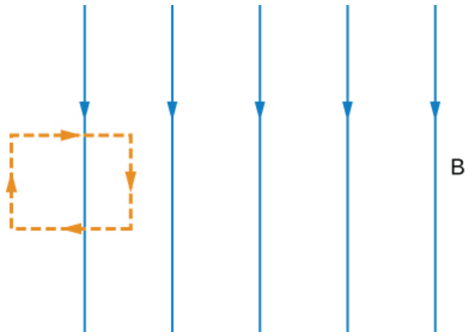


69. (a) Use the result of the previous problem to calculate the magnetic field between, above, and below the pair of infinite sheets shown in the accompanying figure.

(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?

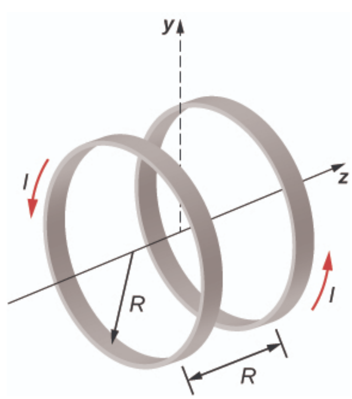
74. A wire element has $vecdl, I\vec{dl} = JAdl = Jdv$, where A and dv are the cross-sectional area and volume of the element, respectively. Use this, the Biot-Savart law, and $J = nev$ to show that the magnetic field of a moving point charge q is given by:

$$\vec{B} = \frac{\mu_0}{4\pi}\frac{qv \times \hat{r}}{r^2}.$$

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_dz2 are both zero at **z = 0.** (These vanishing derivatives demonstrate that the magnetic field varies only slightly near **z = 0**.)



76. A charge of **4.0μC** is distributed uniformly around a thin ring of insulating material. The ring has a radius of 0.20 m and rotates at $2.0 \times 10^4 rev/min$ 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?

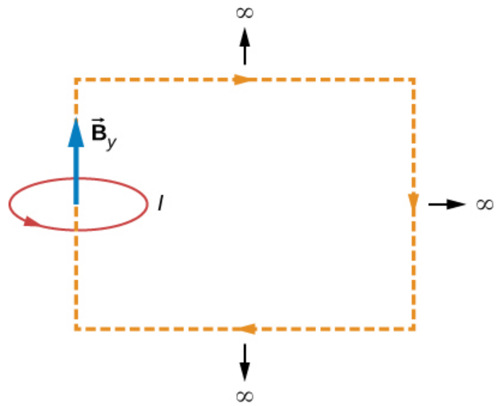
77. A thin, nonconducting disk of radius **R** is free to rotate around the axis that passes through its center and is perpendicular to the face of the disk. The disk is charged uniformly with a total charge **q**. If the disk rotates at a constant angular velocity **ω**, what is the magnetic field at 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.

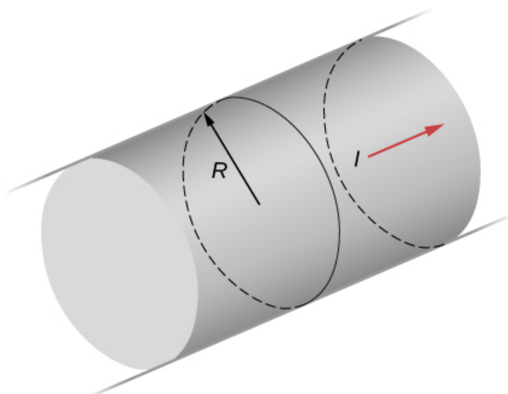
79. Consider the axial magnetic field $B_y = \mu_0 IR^2/2(y^2+R^2)^{3/2}$ of the circular current loop shown below.

(a) Evaluate $\int_{-a}^{a} B_y dy$. Also so show that $\lim_{a\to\infty}\int_{-a}^{a} B_y dy = \mu_0 I$.

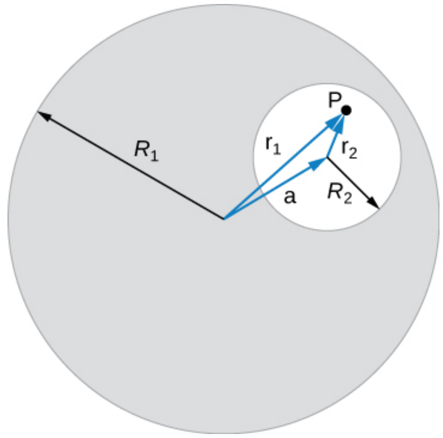
(b) Can you deduce this limit without evaluating the integral? (**Hint:**See the accompanying figure.)



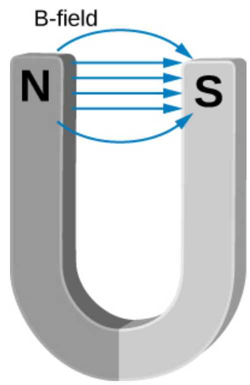
80. The current density in the long, cylindrical wire shown in the accompanying figure varies with distance r from the center of the wire according to $J = cr$, where c is a constant. (a) What is the current through the wire? (b) What is the magnetic field produced by this current for $r \leq R$? For $r \geq R$?



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 $\vec{B} = \frac{1}{2}\mu_0 J_0 k \times a$, where $a = r_1 - r_2$ and $r_1 = r_1\hat{r_1}$ is the position of P relative to the center of the conductor and $2_{=} r_2\overrightarrow{r_2}$ is the position of **P** relative to the center of the cavity.

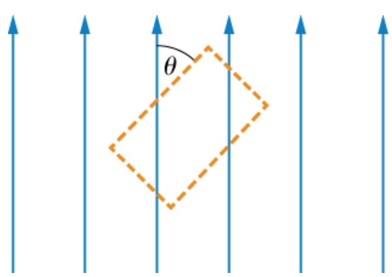


82. Between the two ends of a horseshoe magnet the field is uniform as shown in the diagram. As you move out to outside edges, the field bends. Show by Ampère’s law that the field must bend and thereby the field weakens due to these bends.



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 $\vec{B} \cdot \vec{dl}$ for each side of the loop then find the entire $\oint \vec{B} \cdot \vec{dl}$. 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

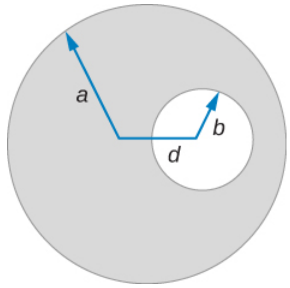
(b) at a point inside the wire. Write your answer in terms of the net current I through the wire.

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.)

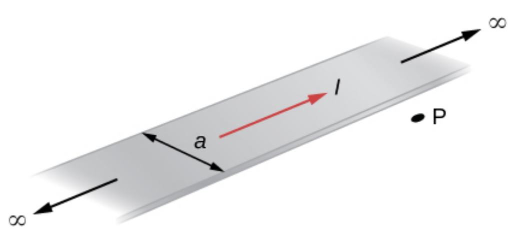


87. Magnetic field inside a torus. Consider a torus of rectangular cross-section with inner radius a and outer radius b. N turns of an insulated thin wire are wound evenly on the torus tightly all around the torus and connected to a battery producing a steady current I in the wire. Assume that the current 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 **r** from the axis.

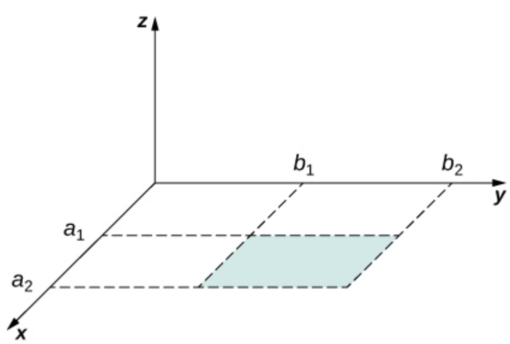
88. Two long coaxial copper tubes, each of length **L**, are connected to a battery of voltage **V**. The inner tube has inner radius **a** and outer radius **b**, and the outer tube has inner radius c and outer radius **d**. The tubes are then disconnected from the battery and rotated in the same direction 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 **r<a**, and (b) at a point between the tubes **b<r<c**, and (c) at a point outside the tubes **r>d**. (Hint: Think of copper tubes as a capacitor and find the charge density based on the voltage applied, **Q=VC,** \(\displaystyle C=\frac{2πε_0L}{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→0**.



90. A hypothetical current flowing in the z-direction creates the field $vecB = C[(x/y^2)\hat{i} + (1/y)\hat{j}]$ 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.

8.13: 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. $A \cdot m^2 \cdot T = A \cdot m^2 . \dfrac{N}{A \cdot m} = N \cdot m$

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. $2.50 \times 10^{-}2m$;

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. $9.8 \times 10^{-5} \hat{j}T$; 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. $0.31 Am^2$;

b. 0.16 Nm

87. $0.024 Am^2$

89. a. $0.16 Am^2$;

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. $R = (mvsin\theta)/qB; p = (\frac{2\pi m}{eB})vcos\theta$

105. $IaL^2/2$

107. $m = \frac{qB_0^2}{8V_{acc}}x^2$

109. 0.01 N

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).

8.14: 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 with 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. $B = \frac{\mu_o I}{8}(\frac{1}{a} - \frac{1}{b})$ 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, $B = \frac{3\mu_o I}{8\pi a}$ 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. $B = \dfrac{\mu_o I a}{2\pi b^2}$ into the page

35. 0.019 m

37. $6.28 \times 10^{-5} T$

39. $B = \dfrac{\mu_o I R^2}{\left(\left(\frac{d}{2}\right)^2 + R^2\right)^{3/2}}$

41. a. $\mu 0 I$;

b. 0;

c. $\mu 0 I$;

d. 0

43. a. $3\mu_0 I$;

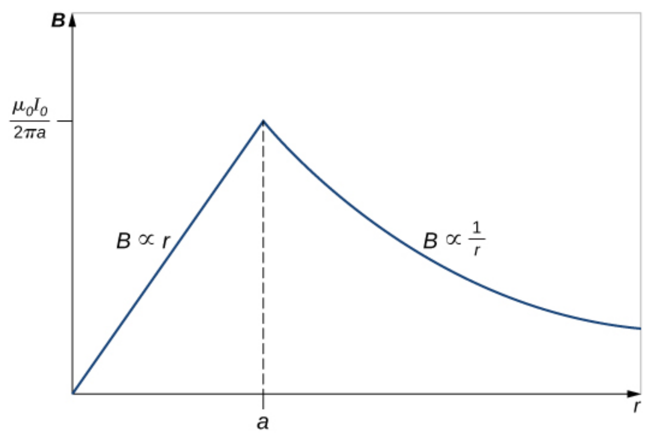
b. 0;

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. $B = \dfrac{1}{2}\mu_0 n I$

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

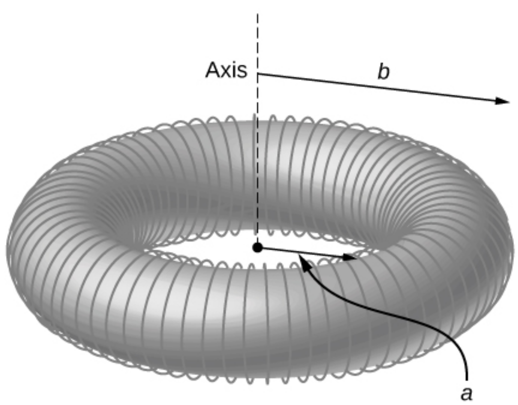
81. derivation

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. $B = \frac{\mu_0 \sigma \omega}{2}[\frac{2h^2 + R^2}{\sqrt{R^2 + h^0}}]$;

b. $B = 4.09 \times 10^{-5} T$, 82% of Earth's magnetic field

CHAPTER OVERVIEW

9: Electromagnetic Induction

9.1: Introduction

The apparatus used by Faraday to demonstrate that magnetic fields can create currents is illustrated in Figure $9.1.1$. When the switch is closed, a magnetic field is produced in the coil on the top part of the iron ring and transmitted to the coil on the bottom part of the ring. The galvanometer is used to detect any current induced in the coil on the bottom. It was found that each time the switch is closed, the galvanometer detects a current in one direction in the coil on the bottom. (You can also observe this in a physics lab.) Each time the switch is opened, the galvanometer detects a current in the opposite direction. Interestingly, if the switch remains closed or open for any length of time, there is no current through the galvanometer. *Closing and opening the switch* induces the current. It is the *change* in magnetic field that creates the current. More basic than the current that flows is the **source voltage** that causes it. The current is a result of an *source voltage induced by a changing magnetic field*, whether or not there is a path for current to flow.

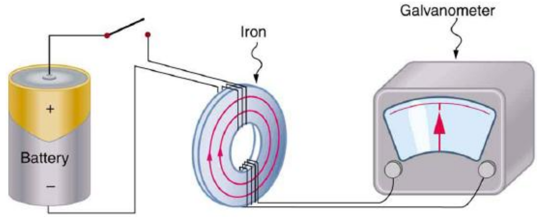


Figure $9.1.1$: Faraday's apparatus for demonstrating that a magnetic field can produce a current. A change in the field produced by the top coil induces a source voltage and, hence, a current in the bottom coil. When the switch is opened and closed, the galvanometer registers currents in opposite directions. No current flows through the galvanometer when the switch remains closed or open.

An experiment easily performed and often done in physics labs is illustrated in Figure $9.1.2$. A source voltage is induced in the coil when a bar magnet is pushed in and out of it. Source voltages of opposite signs are produced by motion in opposite directions, and the source voltages are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet —it is the relative motion that is important. The faster the motion, the greater the source voltage, and there is no source voltage when the magnet is stationary relative to the coil.

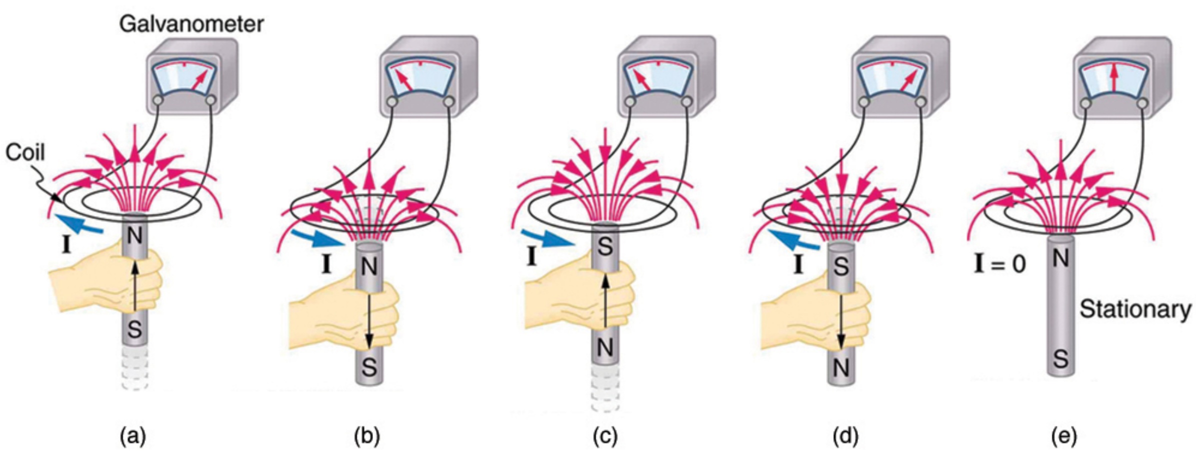


Figure $9.1.2$: Movement of a magnet relative to a coil produces source voltages as shown. The same source voltages are produced if the coil is moved relative to the magnet. The greater the speed, the greater the magnitude of the source voltage, and the source voltage is zero when there is no motion.

9.2: Magnetic Flux

Learning Objectives

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

- Calculate the flux of a uniform magnetic field through a loop of arbitrary orientation.
- Describe methods to produce a source voltage with a magnetic field or magnet and a loop of wire.

The method of inducing an source voltage used in most electric generators is shown in Figure $9.2.1$. A coil is rotated in a magnetic field, producing an alternating current source voltage, which depends on rotation rate and other factors that will be explored in later sections. Note that the generator is remarkably similar in construction to a motor (another symmetry).

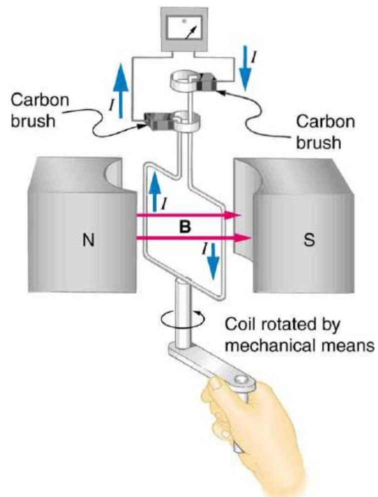


Figure $9.2.1$: Rotation of a coil in a magnetic field produces a source voltage. This is the basic construction of a generator, where work done to turn the coil is converted to electric energy. Note the generator is very similar in construction to a motor.

So we see that changing the magnitude or direction of a magnetic field produces a source voltage. Experiments revealed that there is a crucial quantity called the **magnetic flux**, Φ, given by

$$\Phi = BA\cos\theta, \tag{9.2.1}$$

where B is the magnetic field strength over an area A, at an angle θ with the perpendicular to the area as shown in Figure $9.2.2$.

Any change in magnetic flux Φ induces an source voltage. This process is defined to be **electromagnetic induction**. Units of magnetic flux Φ are $T \cdot m^2$. As seen in Figure $\boxed{\text{\PageIndex{2}}}$, $B\cos\theta = B_\perp$, which is the component of B perpendicular to the area A. Thus magnetic flux is $\Phi = B_\perp A$, the product of the area and the component of the magnetic field perpendicular to it.

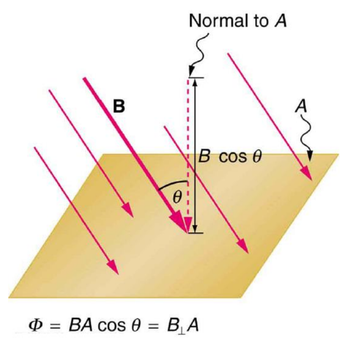


Figure $9.2.2$: Magnetic flux Φ is related to the magnetic field and the area over which it exists. The flux $\Phi = BA\cos\theta$ is related to induction; any change in Φ induces a source voltage.

All induction, including the examples given so far, arises from some change in magnetic flux Φ. For example, Faraday changed B and hence Φ when opening and closing the switch in his apparatus. This is also true for the bar magnet and coil. When rotating the coil of a generator, the angle θ and, hence, Φ is changed. Just how great a source voltage and what direction it takes depend on the change in Φ and how rapidly the change is made, as examined in the next section.

9.3: Faraday's Law

Learning Objectives

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

- Determine the magnetic flux through a surface, knowing the strength of the magnetic field, the surface area, and the angle between the normal to the surface and the magnetic field
- Use Faraday's law to determine the magnitude of induced emf in a closed loop due to changing magnetic flux through the loop

The first productive experiments concerning the effects of time-varying magnetic fields were performed by Michael **Faraday** in 1831. One of his early experiments is represented in Figure $9.3.1$. An **emf** is induced when the magnetic field in the coil is changed by pushing a bar magnet into or out of the coil. Emfs of opposite signs are produced by motion in opposite directions, and the directions of emfs are also reversed by reversing poles. The same results are produced if the coil is moved rather than the magnet—it is the relative motion that is important. The faster the motion, the greater the emf, and there is no emf when the magnet is stationary relative to the coil.

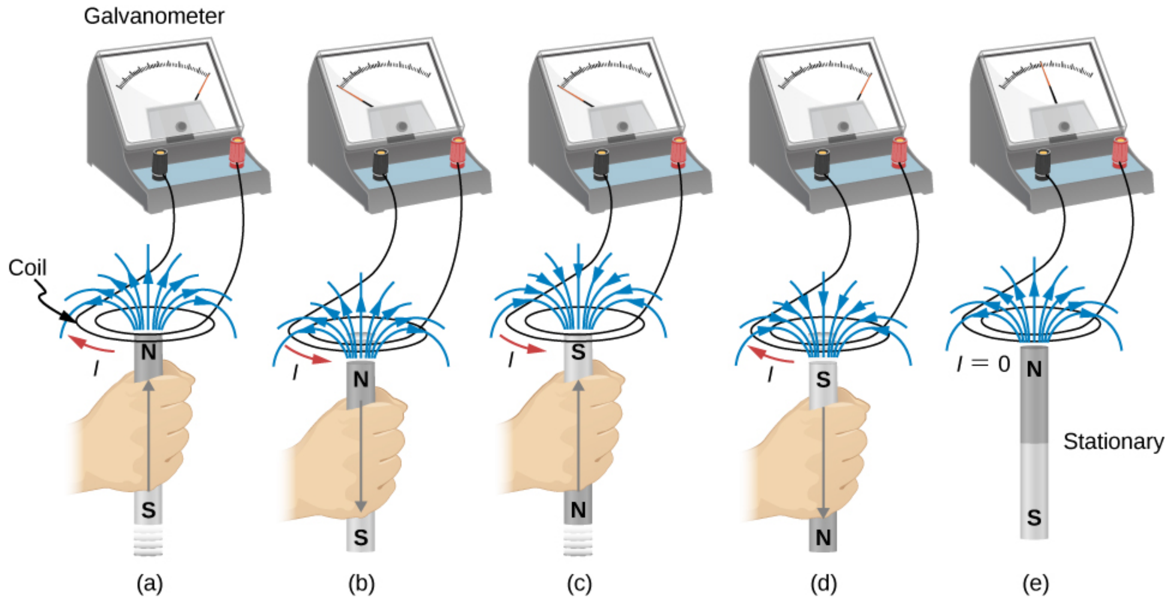


Figure $9.3.1$: Movement of a magnet relative to a coil produces emfs as shown (a–d). The same emfs are produced if the coil is moved relative to the magnet. This short-lived emf is only present during the motion. The greater the speed, the greater the magnitude of the emf, and the emf is zero when there is no motion, as shown in (e).

Faraday also discovered that a similar effect can be produced using two circuits—a changing current in one circuit induces a current in a second, nearby circuit. For example, when the switch is closed in circuit 1 of Figure $9.3.1a$, the ammeter needle of circuit 2 momentarily deflects, indicating that a short-lived current surge has been induced in that circuit. The ammeter needle quickly returns to its original position, where it remains. However, if the switch of circuit 1 is now suddenly opened, another short-lived current surge in the direction opposite from before is observed in circuit 2.

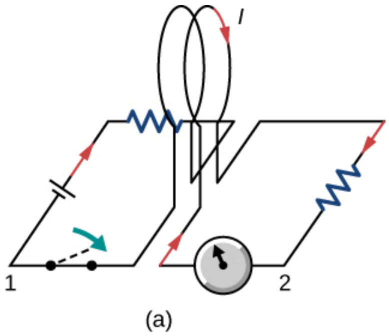


Figure 9.3.2: (a) Closing the switch of circuit 1 produces a short-lived current surge in circuit 2. (b) If the switch remains closed, no current is observed in circuit 2. (c) Opening the switch again produces a short-lived current in circuit 2 but in the opposite direction from before.

Faraday realized that in both experiments, a current flowed in the circuit containing the ammeter only when the magnetic field in the region occupied by that circuit was **changing**. As the magnet of the figure was moved, the strength of its magnetic field at the loop changed; and when the current in circuit 1 was turned on or off, the strength of its magnetic field at circuit 2 changed. Faraday was eventually able to interpret these and all other experiments involving magnetic fields that vary with time in terms of the following law.

Faraday's Law

The emf ϵ induced is the negative change in the magnetic flux Φ_m per unit time. Any change in the magnetic field or change in orientation of the area of the coil with respect to the magnetic field induces a voltage (emf).

The magnetic flux is a measurement of the amount of magnetic field lines through a given surface area, as seen in Figure 9.3.3. This definition is similar to the electric flux studied earlier. This means that if we have

$$\Phi_m = \int_S \vec{B} \cdot \hat{n} dA, \tag{9.3.1}$$

then the **induced emf** or the voltage generated by a conductor or coil moving in a magnetic field is

$$\epsilon = -\frac{d}{dt} \int_S \vec{B} \cdot \hat{n} dA = -\frac{d\Phi_m}{dt}. \tag{9.3.2}$$

The negative sign describes the direction in which the induced emf drives current around a circuit. However, that direction is most easily determined with a rule known as Lenz's law, which we will discuss shortly.

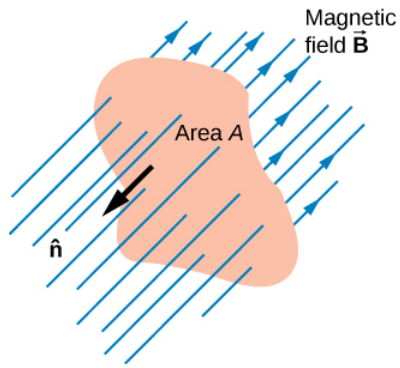


Figure 9.3.3: The magnetic flux is the amount of magnetic field lines cutting through a surface area **A** defined by the unit area vector $\hat{n}$. If the angle between the unit area $\hat{n}$ and magnetic field vector $\vec{B}$ are parallel or antiparallel, as shown in the diagram, the magnetic flux is the highest possible value given the values of area and magnetic field.

$9.3.1a$ depicts a circuit and an arbitrary surface **S** that it bounds. Notice that **S** is an **open surface**. It can be shown that **any** open surface bounded by the circuit in question can be used to evaluate Φ_m. For example, Φ_m is the same for the various surfaces $S_1, S_2, ...$ of part (b) of the figure.

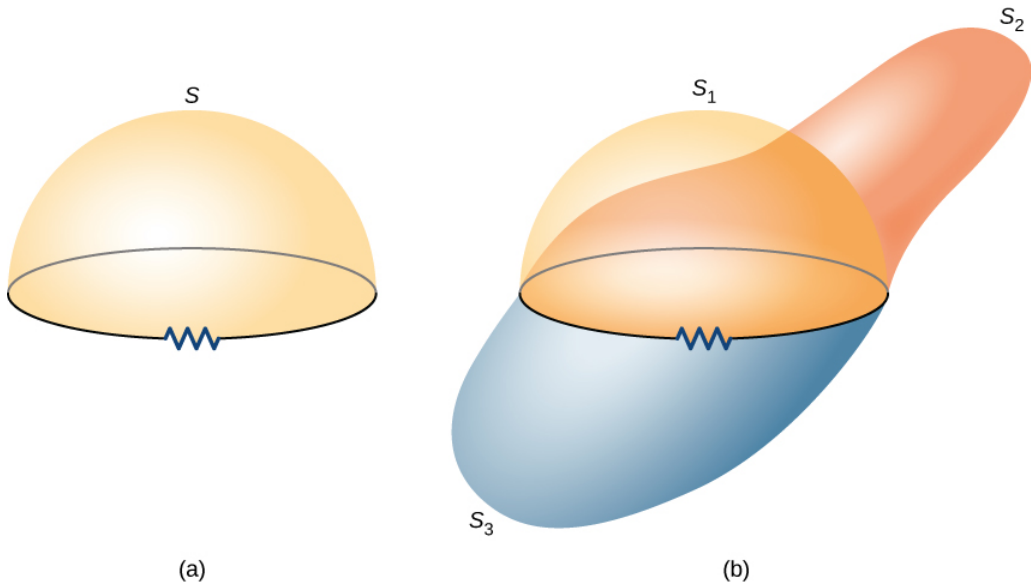


Figure 9.3.4: (a) A circuit bounding an arbitrary open surface **S**. The planar area bounded by the circuit is not part of **S**. (b) Three arbitrary open surfaces bounded by the same circuit. The value of Φ_m is the same for all these surfaces.

The SI unit for magnetic flux is the weber (Wb),

$$1\,Wb = 1\,T\cdot m^2. \tag{9.3.3}$$

Occasionally, the magnetic field unit is expressed as webers per square meter (Wb/m^2) instead of teslas, based on this definition. In many practical applications, the circuit of interest consists of a number **N** of tightly wound turns (Figure 9.3.5). Each turn experiences the same magnetic flux. Therefore, the net magnetic flux through the circuits is **N** times the flux through one turn, and Faraday's law is written as

$$\epsilon = -\frac{d}{dt}(N\Phi_m) = -N\frac{d\Phi_m}{dt}. \tag{9.3.4}$$

A Square Coil in a Changing Magnetic Field

The square coil of Figure 9.3.1 has sides $l = 0.25\,m$ long and is tightly wound with $N = 200$ turns of wire. The resistance of the coil is $R = 5.0\,\Omega$ The coil is placed in a spatially uniform magnetic field that is directed perpendicular to the face of the coil and whose magnitude is decreasing at a rate $dB/dt = -0.040\,T/s$. (a) What is the magnitude of the emf induced in the coil? (b) What is the magnitude of the current circulating through the coil?

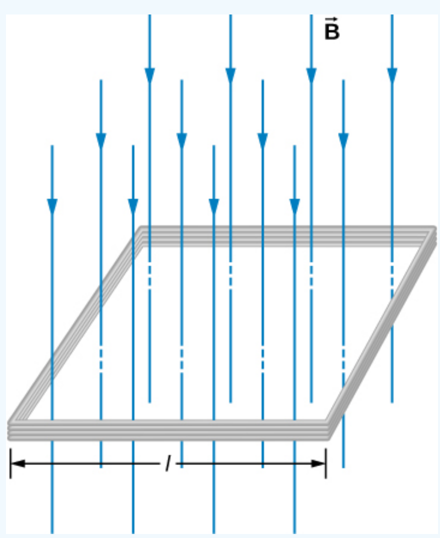


Figure 9.3.5: A square coil with **N** turns of wire with uniform magnetic field $\vec{B}$ directed in the downward direction, perpendicular to the coil.

Strategy

The area vector, or $\hat{n}$ direction, is perpendicular to area covering the loop. We will choose this to be pointing downward so that $\vec{B}$ is parallel to $\hat{n}$ and that the flux turns into multiplication of magnetic field times area. The area of the loop is not changing in time, so it can be factored out of the time derivative, leaving the magnetic field as the only quantity varying in time. Lastly, we can apply Ohm's law once we know the induced emf to find the current in the loop.

Solution

1. The flux through one turn is

$$\Phi_m = BA = Bt^2, \tag{9.3.5}$$

so we can calculate the magnitude of the emf from Faraday's law. The sign of the emf will be discussed in the next section, on Lenz's law:

$$|\epsilon| = \left|-N\frac{d\Phi_m}{dt}\right| = Nl^2\frac{dB}{dt} \tag{9.3.6}$$

$$= (200)(0.25\,m)^2(0.040\,T/s) = 0.50\,V. \tag{9.3.7}$$

- The magnitude of the current induced in the coil is

$$I = \frac{\epsilon}{R} = \frac{0.50\,V}{5.0\,\Omega} = 0.10\,A. \tag{9.3.8}$$

Significance

If the area of the loop were changing in time, we would not be able to pull it out of the time derivative. Since the loop is a closed path, the result of this current would be a small amount of heating of the wires until the magnetic field stops changing. This may increase the area of the loop slightly as the wires are heated.

? Exercise 9.3.1

A closely wound coil has a radius of 4.0 cm, 50 turns, and a total resistance of $40\,\Omega$. At what rate must a magnetic field perpendicular to the face of the coil change in order to produce Joule heating in the coil at a rate of 2.0 mW?

Solution

1.1 T/s

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).

9.4: Lenz's Law

Learning Objectives

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

- Use Lenz's law to determine the direction of induced emf whenever a magnetic flux changes
- Use Faraday's law with Lenz's law to determine the induced emf in a coil and in a solenoid

The direction in which the induced emf drives current around a wire loop can be found through the negative sign. However, it is usually easier to determine this direction with Lenz's law, named in honor of its discoverer, Heinrich Lenz (1804–1865). (Faraday also discovered this law, independently of Lenz.) We state Lenz's law as follows:

Lenz's Law

The direction of the induced emf drives current around a wire loop to always **oppose** the change in magnetic flux that causes the emf.

Lenz's law can also be considered in terms of conservation of energy. If pushing a magnet into a coil causes current, the energy in that current must have come from somewhere. If the induced current causes a magnetic field opposing the increase in field of the magnet we pushed in, then the situation is clear. We pushed a magnet against a field and did work on the system, and that showed up as current. If it were not the case that the induced field opposes the change in the flux, the magnet would be pulled in produce a current without anything having done work. Electric potential energy would have been created, violating the conservation of energy.

To determine an induced emf ϵ, you first calculate the magnetic flux Φ_m and then obtain $d\Phi_m/dt$. The magnitude of ϵ is given by

$$\epsilon = \left|\frac{d\Phi_m}{dt}\right|.$$

Finally, you can apply Lenz's law to determine the sense of ϵ. This will be developed through examples that illustrate the following problem-solving strategy.

Problem-Solving Strategy: Lenz's Law

To use Lenz's law to determine the directions of induced magnetic fields, currents, and emfs:

- Make a sketch of the situation for use in visualizing and recording directions.
- Determine the direction of the applied magnetic field $\vec{B}$.
- Determine whether its magnetic flux is increasing or decreasing.
- Now determine the direction of the induced magnetic field $\vec{B}$. The induced magnetic field tries to reinforce a magnetic flux that is decreasing or opposes a magnetic flux that is increasing. Therefore, the induced magnetic field adds or subtracts to the applied magnetic field, depending on the change in magnetic flux.
- Use **right-hand rule 2** (RHR-2; see Magnetic Forces and Fields) to determine the direction of the induced current **I** that is responsible for the induced magnetic field $\vec{B}$.
- The direction (or polarity) of the induced emf can now drive a conventional current in this direction.

Let's apply Lenz's law to the system of Figure $9.4.1a$. We designate the "front" of the closed conducting loop as the region containing the approaching bar magnet, and the "back" of the loop as the other region. As the north pole of the magnet moves toward the loop, the flux through the loop due to the field of the magnet increases because the strength of field lines directed from the front to the back of the loop is increasing. A current is therefore induced in the loop. By Lenz's law, the direction of the induced current must be such that its own magnetic field is directed in a way to **oppose** the changing flux caused by the field of the approaching magnet. Hence, the induced current circulates so that its magnetic field lines through the loop are directed from the back to the front of the loop. By RHR-2, place your thumb pointing against the magnetic field lines, which is toward the bar magnet. Your fingers wrap in a counterclockwise direction as viewed from the bar magnet. Alternatively, we can determine the direction of the induced current by treating the current loop as an electromagnet that **opposes** the approach of the north pole of the bar magnet. This occurs when the induced current flows as shown, for then the face of the loop nearer the approaching magnet is also a north pole.

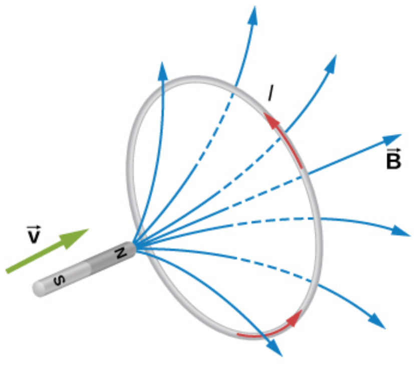


Figure 9.4.1: The change in magnetic flux caused by the approaching magnet induces a current in the loop. (a) An approaching north pole induces a counterclockwise current with respect to the bar magnet. (b) An approaching south pole induces a clockwise current with respect to the bar magnet.

Part (b) of the figure shows the south pole of a magnet moving toward a conducting loop. In this case, the flux through the loop due to the field of the magnet increases because the number of field lines directed from the back to the front of the loop is increasing. To oppose this change, a current is induced in the loop whose field lines through the loop are directed from the front to the back. Equivalently, we can say that the current flows in a direction so that the face of the loop nearer the approaching magnet is a south pole, which then repels the approaching south pole of the magnet. By RHR-2, your thumb points away from the bar magnet. Your fingers wrap in a clockwise fashion, which is the direction of the induced current.

Another example illustrating the use of Lenz's law is shown in Figure $9.4.2$. When the switch is opened, the decrease in current through the solenoid causes a decrease in magnetic flux through its coils, which induces an emf in the solenoid. This emf must oppose the change (the termination of the current) causing it. Consequently, the induced emf has the polarity shown and drives in the direction of the original current. This may generate an arc across the terminals of the switch as it is opened.

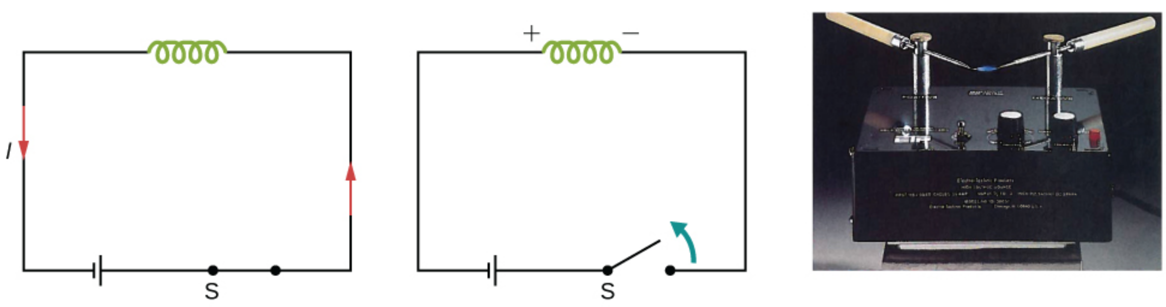


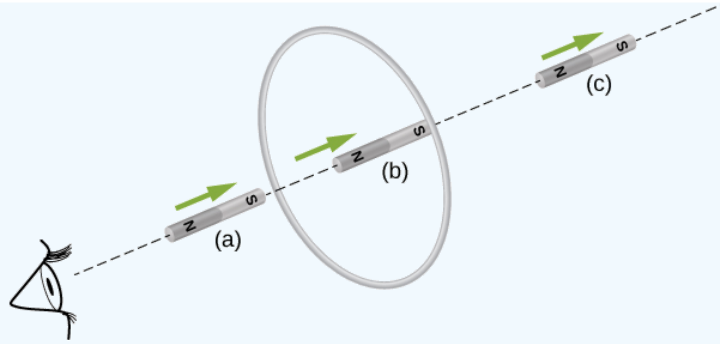
Figure 9.4.2: (a) A solenoid connected to a source of emf. (b) Opening switch S terminates the current, which in turn induces an emf in the solenoid. (c) A potential difference between the ends of the sharply pointed rods is produced by inducing an emf in a coil. This potential difference is large enough to produce an arc between the sharp points.

? Exercise $9.4.1A$

Find the direction of the induced current in the wire loop shown below as the magnet enters, passes through, and leaves the loop.

Solution

To the observer shown, the current flows clockwise as the magnet approaches, decreases to zero when the magnet is centered in the plane of the coil, and then flows counterclockwise as the magnet leaves the coil.



? Exercise $9.4.1B$

Verify the directions of the induced currents in Figure 13.2.2.

✓ Example $9.4.1A$: A Circular Coil in a Changing Magnetic Field

A magnetic field $\vec{B}$ is directed outward perpendicular to the plane of a circular coil of radius $r = 0.50\, m$ (Figure $9.4.3$). The field is cylindrically symmetrical with respect to the center of the coil, and its magnitude decays exponentially according to $B = (1.5T)e^{(5.0s^{-1})t}$, where **B** is in teslas and **t** is in seconds.

a. Calculate the emf induced in the coil at the times $t_1 = 0$, $t_2 = 5.0 \times 10^{-2}\, s$, and $t_3 = 1.0\, s$.
b. Determine the current in the coil at these three times if its resistance is $10\, \Omega$.

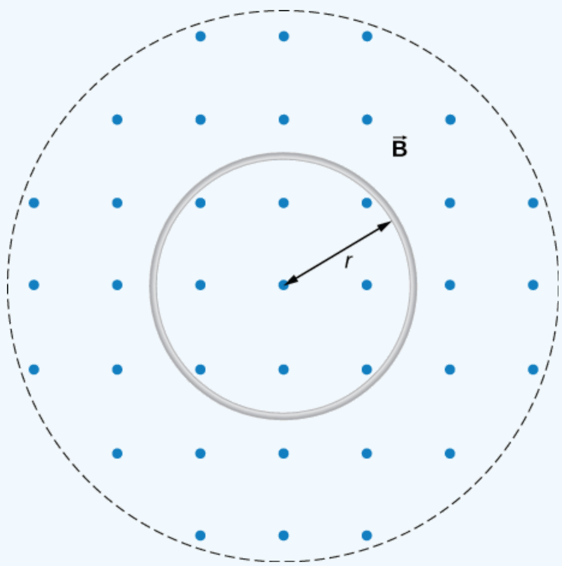


Figure $9.4.3$: A circular coil in a decreasing magnetic field.

Strategy

Since the magnetic field is perpendicular to the plane of the coil and constant over each spot in the coil, the dot product of the magnetic field $\vec{B}$ and normal to the area unit vector $\hat{n}$ turns into a multiplication. The magnetic field can be pulled out of the integration, leaving the flux as the product of the magnetic field times area. We need to take the time derivative of the exponential function to calculate the emf using Faraday's law. Then we use Ohm's law to calculate the current.

Solution

1. Since $\vec{B}$ is perpendicular to the plane of the coil, the magnetic flux is given by

$$\Phi_m = B\pi r^2 = (1.5e^{-5.0t}T)\pi(0.50\, m)^2$$

$$= 1.2e^{-(5.0s^{-1})t}Wb.$$

From Faraday’s law, the magnitude of the induced emf is

$$\epsilon = \left|\frac{d\Phi_m}{dt}\right| = \left|\frac{d}{dt}(1.2e^{-(5.0s^{-1})t}Wb)\right| = 6.0e^{-(5.0s^{-1})t}V.$$

Since $\vec{B}$ is directed out of the page and is decreasing, the induced current must flow counterclockwise when viewed from above so that the magnetic field it produces through the coil also points out of the page. For all three times, the sense of ε is counterclockwise; its magnitudes are

$$\epsilon(t_1) = 6.0V;\ \epsilon(t_2) = 4.7\,V;\ \epsilon(t_3) = 0040\,V.$$

2. From Ohm’s law, the respective currents are

$$I(t_1) = \frac{\epsilon(t_1)}{R} = \frac{6.0\,V}{10\,\Omega} = 0.60\,A;$$
$$I(t_2) = \frac{4.7\,V}{10\,\Omega} = 0.47\,A;$$

and

$$I(t_3) = \frac{0.040\,V}{10\,\Omega} = 4.0\times 10^{-3}\,A.$$

Significance

An emf voltage is created by a changing magnetic flux over time. If we know how the magnetic field varies with time over a constant area, we can take its time derivative to calculate the induced emf.

Example $9.4.1B$: Changing Magnetic Field Inside a Solenoid

The current through the windings of a solenoid with $n = 2000$ turns per meter is changing at a rate $dI/dt = 3.0\,A/s$. (See Sources of Magnetic Fields for a discussion of solenoids.) The solenoid is 50-cm long and has a cross-sectional diameter of 3.0 cm. A small coil consisting of $N = 20$ closely wound turns wrapped in a circle of diameter 1.0 cm is placed in the middle of the solenoid such that the plane of the coil is perpendicular to the central axis of the solenoid. Assuming that the infinite-solenoid approximation is valid at the location of the small coil, determine the magnitude of the emf induced in the coil.

Strategy

The magnetic field in the middle of the solenoid is a uniform value of $\mu_0 nI$. This field is producing a maximum magnetic flux through the coil as it is directed along the length of the solenoid. Therefore, the magnetic flux through the coil is the product of the solenoid’s magnetic field times the area of the coil. Faraday’s law involves a time derivative of the magnetic flux. The only quantity varying in time is the current, the rest can be pulled out of the time derivative. Lastly, we include the number of turns in the coil to determine the induced emf in the coil.

Solution

Since the field of the solenoid is given by $B = \mu_0 nI$, the flux through each turn of the small coil is

$$\Phi_m = \mu_0 nI\left(\frac{\pi d^2}{4}\right),$$

where d is the diameter of the coil. Now from Faraday’s law, the magnitude of the emf induced in the coil is

$$\begin{aligned}\epsilon &= \left|N\frac{d\Phi_m}{dt}\right| \\ &= \left|N\mu_0 n\frac{\pi d^2}{4}\frac{dI}{dt}\right| \\ &= 20(4\pi\times 10^{-7}T\cdot m/s)(2000\,m^{-1})\frac{\pi(0.010\,m)^2}{4}(3.0\,A/s) \\ &= 1.2\times 10^{-5}\,V.\end{aligned}$$

Significance

When the current is turned on in a vertical solenoid, as shown in Figure 9.4.4, the ring has an induced emf from the solenoid's changing magnetic flux that opposes the change. The result is that the ring is fired vertically into the air.

Figure 9.4.4: The jumping ring. When a current is turned on in the vertical solenoid, a current is induced in the metal ring. The stray field produced by the solenoid causes the ring to jump off the solenoid.

Note

A demonstration of the jumping ring from MIT.

9.5: Motional Source Voltage

Learning Objectives

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

- Determine the magnitude of an induced source voltage in a wire moving at a constant speed through a magnetic field.
- Discuss examples that use motional source voltage, such as a rail gun and a tethered satellite

Magnetic flux depends on three factors: the strength of the magnetic field, the area through which the field lines pass, and the orientation of the field with the surface area. If any of these quantities varies, a corresponding variation in magnetic flux occurs. So far, we've only considered flux changes due to a changing field. Now we look at another possibility: a changing area through which the field lines pass including a change in the orientation of the area.

Two examples of this type of flux change are represented in Figure $9.5.1$. In part (a), the flux through the rectangular loop increases as it moves into the magnetic field, and in part (b), the flux through the rotating coil varies with the angle θ.

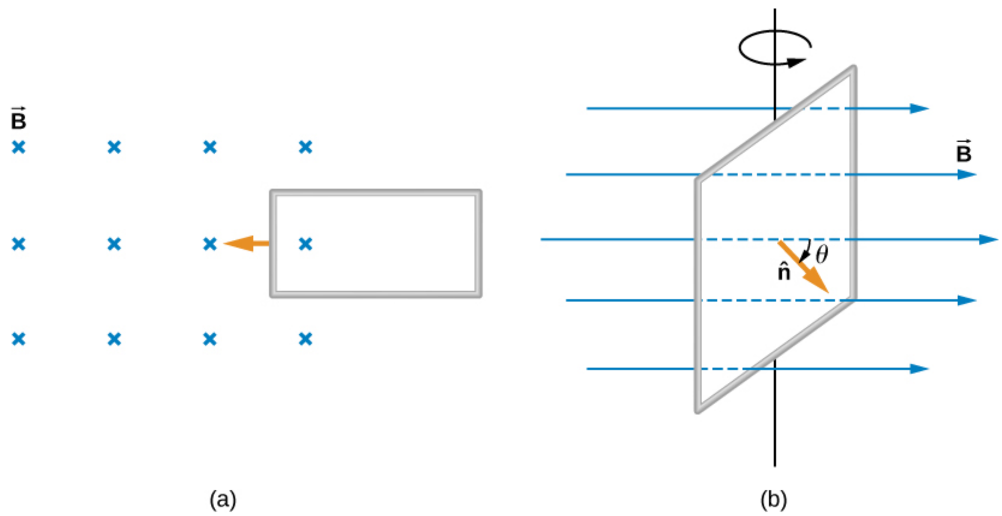


Figure $9.5.1$: (a) Magnetic flux changes as a loop moves into a magnetic field; (b) magnetic flux changes as a loop rotates in a magnetic field.

It's interesting to note that what we perceive as the cause of a particular flux change actually depends on the frame of reference we choose. For example, if you are at rest relative to the moving coils of Figure $9.5.1b$, you would see the flux vary because of a changing magnetic field—in part (a), the field moves from left to right in your reference frame, and in part (b), the field is rotating. It is often possible to describe a flux change through a coil that is moving in one particular reference frame in terms of a changing magnetic field in a second frame, where the coil is stationary. However, reference-frame questions related to magnetic flux are beyond the level of this textbook. We'll avoid such complexities by always working in a frame at rest relative to the laboratory and explain flux variations as due to either a changing field or a changing area.

Now let's look at a conducting rod pulled in a circuit, changing magnetic flux. The area enclosed by the circuit 'MNOP' of Figure $9.5.2$ is **lx** and is perpendicular to the magnetic field, so we can simplify the integration of $\Phi_\text{m} = \int_S \vec{\mathbf{B}} \cdot \hat{\mathbf{n}} dA$ into a multiplication of magnetic field and area. The magnetic flux through the open surface is therefore

$$\Phi_m = Blx. \tag{9.5.1}$$

Since **B** and **l** are constant and the velocity of the rod is $v = dx/dt$, we can now restate Faraday's law, Equation 13.2.2, for the magnitude of the source voltage in terms of the moving conducting rod as

$$\epsilon = \frac{d\Phi_m}{dt} = Bl\frac{dx}{dt} = Blv. \tag{9.5.2}$$

The current induced in the circuit is the source voltage divided by the resistance or

$$I = \frac{Blv}{R}. \tag{9.5.3}$$

Furthermore, the direction of the induced source satisfies Lenz's law, as you can verify by inspection of the figure.

This calculation of motionally induced source voltage is not restricted to a rod moving on conducting rails. With $\vec{F} = q\vec{v} \times \vec{B}$ as the starting point, it can be shown that $\epsilon = -d\Phi_m/dt$ holds for any change in flux caused by the motion of a conductor. We saw in Faraday's Law that the source voltage induced by a time-varying magnetic field obeys this same relationship, which is Faraday's law. Thus Faraday's law **holds for all flux changes**, whether they are produced by a changing magnetic field, by motion, or by a combination of the two.

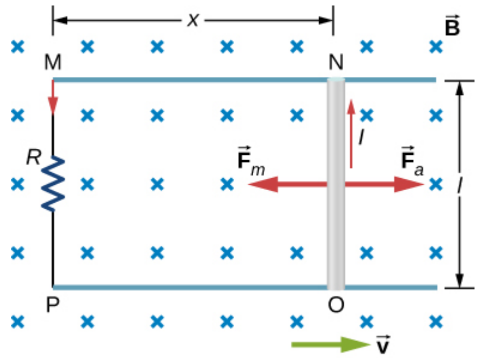


Figure $9.5.2$: A conducting rod is pushed to the right at constant velocity. The resulting change in the magnetic flux induces a current in the circuit.

From an energy perspective, $\vec{F}_a$ produces power $F_a v$, and the resistor dissipates power I^2R. Since the rod is moving at constant velocity, the applied force F_a must balance the magnetic force $F_m = IlB$ on the rod when it is carrying the induced current **I**. Thus the power produced is

$$F_a v = IlBv = \frac{Blv}{R} \cdot lBv = \frac{l^2B^2v^2}{R}. \tag{9.5.4}$$

The power dissipated is

$$P = I^2R = \left(\frac{Blv}{R}\right)^2 R = \frac{l^2B^2v^2}{R}. \tag{9.5.5}$$

In satisfying the principle of energy conservation, the produced and dissipated powers are equal.

This principle can be seen in the operation of a rail gun. A **rail gun** is an electromagnetic projectile launcher that uses an apparatus similar to Figure $9.5.2$ and is shown in schematic form in Figure $9.5.3$. The conducting rod is replaced with a projectile or weapon to be fired. So far, we've only heard about how motion causes a source voltage. In a rail gun, the optimal shutting off/ramping down of a magnetic field decreases the flux in between the rails, causing a current to flow in the rod (armature) that holds the projectile. This current through the armature experiences a magnetic force and is propelled forward. Rail guns, however, are not used widely in the military due to the high cost of production and high currents: Nearly one million amps is required to produce enough energy for a rail gun to be an effective weapon.

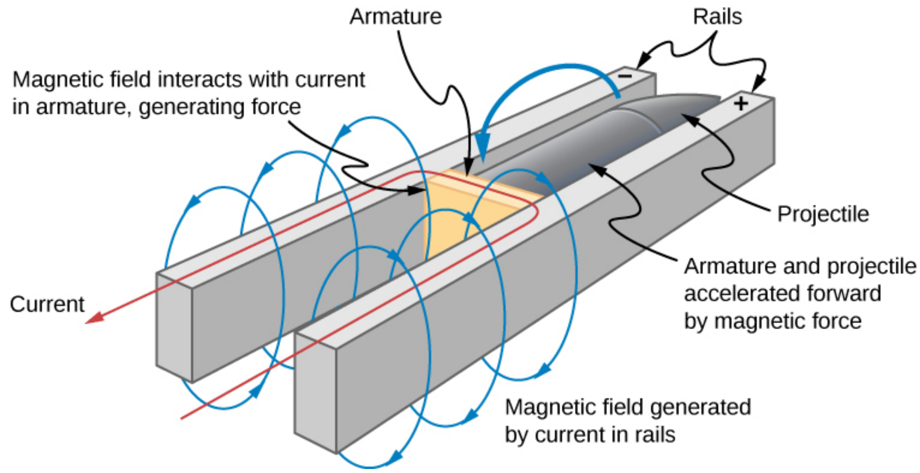


Figure 9.5.3: Current through two rails drives a conductive projectile forward by the magnetic force created.

We can calculate a motionally induced source voltage with Faraday's law **even when an actual closed circuit is not present**. We simply imagine an enclosed area whose boundary includes the moving conductor, calculate Φ_m, and then find the source voltage from Faraday's law. For example, we can let the moving rod of Figure 9.5.4 be one side of the imaginary rectangular area represented by the dashed lines. The area of the rectangle is **lx**, so the magnetic flux through it is $\Phi_m = Blx$. Differentiating this equation, we obtain

$$\frac{d\Phi_m}{dt} = Bl\frac{dx}{dt} = Blv, \tag{9.5.6}$$

which is identical to the potential difference between the ends of the rod that we determined earlier.

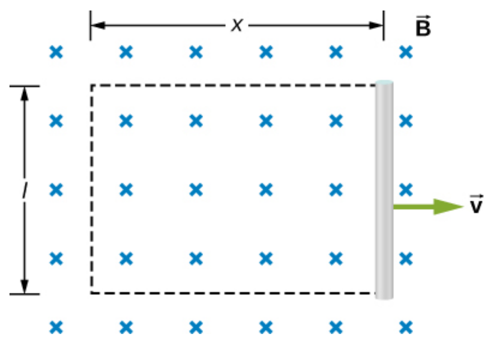


Figure 9.5.4: With the imaginary rectangle shown, we can use Faraday's law to calculate the induced source voltage in the moving rod.

Motional source voltages in Earth's weak magnetic field are not ordinarily very large, or we would notice voltage along metal rods, such as a screwdriver, during ordinary motions. For example, a simple calculation of the motional source voltage of a 1.0-m rod moving at 3.0 m/s perpendicular to the Earth's field gives

$$\text{source voltage} = Blv = (5.0 \times 10^{-5} T)(1.0\, m)(3.0\, m/s) = 150\, \mu V. \tag{9.5.7}$$

This small value is consistent with experience. There is a spectacular exception, however. In 1992 and 1996, attempts were made with the space shuttle to create large motional source voltages. The tethered satellite was to be let out on a 20-km length of wire, as shown in Figure 9.5.5, to create a 5-kV source voltage by moving at orbital speed through Earth's field. This source voltage could be used to convert some of the shuttle's kinetic and potential energy into electrical energy if a complete circuit could be made. To complete the circuit, the stationary ionosphere was to supply a return path through which current could flow. (The ionosphere is the rarefied and partially ionized atmosphere at orbital altitudes. It conducts because of the ionization. The ionosphere serves the same function as the stationary rails and connecting resistor in Figure 9.5.3, without which there would not be a complete circuit.) Drag on the current in the cable due to the magnetic force $F = IlB\sin\theta$ does the work that reduces the shuttle's kinetic and potential energy, and allows it to be converted into electrical energy. Both tests were unsuccessful. In the first, the cable hung up and could only be extended a couple of hundred meters; in the second, the cable broke when almost fully extended. Example 9.5.1 indicates feasibility in principle.

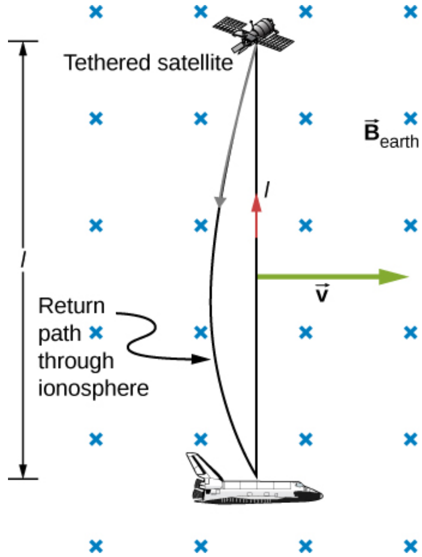


Figure 9.5.5: Motional source voltage as electrical power conversion for the space shuttle was the motivation for the tethered satellite experiment. A 5-kV source voltage was predicted to be induced in the 20-km tether while moving at orbital speed in Earth's magnetic field. The circuit is completed by a return path through the stationary ionosphere.

Example 9.5.1: Calculating the Large Motional Source Voltage of an Object in Orbit

Calculate the motional source voltage induced along a 20.0-km conductor moving at an orbital speed of 7.80 km/s perpendicular to Earth's $5.00 \times 10^{-5} T$ magnetic field.

Strategy

This is a great example of using the equation motional $\epsilon = Blv$.

Solution

Entering the given values into $\epsilon = Blv$ gives

$$\epsilon = Blv \tag{9.5.8}$$

$$= (5.00 \times 10^{-5} T)(2.00 \times 10^4 \, m)(7.80 \times 10^3 \, m/s) \tag{9.5.9}$$

$$= 7.80 \times 10^3 \, V. \tag{9.5.10}$$

Significance

The value obtained is greater than the 5-kV measured voltage for the shuttle experiment, since the actual orbital motion of the tether is not perpendicular to Earth's field. The 7.80-kV value is the maximum source voltage obtained when $\theta = 90^o$ and so $sin\,\theta = 1$.

Example 9.5.2: A Metal Rod Rotating in a Magnetic Field

Part (a) of Figure 9.5.6 shows a metal rod **OS** that is rotating in a horizontal plane around point **O**. The rod slides along a wire that forms a circular arc **PST** of radius **r**. The system is in a constant magnetic field $\vec{B}$ that is directed out of the page. (a) If you rotate the rod at a constant angular velocity ω, what is the current **I** in the closed loop **OPSO**? Assume that the resistor **R** furnishes all of the resistance in the closed loop. (b) Calculate the work per unit time that you do while rotating the rod and show that it is equal to the power dissipated in the resistor.

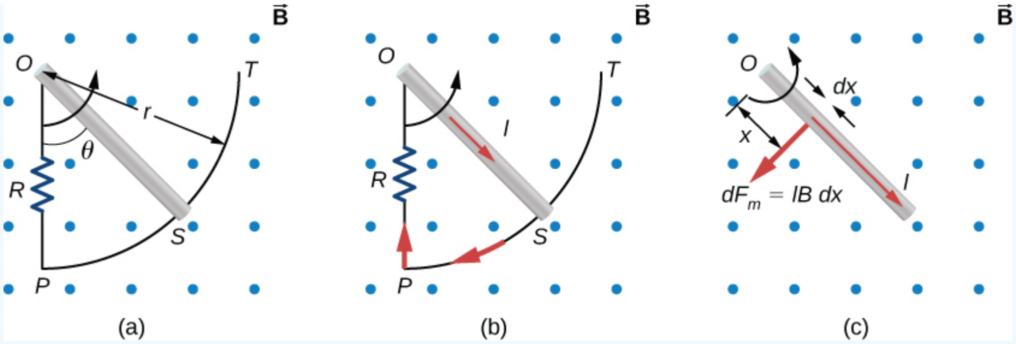


Figure 9.5.6: (a) The end of a rotating metal rod slides along a circular wire in a horizontal plane. (b) The induced current in the rod. (c) The magnetic force on an infinitesimal current segment.

Strategy

The magnetic flux is the magnetic field times the area of the quarter circle or $A = r^2\theta/2$. When finding the source voltage through Faraday's law, all variables are constant in time but θ, with $\omega = d\theta/dt$. To calculate the work per unit time, we know this is related to the torque times the angular velocity. The torque is calculated by knowing the force on a rod and integrating it over the length of the rod.

Solution

1. From geometry, the area of the loop **OPSO** is $A = \frac{r^2\theta}{2}$. Hence, the magnetic flux through the loop is

$$\Phi_m = BA = B\frac{r^2\theta}{2}. \tag{9.5.11}$$

Differentiating with respect to time and using $\omega = d\theta/dt$, we have

$$\epsilon = \left|\frac{d\Phi_m}{dt}\right| = \frac{Br^2\omega}{2}. \tag{9.5.12}$$

When divided by the resistance **R** of the loop, this yields for the magnitude of the induced current

$$I = \frac{\epsilon}{R} = \frac{Br^2\omega}{2R}. \tag{9.5.13}$$

As θ increases, so does the flux through the loop due to $\vec{B}$. To counteract this increase, the magnetic field due to the induced current must be directed into the page in the region enclosed by the loop. Therefore, as part (b) of Figure 9.5.6 illustrates, the current circulates clockwise.

2. You rotate the rod by exerting a torque on it. Since the rod rotates at constant angular velocity, this torque is equal and opposite to the torque exerted on the current in the rod by the original magnetic field. The magnetic force on the infinitesimal segment of length **dx** shown in part (c) of Figure 9.5.6 is $dF_m = IBdx$, so the magnetic torque on this segment is

$$d\tau_m = x \cdot dF_m = IBxdx. \tag{9.5.14}$$

The net magnetic torque on the rod is then

$$\tau_m = \int_0^r d\tau_m = IB\int_0^r x\,dx = \frac{1}{2}IBr^2. \tag{9.5.15}$$

The torque τ that you exert on the rod is equal and opposite to τ_m, and the work that you do when the rod rotates through an angle $d\theta$ is $dW = rd\theta$. Hence, the work per unit time that you do on the rod is

$$\frac{dW}{dt} = \tau\frac{d\theta}{dt} = \frac{1}{2}IBr^2\frac{d\theta}{dt} = \frac{1}{2}\left(\frac{Br^2\omega}{2R}\right)Br^2\omega = \frac{B^2r^4\omega^2}{4R}, \tag{9.5.16}$$

where we have substituted for **I**. The power dissipated in the resister is $P = IR^2$, which can be written as

$$P = \left(\frac{Br^2\omega}{2R}\right)^2 R = \frac{B^2r^4\omega^2}{4R}. \tag{9.5.17}$$

Therefore, we see that

$$P = \frac{dW}{dt}. \tag{9.5.18}$$

Hence, the power dissipated in the resistor is equal to the work per unit time done in rotating the rod.

Significance

An alternative way of looking at the induced source voltage from Faraday's law is to integrate in space instead of time. The solution, however, would be the same. The motional source voltage is

$$|\epsilon| = \int Bvdl. \tag{9.5.19}$$

The velocity can be written as the angular velocity times the radius and the differential length written as **dr**. Therefore,

$$\epsilon| = B\int vdr = B\omega\int_0^l rdr = \frac{1}{2}B\omega l^2, \tag{9.5.20}$$

which is the same solution as before.

Example 9.5.3: A Rectangular Coil Rotating in a Magnetic Field

A rectangular coil of area **A** and **N** turns is placed in a uniform magnetic field $\vec{B} = B\hat{j}$, as shown in Figure 9.5.7. The coil is rotated about the **z**-axis through its center at a constant angular velocity ω. Obtain an expression for the induced source voltage in the coil.

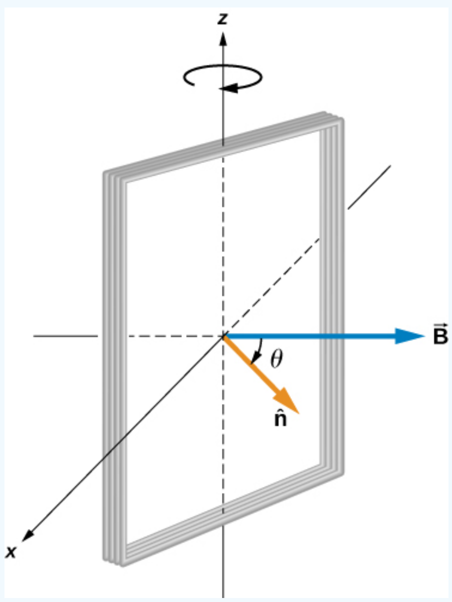


Figure 9.5.7: A rectangular coil rotating in a uniform magnetic field.

Strategy

According to the diagram, the angle between the perpendicular to the surface ((\hat{n}\)) and the magnetic field $(\vec{B})$ is θ. The dot product of $B \cdot \hat{n}$ simplifies to only the $cos\,\theta$ component of the magnetic field, namely where the magnetic field projects onto the unit area vector $\hat{n}$. The magnitude of the magnetic field and the area of the loop are fixed over time, which makes the integration simplify quickly. The induced source voltage is written out using Faraday's law.

Solution

When the coil is in a position such that its normal vector $\hat{n}$ makes an angle θ with the magnetic field $\vec{B}$ the magnetic flux through a single turn of the coil is

$$\Phi_m = \int_S \vec{B} \cdot \hat{n} dA = BA\, cos\, \theta. \tag{9.5.21}$$

From Faraday's law, the source voltage induced in the coil is

$$\epsilon = -N \frac{d\Phi_m}{dt} = NBA\, sin\, \theta \frac{d\theta}{dt}. \tag{9.5.22}$$

The constant angular velocity is $\omega = d\theta/dt$. The angle θ represents the time evolution of the angular velocity or ωt. This is changes the function to time space rather than θ. The induced source voltage therefore varies sinusoidally with time according to

$$\epsilon = \epsilon_0\, sin\, \omega t, \tag{9.5.23}$$

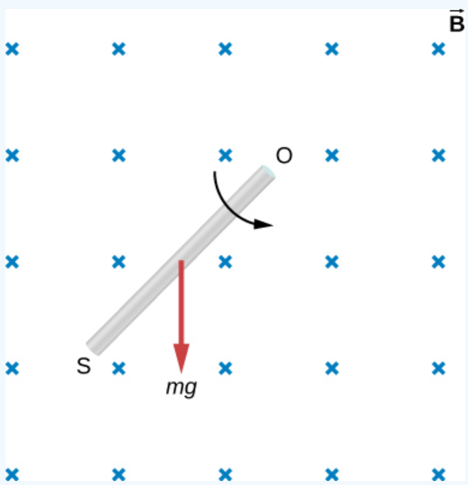
where $\epsilon_0 = NBA\omega$.

Significance

If the magnetic field strength or area of the loop were also changing over time, these variables wouldn't be able to be pulled out of the time derivative to simply the solution as shown. This example is the basis for an electric generator, as we will give a full discussion in Applications of Newton's Law.

? Exercise 9.5.1

Shown below is a rod of length **l** that is rotated counterclockwise around the axis through **O** by the torque due to $m\vec{g}$. Assuming that the rod is in a uniform magnetic field $\vec{B}$, what is the source voltage induced between the ends of the rod when its angular velocity is ω? Which end of the rod is at a higher potential?



Answer

$\epsilon = Bl^2\omega/2$, with **O** at a higher potential than **S**

? Exercise 9.5.2

A rod of length 10 cm moves at a speed of 10 m/s perpendicularly through a 1.5-T magnetic field. What is the potential difference between the ends of the rod?

Answer

1.5 V

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).

9.6: Induced Electric Fields

Learning Objectives

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

- Connect the relationship between an induced emf from Faraday's law to an electric field, thereby showing that a changing magnetic flux creates an electric field
- Solve for the electric field based on a changing magnetic flux in time

The fact that emfs are induced in circuits implies that work is being done on the conduction electrons in the wires. What can possibly be the source of this work? We know that it's neither a battery nor a magnetic field, for a battery does not have to be present in a circuit where current is induced, and magnetic fields never do work on moving charges. The answer is that the source of the work is an electric field $\vec{E}$ that is induced in the wires. The work done by $\vec{E}$ in moving a unit charge completely around a circuit is the induced emf ε; that is,

$$\epsilon = \oint \vec{E} \cdot d\vec{l},$$

where $\oint$ represents the line integral around the circuit. Faraday's law can be written in terms of the induced electric field as

$$\oint \vec{E} \cdot d\vec{l} = -\frac{d\Phi_m}{dt}.$$

There is an important distinction between the electric field induced by a changing magnetic field and the electrostatic field produced by a fixed charge distribution. Specifically, the induced electric field is nonconservative because it does net work in moving a charge over a closed path, whereas the electrostatic field is conservative and does no net work over a closed path. Hence, electric potential can be associated with the electrostatic field, but not with the induced field. The following equations represent the distinction between the two types of electric field:

$$\underbrace{\oint \vec{E} \cdot d\vec{l} \neq 0}_{\text{Induced Electric Field}}$$
$$\underbrace{\oint \vec{E} \cdot d\vec{l} = 0}_{\text{Electrostatic Electric Fields}} \quad .$$

Our results can be summarized by combining these equations:

$$\epsilon = \oint \vec{E} \cdot d\vec{l} = -\frac{d\Phi_m}{dt}. \tag{9.6.1}$$

Example $9.6.1$: Induced Electric Field in a Circular Coil

What is the induced electric field in the circular coil of Example 13.3.1A (and Figure 13.3.3) at the three times indicated?

Strategy

Using cylindrical symmetry, the electric field integral simplifies into the electric field times the circumference of a circle. Since we already know the induced emf, we can connect these two expressions by Faraday's law to solve for the induced electric field.

Solution

The induced electric field in the coil is constant in magnitude over the cylindrical surface, similar to how Ampere's law problems with cylinders are solved. Since $\vec{E}$ is tangent to the coil,

$$\oint \vec{E} \cdot d\vec{l} = \oint E dl = 2\pi r E.$$

When combined with Equation 9.6.1, this gives

$$E = \frac{\epsilon}{2\pi r}.$$

The direction of ϵ is counterclockwise, and $\vec{E}$ circulates in the same direction around the coil. The values of **E** are

$$E(t_1) = \frac{6.0\,V}{2\pi\,(0.50\,m)} = 1.9\,V/m;$$

$$E(t_2) = \frac{4.7\,V}{2\pi\,(0.50\,m)} = 1.5\,V/m;$$

$$E(t_3) = \frac{0.040\,V}{2\pi\,(0.50\,m)} = 0.013\,V/m;$$

Significance

When the magnetic flux through a circuit changes, a nonconservative electric field is induced, which drives current through the circuit. But what happens if $dB/dt \neq 0$ in free space where there isn't a conducting path? The answer is that this case can be treated **as if a conducting path were present**; that is, nonconservative electric fields are induced wherever $dB/dt \neq 0$ whether or not there is a conducting path present.

These nonconservative electric fields always satisfy Equation 9.6.1. For example, if the circular coil were removed, an electric field **in free space** at $r = 0.50\,m$ would still be directed counterclockwise, and its magnitude would still be 1.9 V/m at $t = 0$. 1.5 V/m at $t = 5.0 \times 10^{-2}\,s$, etc. The existence of induced electric fields is certainly **not** restricted to wires in circuits.

Example 9.6.2: Electric Field Induced by the Changing Magnetic Field of a Solenoid

Figure $9.6.1a$ shows a long solenoid with radius R and n turns per unit length; its current decreases with time according to $I = I_0 e^{-\alpha t}$. What is the magnitude of the induced electric field at a point a distance **r** from the central axis of the solenoid (a) when $r > R$ and (b) when $r < R$ [Figure $9.6.1b$]. (c) What is the direction of the induced field at both locations? Assume that the infinite-solenoid approximation is valid throughout the regions of interest.

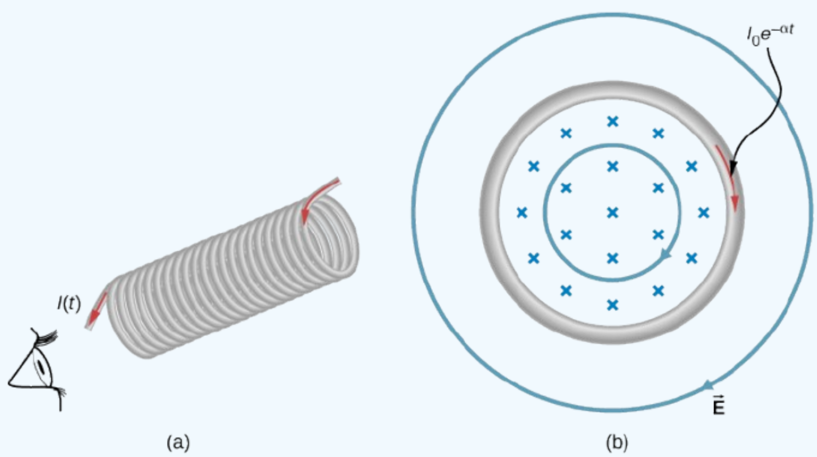


Figure 9.6.1: (a) The current in a long solenoid is decreasing exponentially. (b) A cross-sectional view of the solenoid from its left end. The cross-section shown is near the middle of the solenoid. An electric field is induced both inside and outside the solenoid.

Strategy

Using the formula for the magnetic field inside an infinite solenoid and Faraday's law, we calculate the induced emf. Since we have cylindrical symmetry, the electric field integral reduces to the electric field times the circumference of the integration path. Then we solve for the electric field.

Solution

a. The magnetic field is confined to the interior of the solenoid where

$$B = \mu_0 n I = \mu_0 n I_0 e^{-\alpha t}.$$

Thus, the magnetic flux through a circular path whose radius r is greater than R, the solenoid radius, is

$$\Phi_m = BA = \mu_0 n I_0 \pi R^2 e^{-\alpha t}.$$

The induced field $\vec{E}$ is tangent to this path, and because of the cylindrical symmetry of the system, its magnitude is constant on the path. Hence, we have

$$\left|\oint \vec{E}\cdot d\vec{l}\right| = \left|\frac{d\Phi_m}{dt}\right|,$$

$$E(2\pi r) = \left|\frac{d}{dt}(\mu_0 n I_0 \pi R^2 e^{-\alpha t})\right| = \alpha \mu_0 n I_0 \pi R^2 e^{-\alpha t},$$

$$E = \frac{\alpha \mu_0 n I_0 R^2}{2r} e^{-\alpha t} \; (r > R).$$

b. For a path of radius **r** inside the solenoid, $\Phi_m = B\pi r^2$, so

$$E(2\pi r) = \left|\frac{d}{dt}(\mu_0 n I_0 \pi r^2 e^{-\alpha t})\right| = \alpha \mu_0 n I_0 \pi r^2 e^{-\alpha t},$$

and the induced field is

$$E = \frac{\alpha \mu_0 n I_0 r}{2} e^{-\alpha t} \; (r < R).$$

c. The magnetic field points into the page as shown in part (b) and is decreasing. If either of the circular paths were occupied by conducting rings, the currents induced in them would circulate as shown, in conformity with Lenz's law. The induced electric field must be so directed as well.

Significance

In part (b), note that $|\vec{E}|$ increases with r inside and decreases as $1/r$ outside the solenoid, as shown in Figure $9.6.2$.

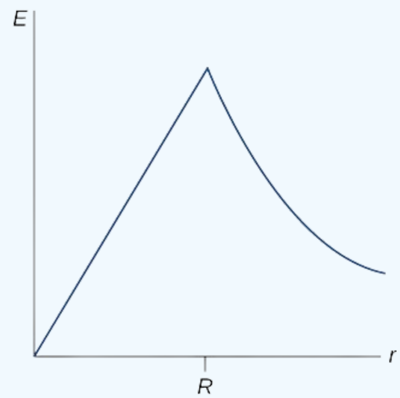


Figure $9.6.2$: The electric field vs. distance **r**. When $r < R$, the electric field rises linearly, whereas when $r > R$, the electric field falls of proportional to 1/**r**.

? Exercise $9.6.1$

Suppose that the coil of Example 13.3.1A is a square rather than circular. Can Equation $9.6.1$ be used to calculate (a) the induced emf and (b) the induced electric field?

Answer

a. yes; b. Yes; however there is a lack of symmetry between the electric field and coil, making $\oint \vec{E}\cdot d\vec{l}$ a more complicated relationship that can't be simplified as shown in the example.

Exercise 9.6.2

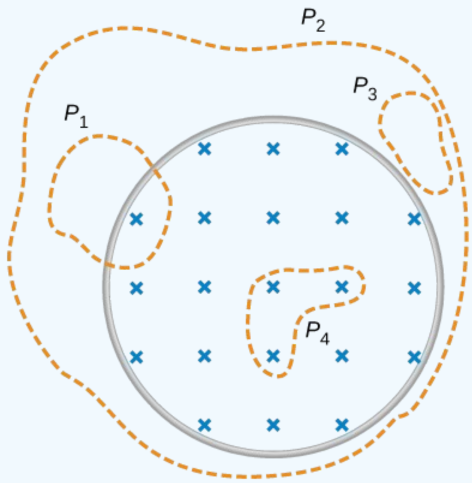
What is the magnitude of the induced electric field in Example 9.6.2 at $t=0$ if $r=6.0\,cm$, $R=2.0\,cm$, $n=2000$ turns per meter, $I_0=2.0\,A$, and $\alpha=200\,s^{-1}$?

Answer

$3.4\times10^{-3}\,V/m$

Exercise 9.6.3

The magnetic field shown below is confined to the cylindrical region shown and is changing with time. Identify those paths for which $\epsilon=\oint\vec{E}\cdot d\vec{l}\neq0$.



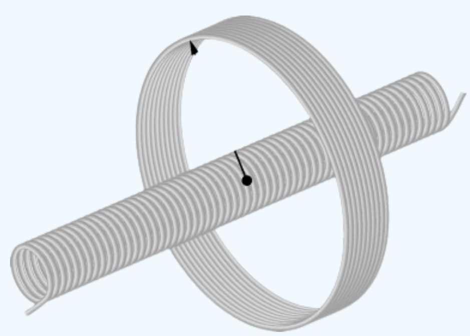
Answer

$P_1,\ P_2,\ P_4$

Exercise 9.6.1

A long solenoid of cross-sectional area $5.0\,cm^2$ is wound with 25 turns of wire per centimeter. It is placed in the middle of a closely wrapped coil of 10 turns and radius 25 cm, as shown below.

a. What is the emf induced in the coil when the current through the solenoid is decreasing at a rate $dI/dt=-0.20\,A/s$?
b. What is the electric field induced in the coil?



Answer

a. $3.1\times10^{-6}V$; b. $2.0\times10^{-7}\,V/m$

9.7: Eddy Currents

Learning Objectives

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

- Explain how eddy currents are created in metals
- Describe situations where eddy currents are beneficial and where they are not helpful

As discussed two sections earlier, a motional emf is induced when a conductor moves in a magnetic field or when a magnetic field moves relative to a conductor. If motional emf can cause a current in the conductor, we refer to that current as an **eddy current**.

Magnetic Damping

Eddy currents can produce significant drag, called **magnetic damping**, on the motion involved. Consider the apparatus shown in Figure $9.7.1$, which swings a pendulum bob between the poles of a strong magnet. (This is another favorite physics demonstration.) If the bob is metal, significant drag acts on the bob as it enters and leaves the field, quickly damping the motion. If, however, the bob is a slotted metal plate, as shown in part (b) of the figure, the magnet produces a much smaller effect. There is no discernible effect on a bob made of an insulator. Why does drag occur in both directions, and are there any uses for magnetic drag?

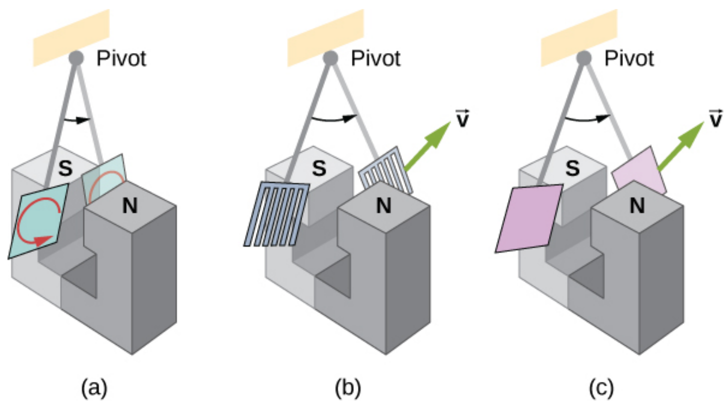


Figure $9.7.1$: A common physics demonstration device for exploring eddy currents and magnetic damping. (a) The motion of a metal pendulum bob swinging between the poles of a magnet is quickly damped by the action of eddy currents. (b) There is little effect on the motion of a slotted metal bob, implying that eddy currents are made less effective. (c) There is also no magnetic damping on a nonconducting bob, since the eddy currents are extremely small.

Figure $9.7.2$ shows what happens to the metal plate as it enters and leaves the magnetic field. In both cases, it experiences a force opposing its motion. As it enters from the left, flux increases, setting up an eddy current (Faraday's law) in the counterclockwise direction (Lenz's law), as shown. Only the right-hand side of the current loop is in the field, so an unopposed force acts on it to the left (RHR-1). When the metal plate is completely inside the field, there is no eddy current if the field is uniform, since the flux remains constant in this region. But when the plate leaves the field on the right, flux decreases, causing an eddy current in the clockwise direction that, again, experiences a force to the left, further slowing the motion. A similar analysis of what happens when the plate swings from the right toward the left shows that its motion is also damped when entering and leaving the field.

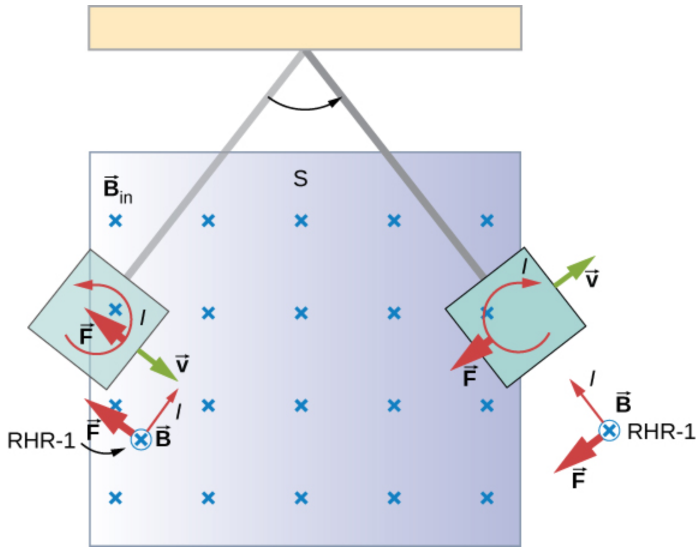


Figure 9.7.2: A more detailed look at the conducting plate passing between the poles of a magnet. As it enters and leaves the field, the change in flux produces an eddy current. Magnetic force on the current loop opposes the motion. There is no current and no magnetic drag when the plate is completely inside the uniform field.

When a slotted metal plate enters the field (Figure 9.7.3), an emf is induced by the change in flux, but it is less effective because the slots limit the size of the current loops. Moreover, adjacent loops have currents in opposite directions, and their effects cancel. When an insulating material is used, the eddy current is extremely small, so magnetic damping on insulators is negligible. If eddy currents are to be avoided in conductors, then they must be slotted or constructed of thin layers of conducting material separated by insulating sheets.

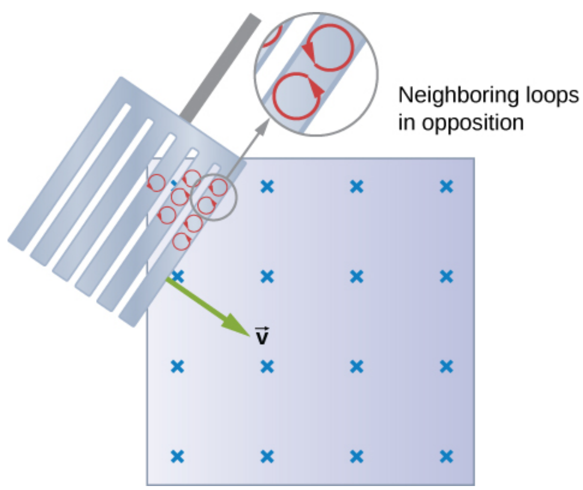


Figure 9.7.3: Eddy currents induced in a slotted metal plate entering a magnetic field form small loops, and the forces on them tend to cancel, thereby making magnetic drag almost zero.

Applications of Magnetic Damping

One use of magnetic damping is found in sensitive laboratory balances. To have maximum sensitivity and accuracy, the balance must be as friction-free as possible. But if it is friction-free, then it will oscillate for a very long time. Magnetic damping is a simple and ideal solution. With magnetic damping, drag is proportional to speed and becomes zero at zero velocity. Thus, the oscillations are quickly damped, after which the damping force disappears, allowing the balance to be very sensitive (Figure 9.7.4). In most balances, magnetic damping is accomplished with a conducting disc that rotates in a fixed field.

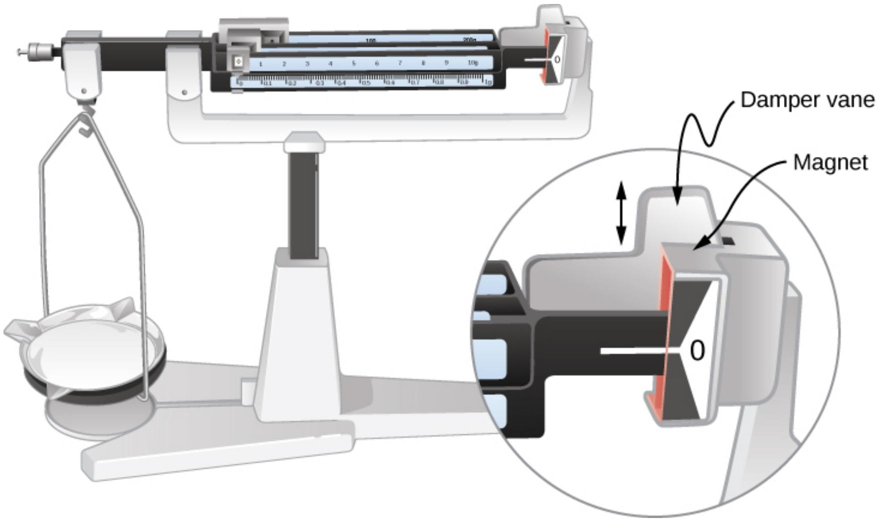


Figure 9.7.4: Magnetic damping of this sensitive balance slows its oscillations. Since Faraday's law of induction gives the greatest effect for the most rapid change, damping is greatest for large oscillations and goes to zero as the motion stops.

Since eddy currents and magnetic damping occur only in conductors, recycling centers can use magnets to separate metals from other materials. Trash is dumped in batches down a ramp, beneath which lies a powerful magnet. Conductors in the trash are slowed by magnetic damping while nonmetals in the trash move on, separating from the metals (Figure 9.7.5). This works for all metals, not just ferromagnetic ones. A magnet can separate out the ferromagnetic materials alone by acting on stationary trash.

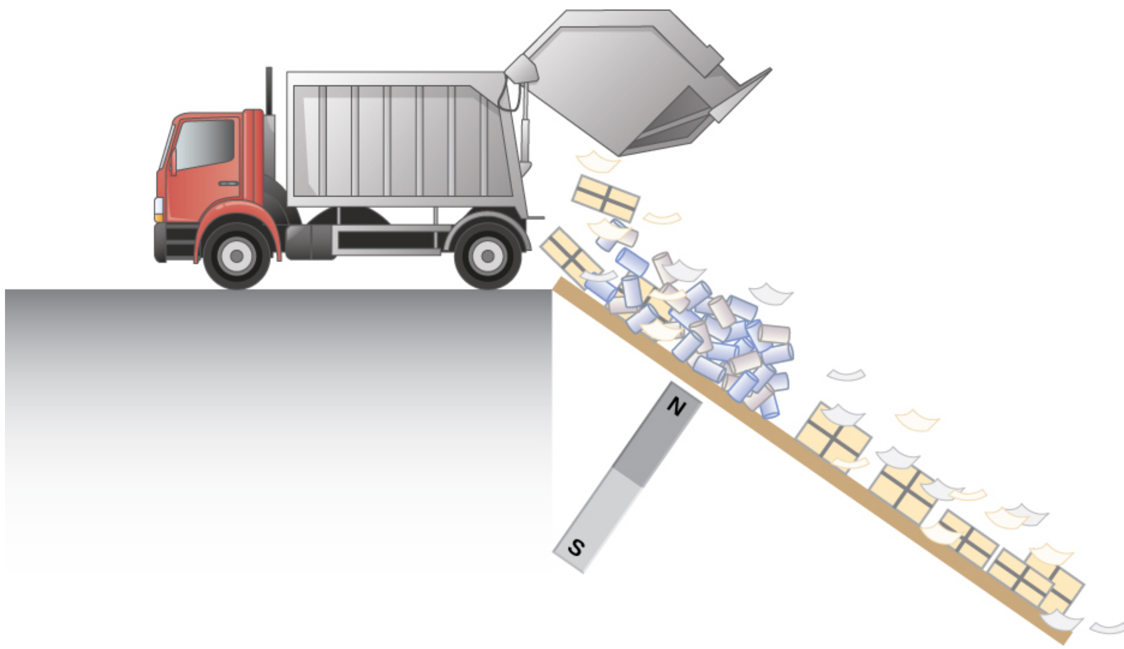


Figure 9.7.5: Metals can be separated from other trash by magnetic drag. Eddy currents and magnetic drag are created in the metals sent down this ramp by the powerful magnet beneath it. Nonmetals move on.

Other major applications of eddy currents appear in **metal detectors** and **braking systems** in trains and roller coasters. Portable metal detectors (Figure 9.7.6) consist of a primary coil carrying an alternating current and a secondary coil in which a current is induced. An eddy current is induced in a piece of metal close to the detector, causing a change in the induced current within the secondary coil. This can trigger some sort of signal, such as a shrill noise.

Figure 9.7.6: A soldier in Iraq uses a metal detector to search for explosives and weapons. (credit: U.S. Army)

Braking using eddy currents is safer because factors such as rain do not affect the braking and the braking is smoother. However, eddy currents cannot bring the motion to a complete stop, since the braking force produced decreases as speed is reduced. Thus, speed can be reduced from say 20 m/s to 5 m/s, but another form of braking is needed to completely stop the vehicle. Generally, powerful rare-earth magnets such as neodymium magnets are used in roller coasters. Figure 9.7.7 shows rows of magnets in such an application. The vehicle has metal fins (normally containing copper) that pass through the magnetic field, slowing the vehicle down in much the same way as with the pendulum bob shown in Figure 9.7.1.

Figure 9.7.7: The rows of rare-earth magnets (protruding horizontally) are used for magnetic braking in roller coasters. (credit: Stefan Scheer)

Induction cooktops have electromagnets under their surface. The magnetic field is varied rapidly, producing eddy currents in the base of the pot, causing the pot and its contents to increase in temperature. Induction cooktops have high efficiencies and good response times but the base of the pot needs to be conductors, such as iron or steel, for induction to work.

9.8: Electric Generators and Back Source Voltage

Learning Objectives

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

- Explain how an electric generator works
- Determine the induced source voltage in a loop at any time interval, rotating at a constant rate in a magnetic field.
- Show that rotating coils have an induced source voltage; in motors this is called back source voltage because it opposes the source voltage input to the motor.

A variety of important phenomena and devices can be understood with Faraday's law. In this section, we examine two of these.

Electric Generators

Electric generators induce a source voltage by rotating a coil in a magnetic field, as briefly discussed in Motional Source Voltage. We now explore generators in more detail. Consider the following example.

Calculating the Source Voltage Induced in a Generator Coil

The generator coil shown in Figure 9.8.1 is rotated through one-fourth of a revolution (from $\theta = 0^o$ to $\theta = 90^o$) in 15.0 ms. The 200-turn circular coil has a 5.00-cm radius and is in a uniform 0.80-T magnetic field. What is the source voltage induced?

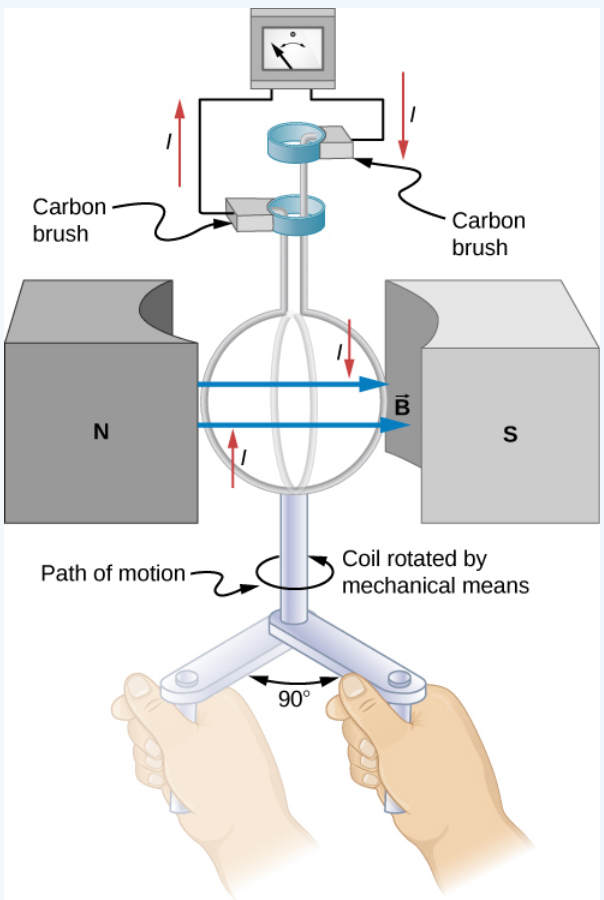


Figure 9.8.1: When this generator coil is rotated through one-fourth of a revolution, the magnetic flux Φ_m changes from its maximum to zero, inducing a source voltage.

Strategy

Faraday's law of induction is used to find the source voltage induced:

$$\epsilon=-N\frac{d\Phi_m}{dt}. \tag{9.8.1}$$

We recognize this situation as the same one in Example 13.4.3. According to the diagram, the projection of the surface normal vector $\hat{n}$ to the magnetic field is initially $cos\,\theta$ and this is inserted by the definition of the dot product. The magnitude of the magnetic field and area of the loop are fixed over time, which makes the integration simplify quickly. The induced source voltage is written out using Faraday's law:

$$\epsilon=NBA\,sin\,\theta\frac{d\theta}{dt}. \tag{9.8.2}$$

Solution We are given that $N=200$, $B=0.80\,T$, $\theta=90^o$, $d\theta=90^o=\pi/2$, and $dt=15.0\,ms$. The area of the loop is

$$A=\pi r^2=(3.14)(0.0500\,m)^2=7.85\times10^{-3}\,m^2. \tag{9.8.3}$$

Entering this value gives

$$\epsilon=(200)(0.80\,T)(7.85\times10^{-3}\,m^2)sin(90^o)\frac{\pi/2}{15.0\times10^{-3}s}=131\,V. \tag{9.8.4}$$

Significance

This is a practical average value, similar to the 120 V used in household power.

The source voltage calculated in Example 9.8.1 is the average over one-fourth of a revolution. What is the source voltage at any given instant? It varies with the angle between the magnetic field and a perpendicular to the coil. We can get an expression for source voltage as a function of time by considering the motional source voltage on a rotating rectangular coil of width **w** and height **l** in a uniform magnetic field, as illustrated in Figure 9.8.2.

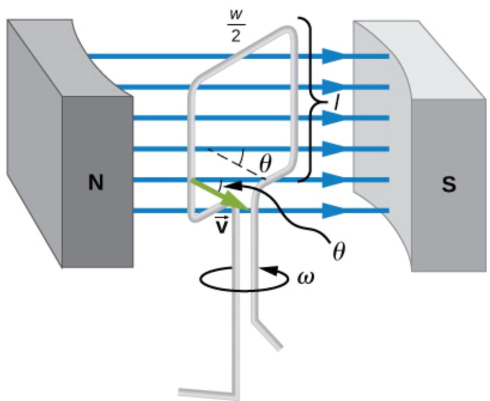


Figure 9.8.2: A generator with a single rectangular coil rotated at constant angular velocity in a uniform magnetic field produces a source voltage that varies sinusoidally in time. Note the generator is similar to a motor, except the shaft is rotated to produce a current rather than the other way around.

Charges in the wires of the loop experience the magnetic force, because they are moving in a magnetic field. Charges in the vertical wires experience forces parallel to the wire, causing currents. But those in the top and bottom segments feel a force perpendicular to the wire, which does not cause a current. We can thus find the induced source voltage by considering only the side wires. Motional source voltage is given to be $\epsilon=Blv$, where the velocity **v** is perpendicular to the magnetic field **B**. Here the velocity is at an angle θ with **B**, so that its component perpendicular to **B** is **v** sin θ (see Figure 9.8.2). Thus, in this case, the source voltage induced on each side is $\epsilon=Blv\,sin\,\theta$ and they are in the same direction. The total source voltage around the loop is then

$$\epsilon=2Blv\,sin\,\theta. \tag{9.8.5}$$

This expression is valid, but it does not give soure voltage as a function of time. To find the time dependence of source voltage, we assume the coil rotates at a constant angular velocity ω. The angle θ is related to angular velocity by $\theta=\omega t$, so that

$$\epsilon=2Blv\,sin(\omega t). \tag{9.8.6}$$

Now, linear velocity **v** is related to angular velocity ω by $v=r\omega$. Here, $r=\omega/2$, so that $v=(\omega/2)\omega$, and

$$\epsilon = 2Bl\frac{\omega}{2}\omega \, sin\, \omega t = (l\omega)Bw \, sin\, \omega t. \tag{9.8.7}$$

Noting that the area of the loop is $A = l\omega$, and allowing for **N** loops, we find that

$$\epsilon = NBAw \, sin\,(\omega t). \tag{9.8.8}$$

This is the source voltage induced in a generator coil of **N** turns and area **A** rotating at a constant angular velocity ω in a uniform magnetic field **B**. This can also be expressed as

$$\epsilon = \epsilon_0 \, sin\, \omega t, \tag{9.8.9}$$

where

$$\epsilon_0 = NAB\omega \tag{9.8.10}$$

is the peak source voltage, since the maximum value of $sin(\omega t) = 1$. Note that the frequency of the oscillation is $f = \omega/2\pi$ and the period is $T = 1/f = 2\pi/\omega$. Figure $9.8.3$ shows a graph of source voltage as a function of time, and it now seems reasonable that ac voltage is sinusoidal.

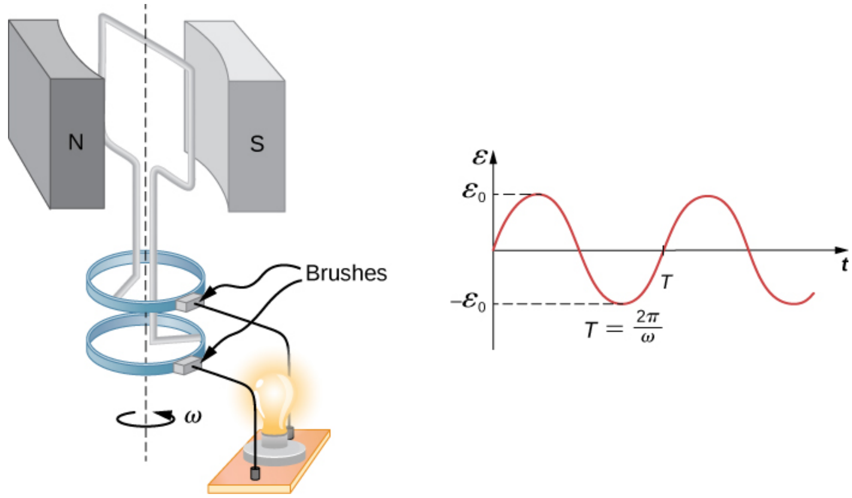


Figure $9.8.3$: The source voltage of a generator is sent to a light bulb with the system of rings and brushes shown. The graph gives the source voltage of the generator as a function of time, where ϵ_0 is the peak source voltage. The period is $T = 1/f = 2\pi/\omega$, where **f** is the frequency.

The fact that the peak source voltage is $\epsilon_0 = NBA\omega$ makes good sense. The greater the number of coils, the larger their area, and the stronger the field, the greater the output voltage. It is interesting that the faster the generator is spun (greater ω), the greater the source voltage. This is noticeable on bicycle generators—at least the cheaper varieties.

Figure $9.8.4$ shows a scheme by which a generator can be made to produce pulsed dc. More elaborate arrangements of multiple coils and split rings can produce smoother dc, although electronic rather than mechanical means are usually used to make ripple-free dc.

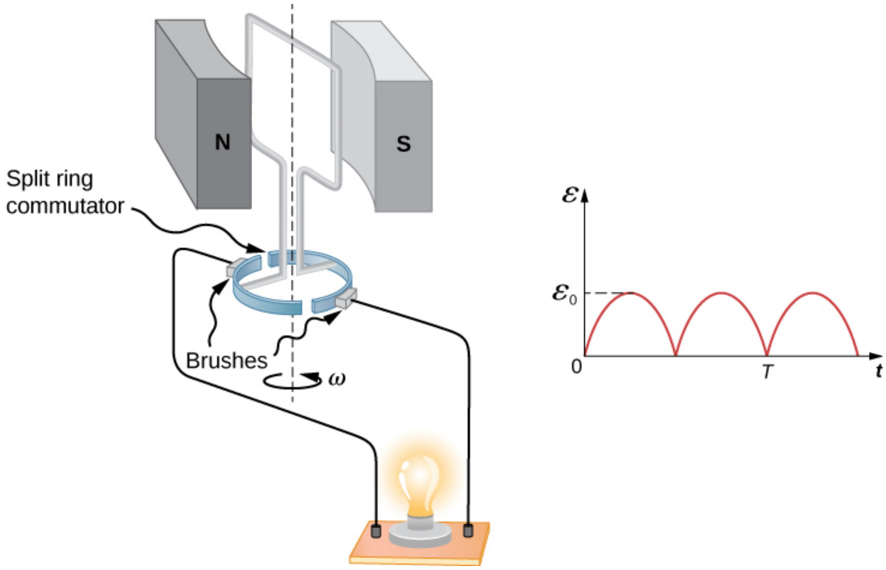


Figure 9.8.4: Split rings, called commutators, produce a pulsed dc source voltage output in this configuration.

In real life, electric generators look a lot different from the figures in this section, but the principles are the same. The source of mechanical energy that turns the coil can be falling water (hydropower), steam produced by the burning of fossil fuels, or the kinetic energy of wind. Figure 9.8.5 shows a cutaway view of a steam turbine; steam moves over the blades connected to the shaft, which rotates the coil within the generator. The generation of electrical energy from mechanical energy is the basic principle of all power that is sent through our electrical grids to our homes.

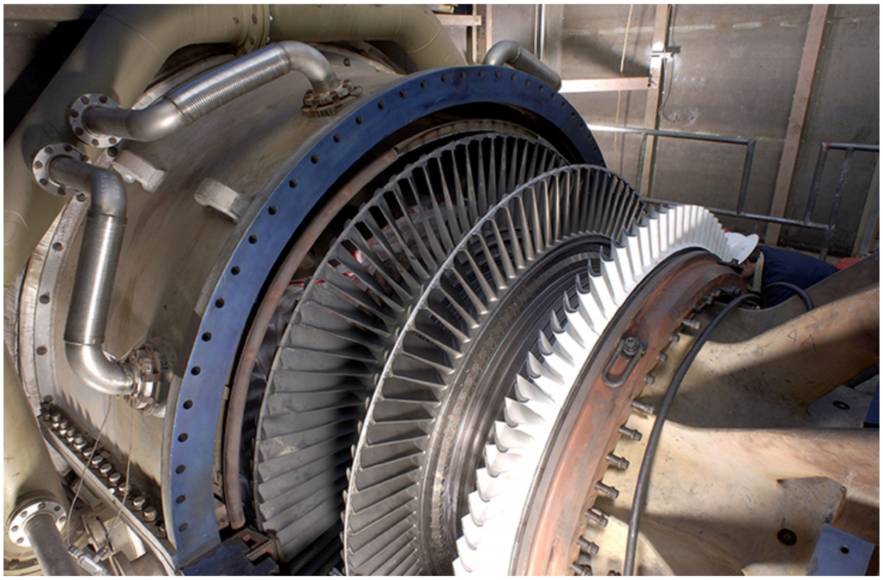


Figure 9.8.5: Steam turbine/generator. The steam produced by burning coal impacts the turbine blades, turning the shaft, which is connected to the generator.

Generators illustrated in this section look very much like the motors illustrated previously. This is not coincidental. In fact, a motor becomes a generator when its shaft rotates. Certain early automobiles used their starter motor as a generator. In the next section, we further explore the action of a motor as a generator.

Back Source Voltage

Generators convert mechanical energy into electrical energy, whereas motors convert electrical energy into mechanical energy. Thus, it is not surprising that motors and generators have the same general construction. A motor works by sending a current through a loop of wire located in a magnetic field. As a result, the magnetic field exerts torque on the loop. This rotates a shaft, thereby extracting mechanical work out of the electrical current sent in initially. (Refer to Force and Torque on a Current Loop for a discussion on motors that will help you understand more about them before proceeding.)

When the coil of a motor is turned, magnetic flux changes through the coil, and a source voltage (consistent with Faraday's law) is induced. The motor thus acts as a generator whenever its coil rotates. This happens whether the shaft is turned by an external input, like a belt drive, or by the action of the motor itself. That is, when a motor is doing work and its shaft is turning, a source voltage is generated. Lenz's law tells us the source voltage opposes any change, so that the motor's self-generated source voltage opposes the input source voltage that powers the motor, called the **back source voltage** of the motor (Figure $9.8.6$).

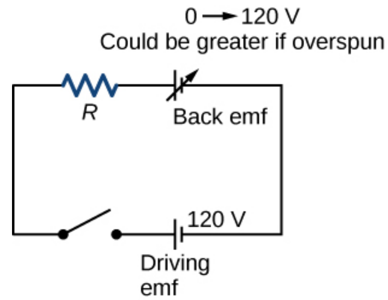


Figure $9.8.6$: The coil of a dc motor is represented as a resistor in this schematic. The back source voltage is represented as a variable source voltage that opposes the source voltage driving the motor. Back source voltage is zero when the motor is not turning and increases proportionally to the motor's angular velocity.

The generator output of a motor is the difference between the supply voltage and the back source voltage. The back source voltage is zero when the motor is first turned on, meaning that the coil receives the full driving voltage and the motor draws maximum current when it is on but not turning. As the motor turns faster, the back source voltage grows, always opposing the driving source voltage, and reduces both the voltage across the coil and the amount of current it draws. This effect is noticeable in many common situations. When a vacuum cleaner, refrigerator, or washing machine is first turned on, lights in the same circuit dim briefly due to the **IR** drop produced in feeder lines by the large current drawn by the motor.

When a motor first comes on, it draws more current than when it runs at its normal operating speed. When a mechanical load is placed on the motor, like an electric wheelchair going up a hill, the motor slows, the back source voltage drops, more current flows, and more work can be done. If the motor runs at too low a speed, the larger current can overheat it (via resistive power in the coil, $P = I^2R$), perhaps even burning it out. On the other hand, if there is no mechanical load on the motor, it increases its angular velocity ω until the back source voltage is nearly equal to the driving source voltage. Then the motor uses only enough energy to overcome friction.

Eddy currents in iron cores of motors can cause troublesome energy losses. These are usually minimized by constructing the cores out of thin, electrically insulated sheets of iron. The magnetic properties of the core are hardly affected by the lamination of the insulating sheet, while the resistive heating is reduced considerably. Consider, for example, the motor coils represented in Figure $9.8.6$. The coils have an equivalent resistance of $0.400\,\Omega$ and are driven by a source voltage of 48.0 V. Shortly after being turned on, they draw a current

$$I = V/R = (48.0\,V)/(0.400\,\Omega) = 120\,A \tag{9.8.11}$$

and thus dissipate $P = I^2R = 5.76\,kW$ of energy as heat transfer. Under normal operating conditions for this motor, suppose the back source voltage is 40.0 V. Then at operating speed, the total voltage across the coils is 8.0 V (48.0 V minus the 40.0 V back source voltage), and the current drawn is

$$I = V/R = (8.0\,V)/(0.400\,\Omega) = 20\,A. \tag{9.8.12}$$

Under normal load, then, the power dissipated is $P = IV = (20\,A)(8.0\,V) = 160\,W$. This does not cause a problem for this motor, whereas the former 5.76 kW would burn out the coils if sustained.

✔ A Series-Wound Motor in Operation

The total resistance $(R_f + R_a)$ of a series-wound dc motor is $2.0\,\Omega$ (Figure $9.8.7$). When connected to a 120-V source (ϵ_S), the motor draws 10 A while running at constant angular velocity. (a) What is the back source voltage induced in the rotating coil, ϵ_i? (b) What is the mechanical power output of the motor? (c) How much power is dissipated in the resistance of the coils? (d) What is the power output of the 120-V source? (e) Suppose the load on the motor increases, causing it to slow down to the point where it draws 20 A. Answer parts (a) through (d) for this situation.

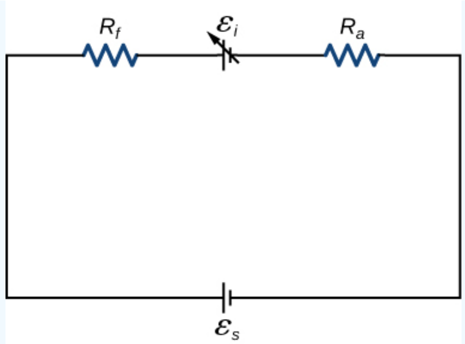


Figure 9.8.7: Circuit representation of a series-wound direct current motor.

Strategy

The back source voltage is calculated based on the difference between the supplied voltage and the loss from the current through the resistance. The power from each device is calculated from one of the power formulas based on the given information.

Solution

1. The back source voltage is

$$\epsilon_i = \epsilon_S - I(R_f + RE_a) = 120\,V - (10\,A)(2.0\,\Omega) = 100\,V. \tag{9.8.13}$$

2. Since the potential across the armature is 100 V when the current through it is 10 A, the power output of the motor is

$$P_m = \epsilon_i I = (100\,V)(10\,A) = 1.0 \times 10^3\,W. \tag{9.8.14}$$

3. A 10-A current flows through coils whose combined resistance is $2.0\,\Omega$, so the power dissipated in the coils is

$$P_R = I^2 R = (10\,A)^2(2.0\,\Omega) = 2.0 \times 10^2\,W. \tag{9.8.15}$$

4. Since 10 A is drawn from the 120-V source, its power output is

$$P_S = \epsilon_S I = (120\,V)(10\,A) = 1.2 \times 10^3\,W. \tag{9.8.16}$$

5. Repeating the same calculations with $I = 20\,A$, we find

$$\epsilon_i = 80\,V,\ P_m = 1.6 \times 10^3\,W,\ P_R = 8.0 \times 10^2\,W,\ and\ P_s = 2.4 \times 10^3\,W. \tag{9.8.17}$$

The motor is turning more slowly in this case, so its power output and the power of the source are larger.

Significance Notice that we have an energy balance in part (d):

$$1.2 \times 10^3\,W = 1.0 \times 10^3\,W + 2.0 \times 10^2\,W. \tag{9.8.18}$$

9.9: Transformers

Learning Objectives

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

- Explain why power plants transmit electricity at high voltages and low currents and how they do this
- Develop relationships among current, voltage, and the number of windings in step-up and step-down transformers

Although ac electric power is produced at relatively low voltages, it is sent through transmission lines at very high voltages (as high as 500 kV). The same power can be transmitted at different voltages because power is the product $I_{rms}V_{rms}$. (For simplicity, we ignore the phase factor cos ϕ.) A particular power requirement can therefore be met with a low voltage and a high current or with a high voltage and a low current. The advantage of the high-voltage/low-current choice is that it results in lower $I^2_{rms}R$ ohmic losses in the transmission lines, which can be significant in lines that are many kilometers long (Figure $9.9.1$).

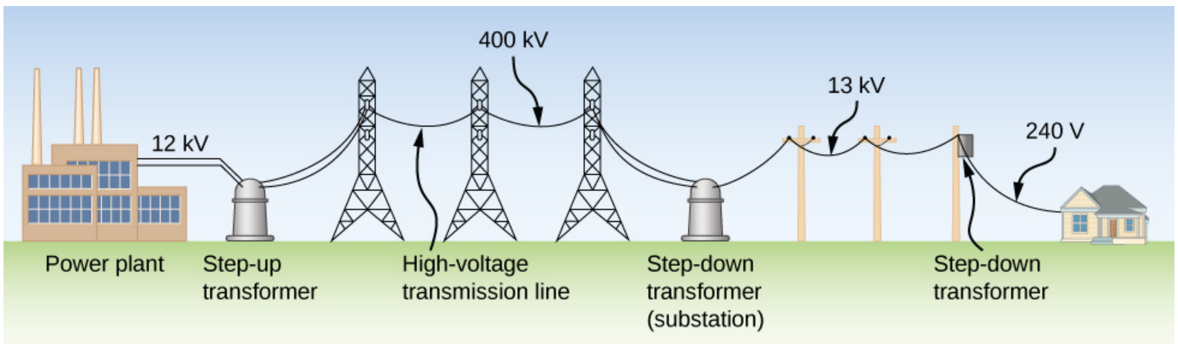


Figure $9.9.1$: The rms voltage from a power plant eventually needs to be stepped down from 12 kV to 240 V so that it can be safely introduced into a home. A high-voltage transmission line allows a low current to be transmitted via a substation over long distances.

Typically, the alternating emfs produced at power plants are “stepped up” to very high voltages before being transmitted through power lines; then, they must be “stepped down” to relatively safe values (110 or 220 V rms) before they are introduced into homes. The device that transforms voltages from one value to another using induction is the **transformer** (Figure $9.9.2$).

Figure 9.9.2: Transformers are used to step down the high voltages in transmission lines to the 110 to 220 V used in homes. (credit: modification of work by “Fortyseven”/Flickr)

As Figure 9.9.3 illustrates, a transformer basically consists of two separated coils, or windings, wrapped around a soft iron core. The primary winding has N_p loops, or turns, and is connected to an alternating voltage $v_p(t)$. The secondary winding has N_s turns and is connected to a load resistor R_s. We assume the ideal case for which all magnetic field lines are confined to the core so that the same magnetic flux permeates each turn of both the primary and the secondary windings. We also neglect energy losses to magnetic hysteresis, to ohmic heating in the windings, and to ohmic heating of the induced eddy currents in the core. A good transformer can have losses as low as 1% of the transmitted power, so this is not a bad assumption.

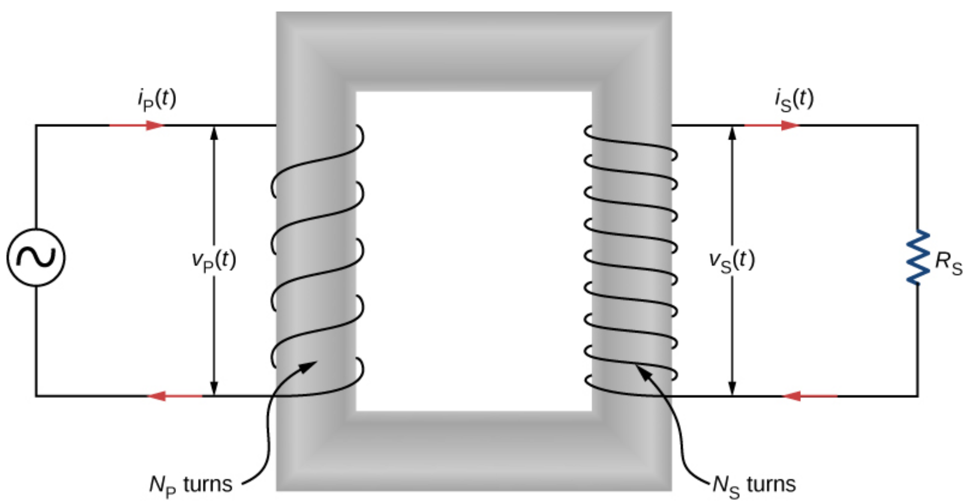


Figure 9.9.3: A step-up transformer (more turns in the secondary winding than in the primary winding). The two windings are wrapped around a soft iron core.

To analyze the transformer circuit, we first consider the primary winding. The input voltage $v_p(t)$ is equal to the potential difference induced across the primary winding. From Faraday’s law, the induced potential difference is $-N_p(d\Phi/dt)$, where Φ is the flux through one turn of the primary winding. Thus,

$$v_p(t) = -N_p \frac{d\Phi}{dt}.$$

Similarly, the output voltage $v_s(t)$ delivered to the load resistor must equal the potential difference induced across the secondary winding. Since the transformer is ideal, the flux through every turn of the secondary winding is also Φ and

$$v_s(t) = -N_s \frac{d\Phi}{dt}.$$

Combining the last two equations, we have

$$v_s(t) = \frac{N_s}{N_p} v_p(t). \tag{9.9.1}$$

Hence, with appropriate values for N_s and N_p, the input voltage $v_p(t)$ may be "stepped up" $(N_s > N_p)$ or "stepped down" $(N_s < N_p)$ to $v_s(t)$, the output voltage. This is often abbreviated as the **transformer equation**,

$$\frac{V_s}{V_p} = \frac{N_s}{N_p}, \tag{9.9.2}$$

which shows that the ratio of the secondary to primary voltages in a transformer equals the ratio of the number of turns in their windings. For a **step-up transformer**, which increases voltage and decreases current, this ratio is greater than one; for a **step-down transformer**, which decreases voltage and increases current, this ratio is less than one.

From the law of energy conservation, the power introduced at any instant by $v_p(t)$ to the primary winding must be equal to the power dissipated in the resistor of the secondary circuit; thus,

$$i_p(t)v_p(t) = i_s(t)v_s(t).$$

When combined with Equation 9.9.2, this gives

$$i_s(t) = \frac{N_p}{N_s} i_p(t). \tag{9.9.3}$$

If the voltage is stepped up, the current is stepped down, and vice versa.

Finally, we can use $i_s(t) = v_s(t)/R_s$, along with Equation 9.9.1 and Equation 9.9.3, to obtain

$$v_p(t) = i_p \left[\left(\frac{N_p}{N_s} \right)^2 R_s \right],$$

which tells us that the input voltage $v_p(t)$ "sees" not a resistance R_s but rather a resistance

$$R_p = \left(\frac{N_p}{N_s} \right)^2 R_s.$$

Our analysis has been based on instantaneous values of voltage and current. However, the resulting equations are not limited to instantaneous values; they hold also for maximum and rms values.

✔ Example 9.9.1: A Step-Down Transformer

A transformer on a utility pole steps the rms voltage down from 12 kV to 240 V.

a. What is the ratio of the number of secondary turns to the number of primary turns?
b. If the input current to the transformer is 2.0 A, what is the output current?
c. Determine the power loss in the transmission line if the total resistance of the transmission line is 200 Ω.
d. What would the power loss have been if the transmission line was at 240 V the entire length of the line, rather than providing voltage at 12 kV? What does this say about transmission lines?

Strategy

The number of turns related to the voltages is found from Equation 9.9.1. The output current is calculated using Equation 9.9.3.

Solution

a. Using Equation 9.9.1 with rms values V_p and V_s we have

$$\frac{N_s}{N_p} = \frac{240\,V}{12 \times 10^3 V} = \frac{1}{50}, \tag{9.9.4}$$

so the primary winding has 50 times the number of turns in the secondary winding.

b. From Equation 9.9.3, the output rms current I_s is found using the transformer equation with current

$$I_S = \frac{N_p}{N_S} I_p \tag{9.9.5}$$

such that

$$I_s = \frac{N_p}{N_s} I_p = (50)(2.0\,A) = 100\,A.$$

c. The power loss in the transmission line is calculated to be

$$P_{loss} = I_p^2 R = (2.0\,A)^2 (200\,\Omega) = 800\,W.$$

d. If there were no transformer, the power would have to be sent at 240 V to work for these houses, and the power loss would be

$$P_{loss} = I_s^2 R = (100\,A)^2 (200\,\Omega) = 2 \times 10^6\,W.$$

Therefore, when power needs to be transmitted, we want to avoid power loss. Thus, lines are sent with high voltages and low currents and adjusted with a transformer before power is sent into homes.

Significance

This application of a step-down transformer allows a home that uses 240-V outlets to have 100 A available to draw upon. This can power many devices in the home.

? Exercise 9.9.1

A transformer steps the line voltage down from 110 to 9.0 V so that a current of 0.50 A can be delivered to a doorbell.

a. What is the ratio of the number of turns in the primary and secondary windings?
b. What is the current in the primary winding?
c. What is the resistance seen by the 110-V source?

Answer a

12:1

Answer b

0.042 A

Answer c

$2.6 \times 10^3\ \Omega$

9.10: Other Applications of Electromagnetic Induction

Learning Objectives

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

- Explain how computer hard drives and graphic tablets operate using magnetic induction
- Explain how hybrid/electric vehicles and transcranial magnetic stimulation use magnetic induction to their advantage

Modern society has numerous applications of Faraday's law of induction, as we will explore in this chapter and others. At this juncture, let us mention several that involve recording information using magnetic fields.

Some computer **hard drives** apply the principle of magnetic induction. Recorded data are made on a coated, spinning disk. Historically, reading these data was made to work on the principle of induction. However, most input information today is carried in digital rather than analog form—a series of 0s or 1s are written upon the spinning hard drive. Therefore, most hard drive readout devices do not work on the principle of induction, but use a technique known as **giant magnetoresistance**. Giant magnetoresistance is the effect of a large change of electrical resistance induced by an applied magnetic field to thin films of alternating ferromagnetic and nonmagnetic layers. This is one of the first large successes of nanotechnology.

Graphics tablets, or **tablet computers** where a specially designed pen is used to draw digital images, also applies induction principles. The tablets discussed here are labeled as passive tablets, since there are other designs that use either a battery-operated pen or optical signals to write with. The passive tablets are different than the touch tablets and phones many of us use regularly, but may still be found when signing your signature at a cash register. Underneath the screen, shown in Figure 9.10.1, are tiny wires running across the length and width of the screen. The pen has a tiny magnetic field coming from the tip. As the tip brushes across the screen, a changing magnetic field is felt in the wires which translates into an induced emf that is converted into the line you just drew.

Figure 9.10.1: A tablet with a specially designed pen to write with is another application of magnetic induction.

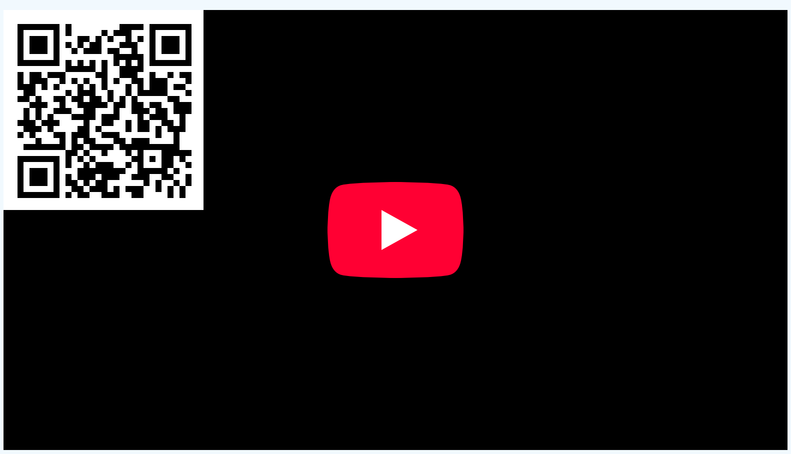
Another application of induction is the magnetic stripe on the back of your personal **credit card** as used at the grocery store or the ATM machine. This works on the same principle as the audio or video tape, in which a playback head reads personal information from your card.

> ✔ Video
>
> Check out this video to see how flashlights can use magnetic induction.
>
> 

A magnet moves by your mechanical work through a wire. The induced current charges a capacitor that stores the charge that will light the lightbulb even while you are not doing this mechanical work.

Electric and **hybrid vehicles** also take advantage of electromagnetic induction. One limiting factor that inhibits widespread acceptance of 100% electric vehicles is that the lifetime of the battery is not as long as the time you get to drive on a full tank of gas. To increase the amount of charge in the battery during driving, the motor can act as a generator whenever the car is braking, taking advantage of the back emf produced. This extra emf can be newly acquired stored energy in the car's battery, prolonging the life of the battery.

Another contemporary area of research in which electromagnetic induction is being successfully implemented is **transcranial magnetic stimulation (TMS)**. A host of disorders, including depression and hallucinations, can be traced to irregular localized electrical activity in the brain. In transcranial magnetic stimulation, a rapidly varying and very localized magnetic field is placed close to certain sites identified in the brain. The usage of TMS as a diagnostic technique is well established.

> ✔ Video
>
> Check out this Youtube video to see how rock-and-roll instruments like electric guitars use electromagnetic induction to get those strong beats.
>
> 

9.11: Electromagnetic Induction (Summary)

Key Terms

back emf	emf generated by a running motor, because it consists of a coil turning in a magnetic field; it opposes the voltage powering the motor
eddy current	current loop in a conductor caused by motional emf
electric generator	device for converting mechanical work into electric energy; it induces an emf by rotating a coil in a magnetic field
Faraday's law	induced emf is created in a closed loop due to a change in magnetic flux through the loop
induced electric field	created based on the changing magnetic flux with time
induced emf	short-lived voltage generated by a conductor or coil moving in a magnetic field
Lenz's law	direction of an induced emf opposes the change in magnetic flux that produced it; this is the negative sign in Faraday's law
magnetic damping	drag produced by eddy currents
magnetic flux	measurement of the amount of magnetic field lines through a given area
motionally induced emf	voltage produced by the movement of a conducting wire in a magnetic field
peak emf	maximum emf produced by a generator

Key Equations

Magnetic flux	$\Phi_m = \int_S \vec{B} \cdot \hat{n} dA$
Faraday's law	$\varepsilon = -N \frac{d\Phi_m}{dt}$
Motionally induced emf	$\varepsilon = Blv$
Motional emf around a circuit	$\varepsilon = \oint \vec{E} \cdot d\vec{l} = -\frac{d\Phi_m}{dt}$
Emf produced by an electric generator	$\varepsilon = NBA\omega sin(\omega t)$

Summary

13.2 Faraday's Law

- The magnetic flux through an enclosed area is defined as the amount of field lines cutting through a surface area A defined by the unit area vector.
- The units for magnetic flux are webers, where $1Wb = 1T \cdot m^2$.
- The induced emf in a closed loop due to a change in magnetic flux through the loop is known as Faraday's law. If there is no change in magnetic flux, no induced emf is created.

13.3 Lenz's Law

- We can use Lenz's law to determine the directions of induced magnetic fields, currents, and emfs.
- The direction of an induced emf always opposes the change in magnetic flux that causes the emf, a result known as Lenz's law.

13.4 Motional Emf

- The relationship between an induced emf εε in a wire moving at a constant speed **v** through a magnetic field **B** is given by $\varepsilon = Blv$.
- An induced emf from Faraday's law is created from a motional emf that opposes the change in flux.

13.5 Induced Electric Fields

- A changing magnetic flux induces an electric field.
- Both the changing magnetic flux and the induced electric field are related to the induced emf from Faraday's law.

13.6 Eddy Currents

- Current loops induced in moving conductors are called eddy currents. They can create significant drag, called magnetic damping.
- Manipulation of eddy currents has resulted in applications such as metal detectors, braking in trains or roller coasters, and induction cooktops.

13.7 Electric Generators and Back Emf

- An electric generator rotates a coil in a magnetic field, inducing an emf given as a function of time by $\varepsilon = NBA\omega sin(\omega t)$ where **A** is the area of an **N**-turn coil rotated at a constant angular velocity ω in a uniform magnetic field $\vec{B}$.
- The peak emf of a generator is $\varepsilon_0 = NBA\omega$.
- Any rotating coil produces an induced emf. In motors, this is called back emf because it opposes the emf input to the motor.

13.8 Applications of Electromagnetic Induction

- Hard drives utilize magnetic induction to read/write information.
- Other applications of magnetic induction can be found in graphics tablets, electric and hybrid vehicles, and in transcranial magnetic stimulation.

9.12: Electromagnetic Induction (Exercises)

Conceptual Questions

13.2 Faraday's Law

1. A stationary coil is in a magnetic field that is changing with time. Does the emf induced in the coil depend on the actual values of the magnetic field?

2. In Faraday's experiments, what would be the advantage of using coils with many turns?

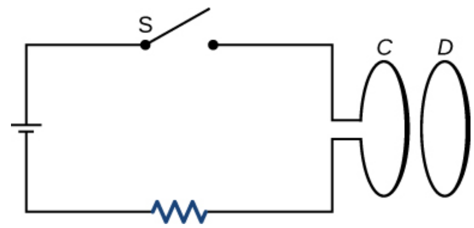
3. A copper ring and a wooden ring of the same dimensions are placed in magnetic fields so that there is the same change in magnetic flux through them. Compare the induced electric fields and currents in the rings.

4. Discuss the factors determining the induced emf in a closed loop of wire.

5. (a) Does the induced emf in a circuit depend on the resistance of the circuit?

(b) Does the induced current depend on the resistance of the circuit?

6. How would changing the radius of loop **D** shown below affect its emf, assuming **C** and **D** are much closer together compared to their radii?



7. Can there be an induced emf in a circuit at an instant when the magnetic flux through the circuit is zero?

8. Does the induced emf always act to decrease the magnetic flux through a circuit?

9. How would you position a flat loop of wire in a changing magnetic field so that there is no induced emf in the loop?

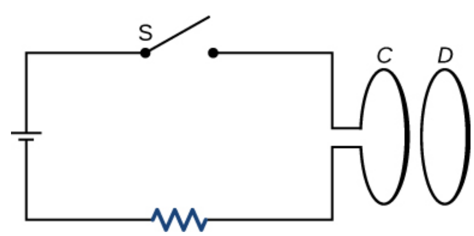
10. The normal to the plane of a single-turn conducting loop is directed at an angle **θ** to a spatially uniform magnetic field $vecB$. It has a fixed area and orientation relative to the magnetic field. Show that the emf induced in the loop is given by $\varepsilon = (dB/dt)(A cos\theta)$,where A is the area of the loop.

13.3 Lenz's Law

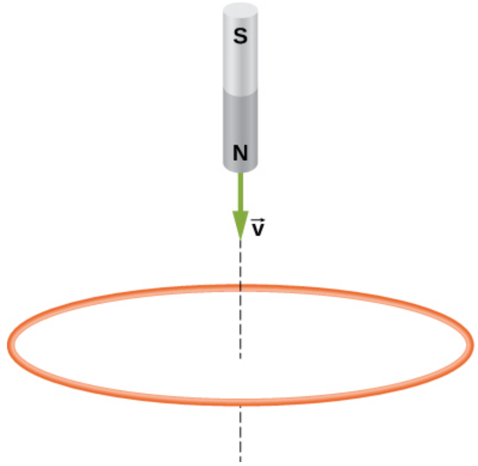
11. The circular conducting loops shown in the accompanying figure are parallel, perpendicular to the plane of the page, and coaxial.

(a) When the switch S is closed, what is the direction of the current induced in **D**?

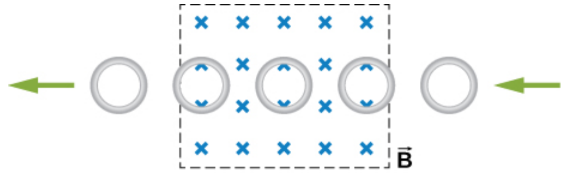
(b) When the switch is opened, what is the direction of the current induced in loop **D**?



12. The north pole of a magnet is moved toward a copper loop, as shown below. If you are looking at the loop from above the magnet, will you say the induced current is circulating clockwise or counterclockwise?

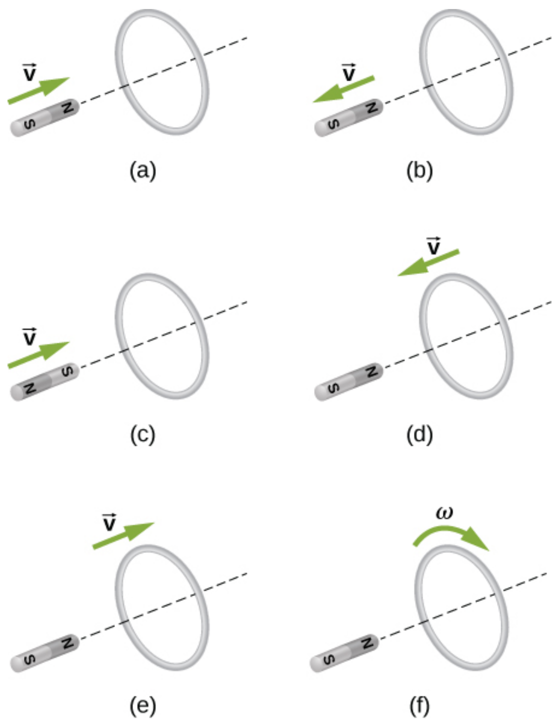


13. The accompanying figure shows a conducting ring at various positions as it moves through a magnetic field. What is the sense of the induced emf for each of those positions?



14. Show that **ε** and $d\Phi_m/dt$ have the same units.

15. State the direction of the induced current for each case shown below, observing from the side of the magnet.



13.4 Motional Emf

16. A bar magnet falls under the influence of gravity along the axis of a long copper tube. If air resistance is negligible, will there be a force to oppose the descent of the magnet? If so, will the magnet reach a terminal velocity?

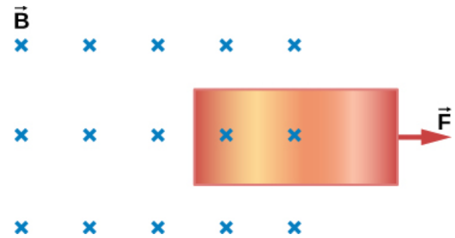
17. Around the geographic North Pole (or magnetic South Pole), Earth's magnetic field is almost vertical. If an airplane is flying northward in this region, which side of the wing is positively charged and which is negatively charged?

18. A wire loop moves translationally (no rotation) in a uniform magnetic field. Is there an emf induced in the loop?

13.5 Induced Electric Fields

19. Is the work required to accelerate a rod from rest to a speed v in a magnetic field greater than the final kinetic energy of the rod? Why?

20. The copper sheet shown below is partially in a magnetic field. When it is pulled to the right, a resisting force pulls it to the left. Explain. What happen if the sheet is pushed to the left?



13.6 Eddy Currents

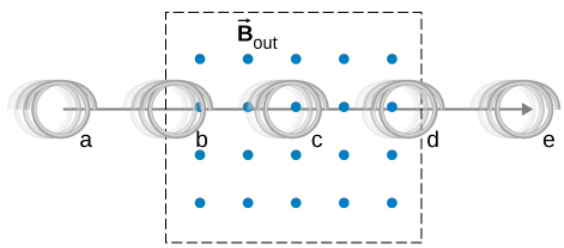
21. A conducting sheet lies in a plane perpendicular to a magnetic field $\vec{B}$ that is below the sheet. If $\vec{B}$ oscillates at a high frequency and the conductor is made of a material of low resistivity, the region above the sheet is effectively shielded from $\vec{B}$. Explain why. Will the conductor shield this region from static magnetic fields?

22. Electromagnetic braking can be achieved by applying a strong magnetic field to a spinning metal disk attached to a shaft.

(a) How can a magnetic field slow the spinning of a disk?

(b) Would the brakes work if the disk was made of plastic instead of metal?

23. A coil is moved through a magnetic field as shown below. The field is uniform inside the rectangle and zero outside. What is the direction of the induced current and what is the direction of the magnetic force on the coil at each position shown?



Problems

13.2 Faraday's Law

24. A 50-turn coil has a diameter of 15 cm. The coil is placed in a spatially uniform magnetic field of magnitude 0.50 T so that the face of the coil and the magnetic field are perpendicular. Find the magnitude of the emf induced in the coil if the magnetic field is reduced to zero uniformly in

(a) 0.10 s,

(b) 1.0 s, and

(c) 60 s.

25. Repeat your calculations of the preceding problem's time of 0.1 s with the plane of the coil making an angle of

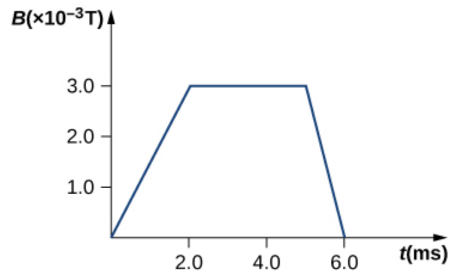
(a) **30°**,

(b) **60°,** and

(c) **90°** with the magnetic field.

26. A square loop whose sides are 6.0-cm long is made with copper wire of radius 1.0 mm. If a magnetic field perpendicular to the loop is changing at a rate of 5.0 mT/s, what is the current in the loop?

27. The magnetic field through a circular loop of radius 10.0 cm varies with time as shown below. The field is perpendicular to the loop. Plot the magnitude of the induced emf in the loop as a function of time.



28. The accompanying figure shows a single-turn rectangular coil that has a resistance of 2.0Ω.2.0Ω. The magnetic field at all points inside the coil varies according to $B = B_0 e^{-\alpha t}$, where $B_0 = 0.25T$ and **α=200Hz**. What is the current induced in the coil at

(a) **t=0.001s**,

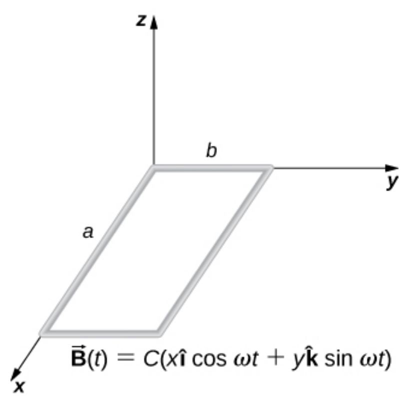
(b) **0.002 s**,

(c) **2.0 s**?

29. How would the answers to the preceding problem change if the coil consisted of 20 closely spaced turns?

30. A long solenoid with **n=10** turns per centimeter has a cross-sectional area of $5.0cm^2$ and carries a current of 0.25 A. A coil with five turns encircles the solenoid. When the current through the solenoid is turned off, it decreases to zero in 0.050 s. What is the average emf induced in the coil?

31. A rectangular wire loop with length a and width b lies in the xy-plane, as shown below. Within the loop there is a time-dependent magnetic field given by $\vec{B}(t) = C((x cos\omega t)\hat{i} + (y sin\omega t)\hat{k})$, with $\vec{B}(t)$ in tesla. Determine the emf induced in the loop as a function of time.



32. The magnetic field perpendicular to a single wire loop of diameter 10.0 cm decreases from 0.50 T to zero. The wire is made of copper and has a diameter of 2.0 mm and length 1.0 cm. How much charge moves through the wire while the field is changing?

13.3 Lenz's Law

33. A single-turn circular loop of wire of radius 50 mm lies in a plane perpendicular to a spatially uniform magnetic field. During a 0.10-s time interval, the magnitude of the field increases uniformly from 200 to 300 mT.

(a) Determine the emf induced in the loop.

(b) If the magnetic field is directed out of the page, what is the direction of the current induced in the loop?

34. When a magnetic field is first turned on, the flux through a 20-turn loop varies with time according to $\Phi_m = 5.0t^2 - 2.0t$, where Φ_m is in milliwebers, t is in seconds, and the loop is in the plane of the page with the unit normal pointing outward.

(a) What is the emf induced in the loop as a function of time? What is the direction of the induced current at

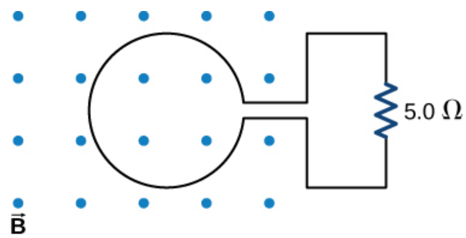
(b) t = 0,

(c) 0.10,

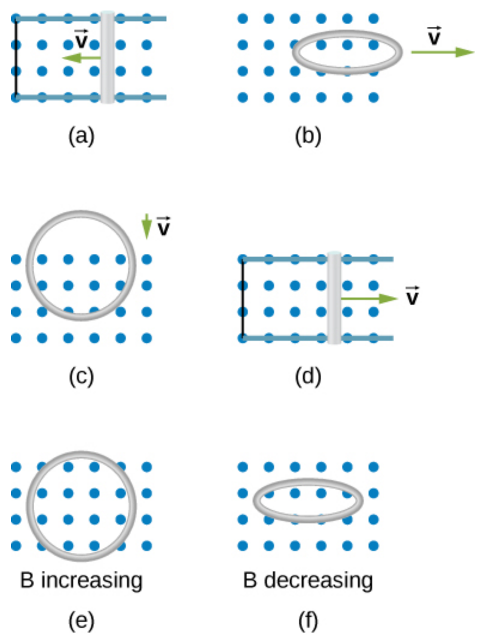
(d) 1.0, and

(e) 2.0 s?

35. The magnetic flux through the loop shown in the accompanying figure varies with time according to $\Phi_m = 2.00e^{-3t} sin(120\pi t)$,where Φ_m is in milliwebers. What are the direction and magnitude of the current through the **5.00-Ω** resistor at (a) **t=0**t=0; (b) $t = 2.17 \times 10^{-2} s$, and (c) **t=3.00s**?



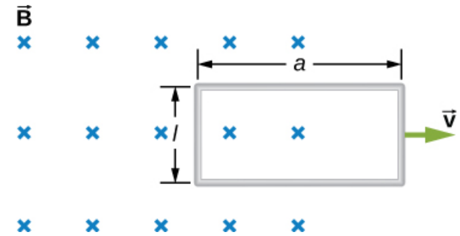
36. Use Lenz’s law to determine the direction of induced current in each case.



13.4 Motional Emf

37. An automobile with a radio antenna 1.0 m long travels at 100.0 km/h in a location where the Earth’s horizontal magnetic field is $5.5 \times 10^{-5} T$. What is the maximum possible emf induced in the antenna due to this motion?

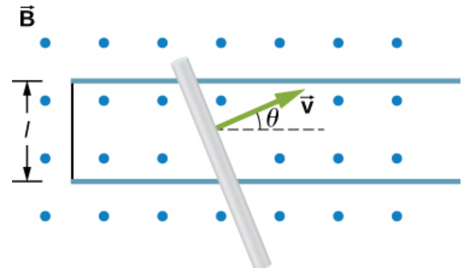
38. The rectangular loop of N turns shown below moves to the right with a constant velocity $\vec{v}$ while leaving the poles of a large electromagnet. (a) Assuming that the magnetic field is uniform between the pole faces and negligible elsewhere, determine the induced emf in the loop. (b) What is the source of work that produces this emf?



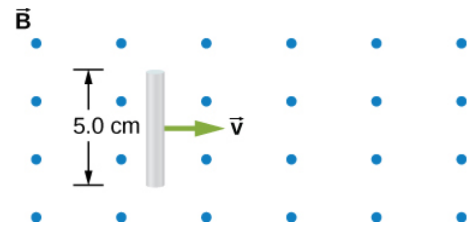
39. Suppose the magnetic field of the preceding problem oscillates with time according to $B = B_0 \sin \omega t$. What then is the emf induced in the loop when its trailing side is a distance d from the right edge of the magnetic field region?

40. A coil of 1000 turns encloses an area of$25cm^2$. It is rotated in 0.010 s from a position where its plane is perpendicular to Earth's magnetic field to one where its plane is parallel to the field. If the strength of the field is $6.0 \times 10^{-5}T$, what is the average emf induced in the coil?

41. In the circuit shown in the accompanying figure, the rod slides along the conducting rails at a constant velocity $\vec{v}$. The velocity is in the same plane as the rails and directed at an angle θθ to them. A uniform magnetic field $\vec{B}$ is directed out of the page. What is the emf induced in the rod?



42. The rod shown in the accompanying figure is moving through a uniform magnetic field of strength $B = 0.50T$ with a constant velocity of magnitude $v = 8.0m/s.$. What is the potential difference between the ends of the rod? Which end of the rod is at a higher potential?



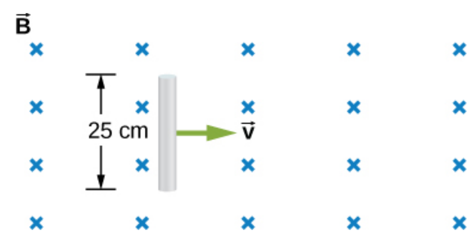
43. A 25-cm rod moves at 5.0 m/s in a plane perpendicular to a magnetic field of strength 0.25 T. The rod, velocity vector, and magnetic field vector are mutually perpendicular, as indicated in the accompanying figure. Calculate

(a) the magnetic force on an electron in the rod,

(b) the electric field in the rod, and

(c) the potential difference between the ends of the rod.

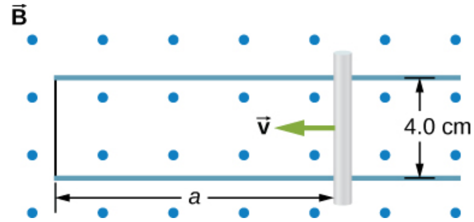
(d) What is the speed of the rod if the potential difference is 1.0 V?



44. In the accompanying figure, the rails, connecting end piece, and rod all have a resistance per unit length of **2.0Ω/cm**. The rod moves to the left at **v=3.0m/s**. If **B=0.75T** everywhere in the region, what is the current in the circuit

(a) when **a=8.0cm**?

(b) when **a=5.0cm**? Specify also the sense of the current flow.



45. The rod shown below moves to the right on essentially zero-resistance rails at a speed of **v=3.0m/s**. If **B=0.75T** everywhere in the region, what is the current through the **5.0-Ω** resistor? Does the current circulate clockwise or counterclockwise?

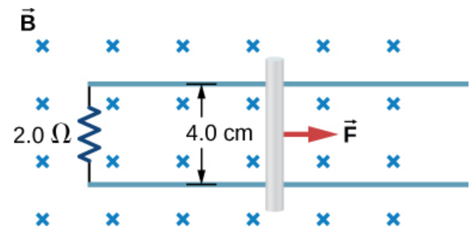
46. Shown below is a conducting rod that slides along metal rails. The apparatus is in a uniform magnetic field of strength 0.25 T, which is directly into the page. The rod is pulled to the right at a constant speed of 5.0 m/s by a force $\vec{F}$. The only significant resistance in the circuit comes from the **2.0-Ω** resistor shown.

(a) What is the emf induced in the circuit?

(b) What is the induced current? Does it circulate clockwise or counter clockwise?

(c) What is the magnitude of $\vec{F}$?

(d) What are the power output of $\vec{F}$ and the power dissipated in the resistor?



13.5 Induced Electric Fields

47. Calculate the induced electric field in a 50-turn coil with a diameter of 15 cm that is placed in a spatially uniform magnetic field of magnitude 0.50 T so that the face of the coil and the magnetic field are perpendicular. This magnetic field is reduced to zero in 0.10 seconds. Assume that the magnetic field is cylindrically symmetric with respect to the central axis of the coil.

48. The magnetic field through a circular loop of radius 10.0 cm varies with time as shown in the accompanying figure. The field is perpendicular to the loop. Assuming cylindrical symmetry with respect to the central axis of the loop, plot the induced electric field in the loop as a function of time.

49. The current I through a long solenoid with n turns per meter and radius R is changing with time as given by dI/dt. Calculate the induced electric field as a function of distance r from the central axis of the solenoid.

50. Calculate the electric field induced both inside and outside the solenoid of the preceding problem if $I = I_0 sin\omega t.$.

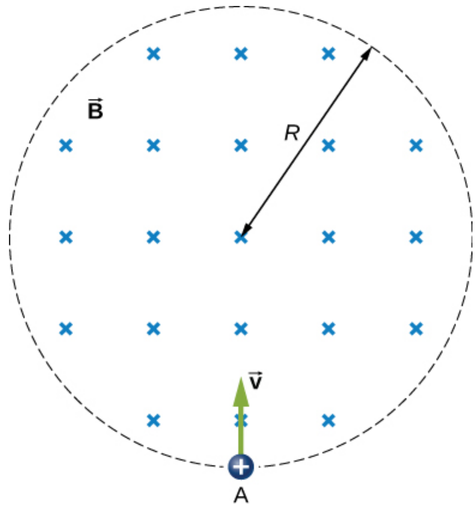
51. Over a region of radius R, there is a spatially uniform magnetic field $\vec{B}$. (See below.) At **t=0**, **B=1.0T**, after which it decreases at a constant rate to zero in 30 s.

(a) What is the electric field in the regions where $r \leq R$ and $r \geq R$ during that 30-s interval?

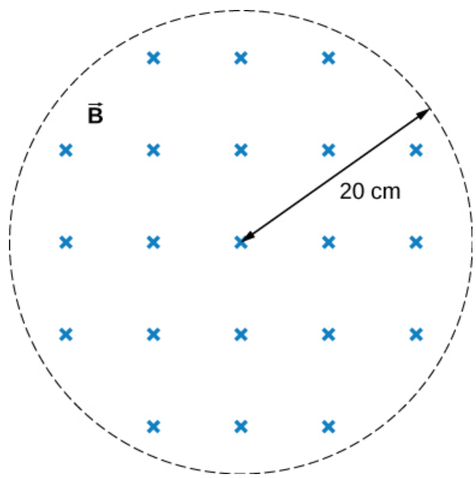
(b) Assume that **R=10.0cm**. How much work is done by the electric field on a proton that is carried once clock wise around a circular path of radius 5.0 cm?

(c) How much work is done by the electric field on a proton that is carried once counterclockwise around a circular path of any radius $r \geq R$?

(d) At the instant when **B=0.50T**, a proton enters the magnetic field at A, moving a velocity $\vec{v}(v=5.0\times10^6 m/s)$ as shown. What are the electric and magnetic forces on the proton at that instant?



52. The magnetic field at all points within the cylindrical region whose cross-section is indicated in the accompanying figure starts at 1.0 T and decreases uniformly to zero in 20 s. What is the electric field (both magnitude and direction) as a function of **r**, the distance from the geometric center of the region?



53. The current in a long solenoid of radius 3 cm is varied with time at a rate of 2 A/s. A circular loop of wire of radius 5 cm and resistance **2Ω** surrounds the solenoid. Find the electrical current induced in the loop.

54. The current in a long solenoid of radius 3 cm and 20 turns/cm is varied with time at a rate of 2 A/s. Find the electric field at a distance of 4 cm from the center of the solenoid.

13.7 Electric Generators and Back Emf

55. Design a current loop that, when rotated in a uniform magnetic field of strength 0.10 T, will produce an emf $\varepsilon=\varepsilon_0 sin\omega t,$, where $\varepsilon_0=110V$ and $\omega=120\pi rad/s$.

56. A flat, square coil of 20 turns that has sides of length 15.0 cm is rotating in a magnetic field of strength 0.050 T. If the maximum emf produced in the coil is 30.0 mV, what is the angular velocity of the coil?

57. A 50-turn rectangular coil with dimensions **0.15m×0.40m** rotates in a uniform magnetic field of magnitude 0.75 T at 3600 rev/min.

(a) Determine the emf induced in the coil as a function of time.

(b) If the coil is connected to a **1000-Ω** resistor, what is the power as a function of time required to keep the coil turning at 3600 rpm?

(c) Answer part (b) if the coil is connected to a **2000-Ω** resistor.

58. The square armature coil of an alternating current generator has 200 turns and is 20.0 cm on side. When it rotates at 3600 rpm, its peak output voltage is 120 V.

(a) What is the frequency of the output voltage?

(b) What is the strength of the magnetic field in which the coil is turning?

59. A flip coil is a relatively simple device used to measure a magnetic field. It consists of a circular coil of N turns wound with fine conducting wire. The coil is attached to a ballistic galvanometer, a device that measures the total charge that passes through it. The coil is placed in a magnetic field $\vec{B}$ such that its face is perpendicular to the field. It is then flipped through 180°,180°, and the total chargeQ that flows through the galvanometer is measured.

(a) If the total resistance of the coil and galvanometer is R, what is the relationship between B and Q? Because the coil is very small, you can assume that$\vec{B}$ is uniform over it.

(b) How can you determine whether or not the magnetic field is perpendicular to the face of the coil?

60. The flip coil of the preceding problem has a radius of 3.0 cm and is wound with 40 turns of copper wire. The total resistance of the coil and ballistic galvanometer is **0.20Ω**. When the coil is flipped through **180°** in a magnetic field $\vec{B}$, a change of 0.090 C flows through the ballistic galvanometer.

(a) Assuming that $\vec{B}$ and the face of the coil are initially perpendicular, what is the magnetic field?

(b) If the coil is flipped through **90°,** what is the reading of the galvanometer?

61. A 120-V, series-wound motor has a field resistance of 80 **Ω** and an armature resistance of 10 **Ω**. When it is operating at full speed, a back emf of 75 V is generated.

(a) What is the initial current drawn by the motor? When the motor is operating at full speed, where are

(b) the current drawn by the motor,

(c) the power output of the source,

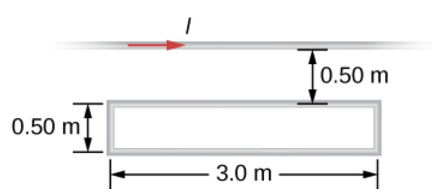
(d) the power output of the motor, and

(e) the power dissipated in the two resistances?

62. A small series-wound dc motor is operated from a 12-V car battery. Under a normal load, the motor draws 4.0 A, and when the armature is clamped so that it cannot turn, the motor draws 24 A. What is the back emf when the motor is operating normally?

Additional Problems

63. Shown in the following figure is a long, straight wire and a single-turn rectangular loop, both of which lie in the plane of the page. The wire is parallel to the long sides of the loop and is 0.50 m away from the closer side. At an instant when the emf induced in the loop is 2.0 V, what is the time rate of change of the current in the wire?



64. A metal bar of mass 500 g slides outward at a constant speed of 1.5 cm/s over two parallel rails separated by a distance of 30 cm which are part of a U-shaped conductor. There is a uniform magnetic field of magnitude 2 T pointing out of the page over the entire area. The railings and metal bar have an equivalent resistance of **150Ω**.

(a) Determine the induced current, both magnitude and direction.

(b) Find the direction of the induced current if the magnetic field is pointing into the page.

(c) Find the direction of the induced current if the magnetic field is pointed into the page and the bar moves inwards.

65. A current is induced in a circular loop of radius 1.5 cm between two poles of a horseshoe electromagnet when the current in the electromagnet is varied. The magnetic field in the area of the loop is perpendicular to the area and has a uniform magnitude. If the rate of change of magnetic field is 10 T/s, find the magnitude and direction of the induced current if resistance of the loop is **25Ω**.

66. A metal bar of length 25 cm is placed perpendicular to a uniform magnetic field of strength 3 T.

(a) Determine the induced emf between the ends of the rod when it is not moving.

(b) Determine the emf when the rod is moving perpendicular to its length and magnetic field with a speed of 50 cm/s.

67. A coil with 50 turns and area 10 cm^2 is oriented with its plane perpendicular to a 0.75-T magnetic field. If the coil is flipped over (rotated through **180°**) in 0.20 s, what is the average emf induced in it?

68. A 2-turn planer loop of flexible wire is placed inside a long solenoid of n turns per meter that carries a constant current I_0. The areaA of the loop is changed by pulling on its sides while ensuring that the plane of the loop always remains perpendicular to the axis of the solenoid. If **n=500turns** per meter, $I_0 = 20A$, and $A = 20cm^2$, what is the emf induced in the loop when **dA/dt=100**?

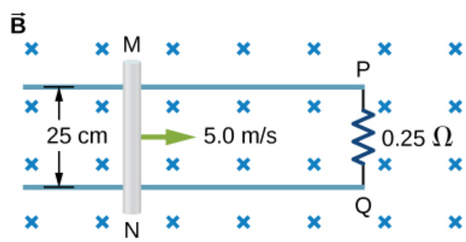
69. The conducting rod shown in the accompanying figure moves along parallel metal rails that are 25-cm apart. The system is in a uniform magnetic field of strength 0.75 T, which is directed into the page. The resistances of the rod and the rails are negligible, but the section PQ has a resistance of **0.25Ω**.

(a) What is the emf (including its sense) induced in the rod when it is moving to the right with a speed of 5.0 m/s?

(b) What force is required to keep the rod moving at this speed?

(c) What is the rate at which work is done by this force?

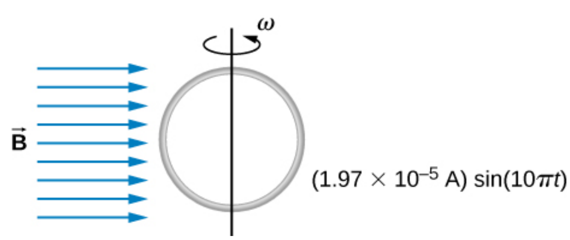
(d) What is the power dissipated in the resistor?



70. A circular loop of wire of radius 10 cm is mounted on a vertical shaft and rotated at a frequency of 5 cycles per second in a region of uniform magnetic field of 2 Gauss perpendicular to the axis of rotation.

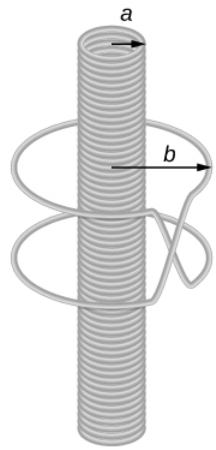
(a) Find an expression for the time-dependent flux through the ring.

(b) Determine the time-dependent current through the ring if it has a resistance of 10 **Ω**.



71. The magnetic field between the poles of a horseshoe electromagnet is uniform and has a cylindrical symmetry about an axis from the middle of the South Pole to the middle of the North Pole. The magnitude of the magnetic field changes as a rate of dB/dt due to the changing current through the electromagnet. Determine the electric field at a distance r from the center.

72. A long solenoid of radius a with n turns per unit length is carrying a time-dependent current $I(t) = I_0 sin(\omega t)$, where I_0 and **ω** are constants. The solenoid is surrounded by a wire of resistance **R** that has two circular loops of radius **b** with **b>a** (see the following figure). Find the magnitude and direction of current induced in the outer loops at time **t=0**.



73. A 120-V, series-wound dc motor draws 0.50 A from its power source when operating at full speed, and it draws 2.0 A when it starts. The resistance of the armature coils is **10Ω**.

(a) What is the resistance of the field coils?

(b) What is the back emf of the motor when it is running at full speed?

(c) The motor operates at a different speed and draws 1.0 A from the source. What is the back emf in this case?

74. The armature and field coils of a series-wound motor have a total resistance of **3.0Ω**. When connected to a 120-V source and running at normal speed, the motor draws 4.0 A.

(a) How large is the back emf?

(b) What current will the motor draw just after it is turned on? Can you suggest a way to avoid this large initial current?

Challenge Problems

75. A copper wire of length L is fashioned into a circular coil with **N** turns. When the magnetic field through the coil changes with time, for what value of N is the induced emf a maximum?

76. A 0.50-kg copper sheet drops through a uniform horizontal magnetic field of 1.5 T, and it reaches a terminal velocity of 2.0 m/s.

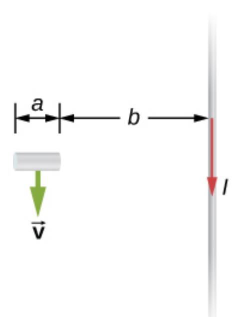
(a) What is the net magnetic force on the sheet after it reaches terminal velocity?

(b) Describe the mechanism responsible for this force.

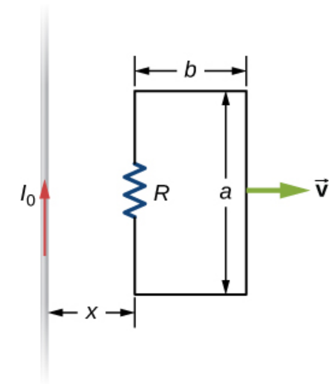
(c) How much power is dissipated as Joule heating while the sheet moves at terminal velocity?

77. A circular copper disk of radius 7.5 cm rotates at 2400 rpm around the axis through its center and perpendicular to its face. The disk is in a uniform magnetic field $\vec{B}$ of strength 1.2 T that is directed along the axis. What is the potential difference between the rim and the axis of the disk?

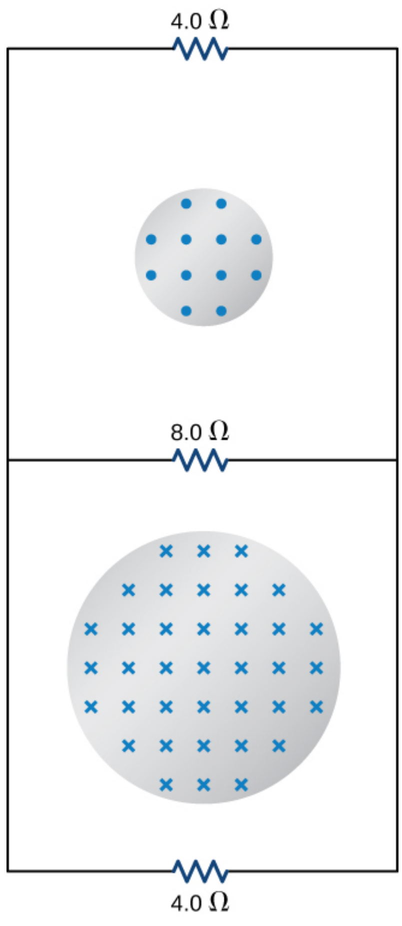
78. A short rod of length a moves with its velocity $\vec{v}$ parallel to an infinite wire carrying a current **I** (see below). If the end of the rod nearer the wire is a distance b from the wire, what is the emf induced in the rod?



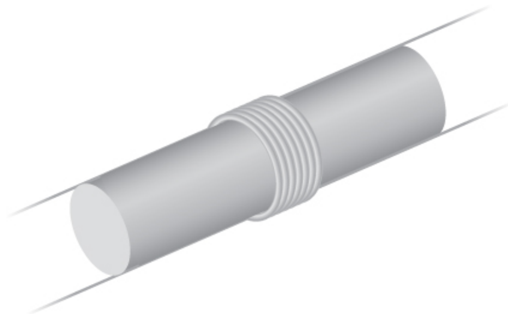
79. A rectangular circuit containing a resistance R is pulled at a constant velocity $\vec{v}$ away from a long, straight wire carrying a current I_0 (see below). Derive an equation that gives the current induced in the circuit as a function of the distance x between the near side of the circuit and the wire.



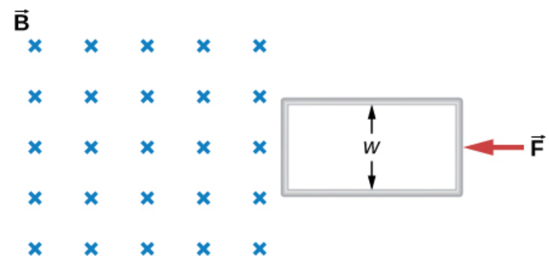
80. Two infinite solenoids cross the plane of the circuit as shown below. The radii of the solenoids are 0.10 and 0.20 m, respectively, and the current in each solenoid is changing such that **dB/dt=50.0T/s.** What are the currents in the resistors of the circuit?



81. An eight-turn coil is tightly wrapped around the outside of the long solenoid as shown below. The radius of the solenoid is 2.0 cm and it has 10 turns per centimeter. The current through the solenoid increases according to $I=I_0(1-e^{-\alpha t})$, where $I_0=4.0A$ and $\alpha=2.0\times10^{-2}s^{-1}$. What is the emf induced in the coil when (a) $t=0$, (b) $t=1.0\times10^2 s$, and (c) $t\to\infty$?



82. Shown below is a long rectangular loop of width **w**, length **l**, mass **m**, and resistance **R**. The loop starts from rest at the edge of a uniform magnetic field $\vec{B}$ and is pushed into the field by a constant force $\vec{F}$. Calculate the speed of the loop as a function of time.



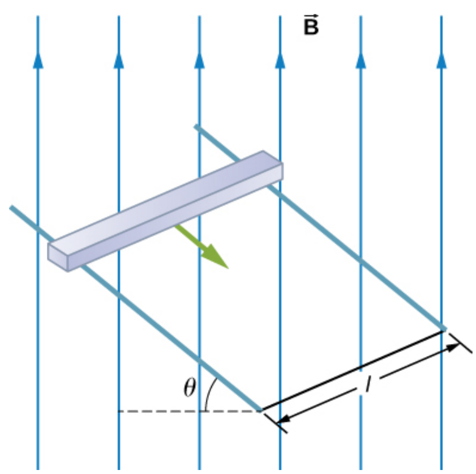
83. A square bar of mass **m** and resistance **R** is sliding without friction down very long, parallel conducting rails of negligible resistance (see below). The two rails are a distance l apart and are connected to each other at the bottom of the incline by a zero-resistance wire. The rails are inclined at an angle **θ**, and there is a uniform vertical magnetic field $\vec{B}$ throughout the region.

(a) Show that the bar acquires a terminal velocity given by $v = \frac{mgRsin\theta}{B^2l^2cos^2\theta}$.

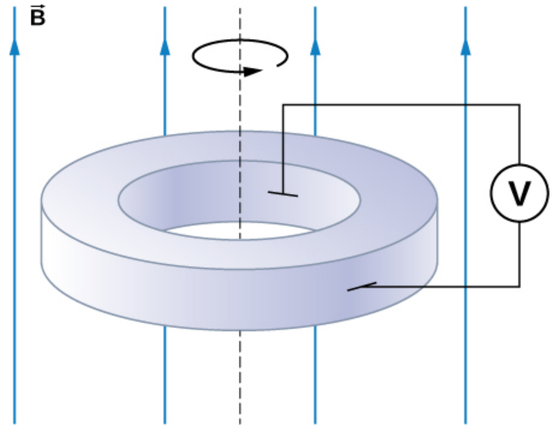
(b) Calculate the work per unit time done by the force of gravity.

(c) Compare this with the power dissipated in the Joule heating of the bar.

(d) What would happen if $\vec{B}$ were reversed?



84. The accompanying figure shows a metal disk of inner radius r_1 and other radius r_2 rotating at an angular velocity $\vec{\omega}$ while in a uniform magnetic field directed parallel to the rotational axis. The brush leads of a voltmeter are connected to the dark's inner and outer surfaces as shown. What is the reading of the voltmeter?



85. A long solenoid with 10 turns per centimeter is placed inside a copper ring such that both objects have the same central axis. The radius of the ring is 10.0 cm, and the radius of the solenoid is 5.0 cm.

(a) What is the emf induced in the ring when the current I through the solenoid is 5.0 A and changing at a rate of 100 A/s?

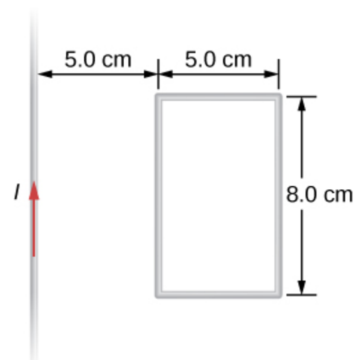
(b) What is the emf induced in the ring when **I=2.0A** and **dI/dt=100A/s**?

(c) What is the electric field inside the ring for these two cases?

(d) Suppose the ring is moved so that its central axis and the central axis of the solenoid are still parallel but no longer coincide. (You should assume that the solenoid is still inside the ring.) Now what is the emf induced in the ring?

(e) Can you calculate the electric field in the ring as you did in part (c)?

86. The current in the long, straight wire shown in the accompanying figure is given by $I=I_0 sin\omega t$, where $I_0=15A$ and $\omega=120\pi rad/s$. What is the current induced in the rectangular loop at (a) **t=0** and (b) $t=2.1\times10^{-3}s$? The resistance of the loop is **2.0Ω**.



87. A 500-turn coil with a $0.250-m^2$ area is spun in Earth's $5.00\times10^{-5}T$ magnetic field, producing a 12.0-kV maximum emf.

(a) At what angular velocity must the coil be spun?

(b) What is unreasonable about this result?

(c) Which assumption or premise is responsible?

88. A circular loop of wire of radius 10 cm is mounted on a vertical shaft and rotated at a frequency of 5 cycles per second in a region of uniform magnetic field of $2\times10^{-4}T$ perpendicular to the axis of rotation.

(a) Find an expression for the time-dependent flux through the ring

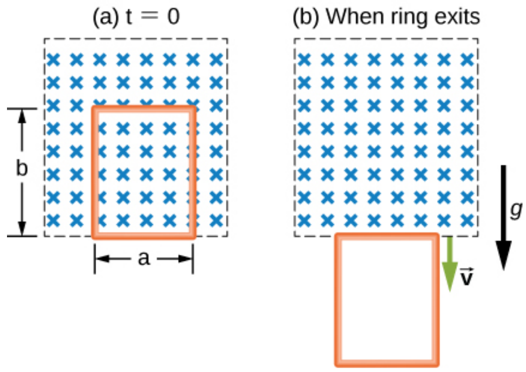
(b) Determine the time-dependent current through the ring if it has a resistance of **10Ω**.

89. A long solenoid of radius aa with nn turns per unit length is carrying a time-dependent current $I(t)=I_0 sin\omega t$ where I_0 and ωω are constants. The solenoid is surrounded by a wire of resistance R that has two circular loops of radius **b** with **b>a**. Find the magnitude and direction of current induced in the outer loops at time **t=0**.

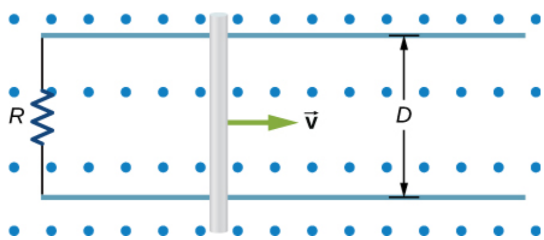
90. A rectangular copper loop of mass 100 g and resistance **0.2Ω** is in a region of uniform magnetic field that is perpendicular to the area enclosed by the ring and horizontal to Earth's surface (see below). The loop is let go from rest when it is at the edge of the nonzero magnetic field region.

(a) Find an expression for the speed when the loop just exits the region of uniform magnetic field.

(b) If it was let go at t=0t=0, what is the time when it exits the region of magnetic field for the following values: $a=25cm, b=50cm, B=3T, g=9.8m/s^2$? Assume that the magnetic field of the induced current is negligible compared to 3 T.



91. A metal bar of mass **m** slides without friction over two rails a distance **D** apart in the region that has a uniform magnetic field of magnitude B_0 and direction perpendicular to the rails (see below). The two rails are connected at one end to a resistor whose resistance is much larger than the resistance of the rails and the bar. The bar is given an initial speed of v_0. It is found to slow down. How far does the bar go before coming to rest? Assume that the magnetic field of the induced current is negligible compared to B_0.



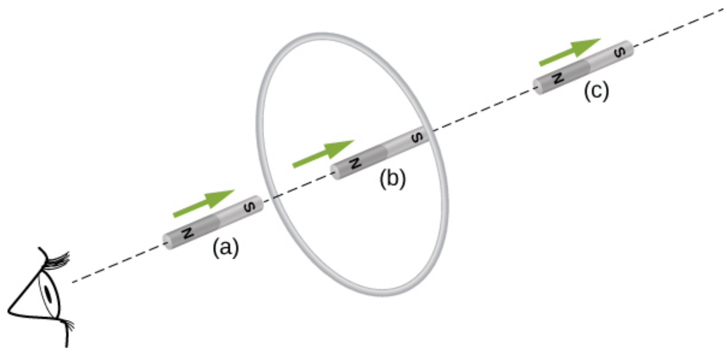
92. A time-dependent uniform magnetic field of magnitude **B(t)** is confined in a cylindrical region of radius **R**. A conducting rod of length **2D** is placed in the region, as shown below. Show that the emf between the ends of the rod is given by $\frac{dB}{dt}D\sqrt{R^2-D^2}$. (**Hint**: To find the emf between the ends, we need to integrate the electric field from one end to the other. To find the electric field, use Faraday's law as "Ampère's law for **E**.")

9.13: Electromagnetic Induction (Answers)

Check Your Understanding

13.1. 1.1 T/s

13.2. To the observer shown, the current flows clockwise as the magnet approaches, decreases to zero when the magnet is centered in the plane of the coil, and then flows counterclockwise as the magnet leaves the coil.



13.4. $\varepsilon = Bl^2\omega/2$, with **O** at a higher potential than **S**

13.5. 1.5 V

13.6. a. yes;

b. Yes; however there is a lack of symmetry between the electric field and coil, making $\oint \vec{E} \cdot d\vec{l}$ a more complicated relationship that can't be simplified as shown in the example.

13.7. $3.4 \times 10^{-3} V/m$

13.8. $P_1, P)2, P_4$

13.9. a. $3.1 \times 10^{-6} V$;

b. $2.0 \times 10^{-7} V/m$

Conceptual Questions

1. The emf depends on the rate of change of the magnetic field.

3. Both have the same induced electric fields; however, the copper ring has a much higher induced emf because it conducts electricity better than the wooden ring.

5. a. no; b. yes

7. As long as the magnetic flux is changing from positive to negative or negative to positive, there could be an induced emf.

9. Position the loop so that the field lines run perpendicular to the area vector or parallel to the surface.

11. a. CW as viewed from the circuit; b. CCW as viewed from the circuit

13. As the loop enters, the induced emf creates a CCW current while as the loop leaves the induced emf creates a CW current. While the loop is fully inside the magnetic field, there is no flux change and therefore no induced current.

15. a. CCW viewed from the magnet;

b. CW viewed from the magnet;

c. CW viewed from the magnet;

d. CCW viewed from the magnet;

e. CW viewed from the magnet;

f. no current

17. Positive charges on the wings would be to the west, or to the left of the pilot while negative charges would be pulled east or to the right of the pilot. Thus, the left hand tips of the wings would be positive and the right hand tips would be negative.

19. The work is greater than the kinetic energy because it takes energy to counteract the induced emf.

21. The conducting sheet is shielded from the changing magnetic fields by creating an induced emf. This induced emf creates an induced magnetic field that opposes any changes in magnetic fields from the field underneath. Therefore, there is no net magnetic field in the region above this sheet. If the field were due to a static magnetic field, no induced emf will be created since you need a changing magnetic flux to induce an emf. Therefore, this static magnetic field will not be shielded.

23. a. zero induced current, zero force; b. clockwise induced current, force is to the left; c. zero induced current, zero force; d. counterclockwise induced current, force is to the left; e. zero induced current, zero force.

Problems

25. a. 3.8 V;

b. 2.2 V;

c. 0 V

27. $B = 1.5t, 0 \leq t < 2.0ms, B = 3.0mT, 2.0ms \leq t \leq 5.0ms,$

$B = -3.0t + 18mT, 5.0ms < t \leq 6.0ms,$

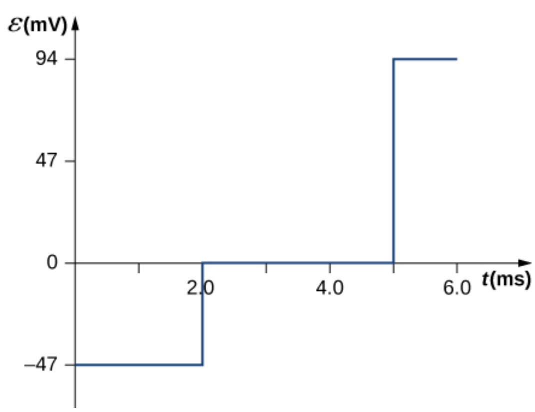
$\varepsilon = -\frac{d\Phi m}{dt} = -\frac{d(BA)}{dt} = -A\frac{dB}{dt},$

$\varepsilon = -\pi(0.100m)^2(1.5T/s)$

$= -47mV(0 \leq t < 2.0ms),$

$\varepsilon = \pi(0.100m)^2(0) = 0(2.0ms \leq t \leq 5.0ms),$

$\varepsilon = -\pi(0.100m)^2(-3.0T/s) = 94mV(5.0ms < t < 6.0ms).$



29. Each answer is 20 times the previously given answers.

31. $\hat{n} = \hat{k}, d\Phi_m = Cysin(\omega t)dxdy,$

$\Phi_m = \frac{Cab^2 sin(\omega t)}{2},$

$\varepsilon = -\frac{Cab^2 \omega cos(\omega t)}{2}.$

33. a. $7.8 \times 10^{-3} V$;

b. CCW from the same view as the magnetic field

35. a. 150 A downward through the resistor;

b. 232 A upward through the resistor;

c. 0.093 A downward through the resistor

37. 0.0015 V

39. $\varepsilon = -B_0 l d\omega \cos(\Omega t)\text{ld} + B_0 \sin(\Omega t)\text{lv}$

41. $\varepsilon = Blvcos\theta$

43. a. $2\times 10^{-19}T$;

b. 1.25 V/m;

c. 0.3125 V;

d. 16 m/s

45. 0.018 A, CW as seen in the diagram

47. 9.375 V/m

49. Inside, $B = \mu_0 nI, \oint \vec{E}\cdot d\vec{l} = (\pi r^2)\mu_0 n\frac{dI}{dt}$, so, $E = \frac{\mu_0 nr}{2}\cdot\frac{dI}{dt}$ (inside). Outside, $E(2\pi r) = \pi R^2\mu_0 n\frac{dI}{dt}$, so, $E = \frac{\mu_0 nR^2}{2r\cdot\frac{dI}{dt}}$ (outside)

51. a. $E_{inside} = \frac{r}{2}\frac{dB}{dt}, E_{outside} = \frac{r^2}{2R}\frac{dB}{dt}$;

b. $W = 4.19\times 10^{-23}J$;

c. 0 J;

d. $F_{mag} = 4\times 10^{-13}N, F_{elec} = 2.7\times 10^{-22}N$

53. $7.1\mu A$

55. Three turns with an area of $1m^2$

57. a. $\omega = 120\pi rad/s, \varepsilon = 850sin120\pi tV$;

b. $P = 720sin^2 120\pi tW$**;**

c. $P = 360sin^2 120\pi tW$

59. a. **B** is proportional to **Q**;

b. If the coin turns easily, the magnetic field is perpendicular. If the coin is at an equilibrium position, it is parallel.

61. a. 1.33 A;

b. 0.50 A;

c. 60 W;

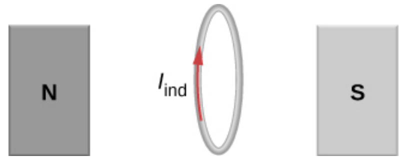
d. 22.5 W;

e. 2.5W

Additional Problems

63. 4.8×10^6 A/s

65. $2.83\times 10^{-4}A$, the direction as follows for increasing magnetic field:



67. 0.375 V

69. a. 0.94 V;

b. 0.70 N;

c. 3.52 J/s;

d. 3.52 W

71. $(\frac{dB}{dt})\frac{A}{2\pi r}$

73. a. $R_f + R_a = \frac{120V}{2.0A} = 60\Omega$, so $R_f = 50\Omega;$

b. $I = \frac{\varepsilon_s - \varepsilon_i}{R_f + R_a}, \Rightarrow \varepsilon_i = 90V$;

c. $\varepsilon_i = 60V$

Challenge Problems

75. N is a maximum number of turns allowed.

77. 5.3 V

79. $\Phi = \frac{\mu_0 I_0 a}{2\pi} ln(1 + \frac{b}{x})$, so $I = \frac{\mu_0 I_0 abv}{2\pi Rx(x+b)}\varepsilon = \frac{\mu_0 I_0 abv}{2\pi x(x+b)}$

81. a. $1.01 \times 10^{-6} V;$

b. $1.37 \times 10^{-7} V;$

c. 0 V

83. a. $v = \frac{mgRsin\theta}{B^2 l^2 cos^2\theta}$;

b. $mgvsin\theta;$

c. $mc\Delta T;$

d. current would reverse direction but bar would still slide at the same speed

85. a. $B = \mu_0 nI, \Phi_m = BA = \mu_0 nIA$,

$\varepsilon = 9.9 \times 10^{-4} V$;

b. $9.9 \times 10^{-4} V;$

c. $\oint \vec{E} \cdot d\vec{l} = \varepsilon, \Rightarrow E = 1.6 \times 10^{-3} V/m$

d. $9.9 \times 10^{-4} V;$

e. no, because there is no cylindrical symmetry

87. a. $1.92 \times 10^6 rad/s = 1.83 \times 10^7 rpm$;

b. This angular velocity is unreasonably high, higher than can be obtained for any mechanical system.

c. The assumption that a voltage as great as 12.0 kV could be obtained is unreasonable.

89. $\frac{2\mu_0 \pi a^2 I_0 n\omega}{R}$

91. $\frac{mRv_o}{B^2 D^2}$

- **13.11: Electromagnetic Induction (Answers)** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.

CHAPTER OVERVIEW

10: Inductance

10.1: Inductance

In Electromagnetic Induction, we discussed how a time-varying magnetic flux induces an source voltage in a circuit. In many of our calculations, this flux was due to an applied time-dependent magnetic field. The reverse of this phenomenon also occurs: The current flowing in a circuit produces its own magnetic field.

Figure 10.1.1: A smartphone charging mat contains a coil that receives alternating current, or current that is constantly increasing and decreasing. The varying current induces a source voltage in the smartphone, which charges its battery. Note that the black box containing the electrical plug also contains a transformer that modifies the current from the outlet to suit the needs of the smartphone. (credit: modification of work by "LG"/Flickr)

In the chapter on The Electric Field, we saw that induction is the process by which a source voltage is induced by changing electric flux and separation of a dipole. So far, we have discussed some examples of induction, although some of these applications are more effective than others. The smartphone charging mat in the chapter opener photo also works by induction. Is there a useful physical quantity related to how "effective" a given device is? The answer is yes, and that physical quantity is inductance. In this chapter, we look at the applications of inductance in electronic devices and how inductors are used in circuits.

10.2: Mutual Inductance

LEARNING OBJECTIVES

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

- Correlate two nearby circuits that carry time-varying currents with the emf induced in each circuit
- Describe examples in which mutual inductance may or may not be desirable

Inductance is the property of a device that tells us how effectively it induces an emf in another device. In other words, it is a physical quantity that expresses the effectiveness of a given device.

When two circuits carrying time-varying currents are close to one another, the magnetic flux through each circuit varies because of the changing current **I** in the other circuit. Consequently, an emf is induced in each circuit by the changing current in the other. This type of emf is therefore called a **mutually induced emf**, and the phenomenon that occurs is known as **mutual inductance (M)**. As an example, let's consider two tightly wound coils (Figure $10.2.1$). Coils 1 and 2 have N_1 and N_2 turns and carry currents I_1 and I_2 respectively. The flux through a single turn of coil 2 produced by the magnetic field of the current in coil 1 is Φ_{12}, whereas the flux through a single turn of coil 1 due to the magnetic field of I_2 is Φ_{12}.

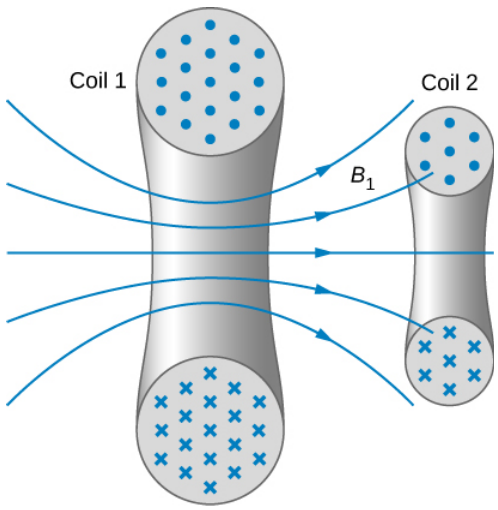


Figure 10.2.1: Some of the magnetic field lines produced by the current in coil 1 pass through coil 2.

The mutual inductance M_{21} of coil 2 with respect to coil 1 is the ratio of the flux through the N_2 turns of coil 2 produced by the magnetic field of the current in coil 1, divided by that current, that is,

$$M_{21}=\frac{N_2\Phi_{21}}{I_1}. \tag{10.2.1}$$

Similarly, the mutual inductance of coil 1 with respect to coil 2 is

$$M_{12}=\frac{N_1\Phi_{12}}{I_2}. \tag{10.2.2}$$

Like capacitance, mutual inductance is a geometric quantity. It depends on the shapes and relative positions of the two coils, and it is independent of the currents in the coils. The SI unit for mutual inductance **M** is called the **henry (H)** in honor of Joseph **Henry** (1799–1878), an American scientist who discovered induced emf independently of Faraday. Thus, we have $1\,H=1\,V\cdot s/A$. From Equations 10.2.1 and 10.2.2, we can show that $M_{21}=M_{12}$, so we usually drop the subscripts associated with mutual inductance and write

$$M=\frac{N_2\Phi_{21}}{I_1}=\frac{N_1\Phi_{12}}{I_2}. \tag{10.2.3}$$

The emf developed in either coil is found by combining Faraday's law and the definition of mutual inductance. Since $N_2\Phi_{21}$ is the total flux through coil 2 due to I_1, we obtain

$$\epsilon_2 = -\frac{d}{dt}(N_2\Phi_{21}) \tag{10.2.4}$$

$$= -\frac{d}{dt}(MI_1) \tag{10.2.5}$$

$$= -M\frac{dI_1}{dt} \tag{10.2.6}$$

where we have used the fact that M is a time-independent constant because the geometry is time-independent. Similarly, we have

$$\epsilon_1 = -M\frac{dI_2}{dt}. \tag{10.2.7}$$

In Equation 10.2.7, we can see the significance of the earlier description of mutual inductance (M) as a geometric quantity. The value of M neatly encapsulates the physical properties of circuit elements and allows us to separate the physical layout of the circuit from the dynamic quantities, such as the emf and the current. Equation 10.2.7 defines the mutual inductance in terms of properties in the circuit, whereas the previous definition of mutual inductance in Equation 10.2.1 is defined in terms of the magnetic flux experienced, regardless of circuit elements. You should be careful when using Equations 10.2.6 and 10.2.6 because ϵ_1 and ϵ_2 do not necessarily represent the total emfs in the respective coils. Each coil can also have an emf induced in it because of its **self-inductance** (self-inductance will be discussed in more detail in a later section).

A large mutual inductance **M** may or may not be desirable. We want a transformer to have a large mutual inductance. But an appliance, such as an electric clothes dryer, can induce a dangerous emf on its metal case if the mutual inductance between its coils and the case is large. One way to reduce mutual inductance is to counter-wind coils to cancel the magnetic field produced (Figure 10.2.2).

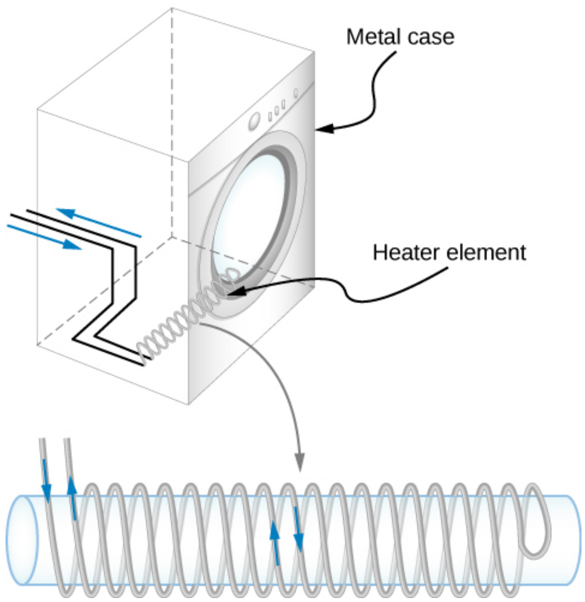


Figure 10.2.2: The heating coils of an electric clothes dryer can be counter-wound so that their magnetic fields cancel one another, greatly reducing the mutual inductance with the case of the dryer.

Digital signal processing is another example in which mutual inductance is reduced by counter-winding coils. The rapid on/off emf representing 1s and 0s in a digital circuit creates a complex time-dependent magnetic field. An emf can be generated in neighboring conductors. If that conductor is also carrying a digital signal, the induced emf may be large enough to switch 1s and 0s, with consequences ranging from inconvenient to disastrous.

Example 10.2.1: Mutual Inductance

Figure 10.2.3 shows a coil of N_2 turns and radius R_2 surrounding a long solenoid of length l_1, radius R_1, and N_1 turns.

a. What is the mutual inductance of the two coils?
b. If $N_1 = 500\,turns$, $N_2 = 10\,turns$, $R_1 = 3.10\,cm$, $l_1 = 75.0\,cm$, and the current in the solenoid is changing at a rate of 200 A/s, what is the emf induced in the surrounding coil?

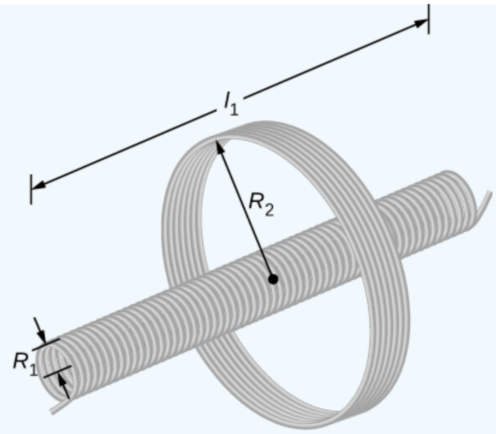


Figure 10.2.3: A solenoid surrounded by a coil.

Strategy

There is no magnetic field outside the solenoid, and the field inside has magnitude $B_1 = \mu_0(N_1/l_1)I_1$ and is directed parallel to the solenoid's axis. We can use this magnetic field to find the magnetic flux through the surrounding coil and then use this flux to calculate the mutual inductance for part (a), using Equation 10.2.3. We solve part (b) by calculating the mutual inductance from the given quantities and using Equation 10.2.6 to calculate the induced emf.

Solution

1. The magnetic flux Φ_{21} through the surrounding coil is

$$\begin{aligned}\Phi_{21} &= B_1 \pi R_1^2 \\ &= \frac{\mu_0 N_1 I_1}{l_1} \pi R_1^2.\end{aligned}$$

Now from Equation 10.2.3, the mutual inductance is

$$\begin{aligned}M &= \frac{N_2 \Phi_{21}}{I_1} \\ &= \left(\frac{N_2}{I_1}\right)\left(\frac{\mu_0 N_1 I_1}{l_1}\right)\pi R_1^2 \\ &= \frac{\mu_0 N_1 N_2 \pi R_1^2}{l_1}.\end{aligned}$$

2. Using the previous expression and the given values, the mutual inductance is

$$\begin{aligned}M &= \frac{(4\pi \times 10^{-7}\, T\cdot m/A)(500)(10)\pi(0.0310\, m)^2}{0.750\, m} \\ &= 2.53 \times 10^{-5}\, H.\end{aligned}$$

Thus, from Equation 10.2.6, the emf induced in the surrounding coil is

$$\begin{aligned}\epsilon_2 &= -M\frac{dI_1}{dt} \\ &= -(2.53 \times 10^{-5} H)(200\, A/s) \\ &= -5.06 \times 10^{-3} V.\end{aligned}$$

Significance

Notice that **M** in part (a) is independent of the radius R_2 of the surrounding coil because the solenoid's magnetic field is confined to its interior. In principle, we can also calculate **M** by finding the magnetic flux through the solenoid produced by the current in the surrounding coil. This approach is much more difficult because Φ_{12} is so complicated. However, since $M_{12} = M_{21}$, we do know the result of this calculation.

Exercise 10.2.1

A current $I(t) = (5.0\ A)\, sin\,((120\pi\, rad/s)t)$ flows through the solenoid of part (b) of Example $10.2.1$. What is the maximum emf induced in the surrounding coil?

Solution

$4.77 \times 10^{-2}\ V$

10.3: Self-Inductance and Inductors

Learning Objectives

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

- Correlate the rate of change of current to the induced source voltage created by that current in the same circuit.
- Derive the self-inductance for a cylindrical solenoid.
- Derive the self-inductance for a rectangular toroid.

Mutual inductance arises when a current in one circuit produces a changing magnetic field that induces an source voltage in another circuit. But can the magnetic field affect the current in the original circuit that produced the field? The answer is yes, and this is the phenomenon called **self-inductance**.

Inductors

Figure $10.3.1$ shows some of the magnetic field lines due to the current in a circular loop of wire. If the current is constant, the magnetic flux through the loop is also constant. However, if the current **I** were to vary with time—say, immediately after switch S is closed—then the magnetic flux Φ_m would correspondingly change. Then Faraday's law tells us that a source voltage ϵ would be induced in the circuit, where

$$\epsilon = -\frac{d\Phi_m}{dt}. \tag{10.3.1}$$

Since the magnetic field due to a current-carrying wire is directly proportional to the current, the flux due to this field is also proportional to the current; that is,

$$\Phi_m \propto I. \tag{10.3.2}$$

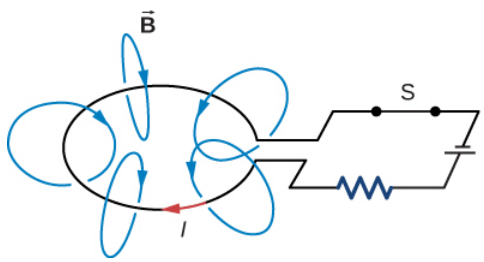


Figure 10.3.1: A magnetic field is produced by the current **I** in the loop. If **I** were to vary with time, the magnetic flux through the loop would also vary and a source voltage would be induced in the loop.

This can also be written as

$$\Phi_m = LI \tag{10.3.3}$$

where the constant of proportionality **L** is known as the **self-inductance** of the wire loop. If the loop has **N** turns, this equation becomes

$$\boxed{N\Phi_m = LI} \tag{10.3.4}$$

By convention, the positive sense of the normal to the loop is related to the current by the right-hand rule, so in Figure $10.3.1$, the normal points downward. With this convention, Φ_m is positive in Equation $10.3.4$, so **L always has a positive value**.

For a loop with **N** turns, $\epsilon = -N d\Phi_m/dt$, so the induced source voltage may be written in terms of the self-inductance as

$$\boxed{\epsilon = -L\frac{dI}{dt}.} \tag{10.3.5}$$

When using this equation to determine **L**, it is easiest to ignore the signs of ϵ and dI/dt, and calculate **L** as

$$L = \frac{|\epsilon|}{|dI/dt|}. \tag{10.3.6}$$

Since self-inductance is associated with the magnetic field produced by a current, any configuration of conductors possesses self-inductance. For example, besides the wire loop, a long, straight wire has self-inductance, as does a coaxial cable. A coaxial cable is most commonly used by the cable television industry and may also be found connecting to your cable modem. Coaxial cables are used due to their ability to transmit electrical signals with minimal distortions. Coaxial cables have two long cylindrical conductors that possess current and a self-inductance that may have undesirable effects.

Figure $10.3.2$: Symbol used to represent an inductor in a circuit.

A circuit element used to provide self-inductance is known as an **inductor**. It is represented by the symbol shown in Figure $10.3.2$, which resembles a coil of wire, the basic form of the inductor. Figure $10.3.3$ shows several types of inductors commonly used in circuits.

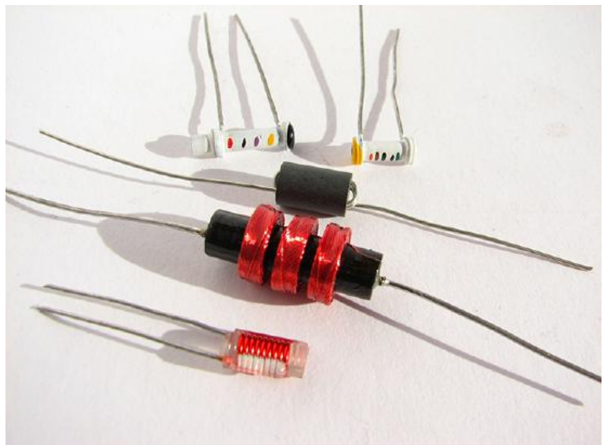


Figure $10.3.3$: A variety of inductors. Whether they are encapsulated like the top three shown or wound around in a coil like the bottom-most one, each is simply a relatively long coil of wire. (credit: Windell Oskay)

In accordance with Lenz's law, the negative sign in Equation $10.3.5$ indicates that the induced source voltage across an inductor always has a polarity that **opposes** the change in the current. For example, if the current flowing from **A** to **B** in Figure $10.3.4a$ were increasing, the induced source voltage (represented by the imaginary battery) would have the polarity shown in order to oppose the increase. If the current from **A** to **B** were decreasing, then the induced source voltage would have the opposite polarity, again to oppose the change in current (Figure $10.3.4b$). Finally, if the current through the inductor were constant, no source voltage would be induced in the coil.

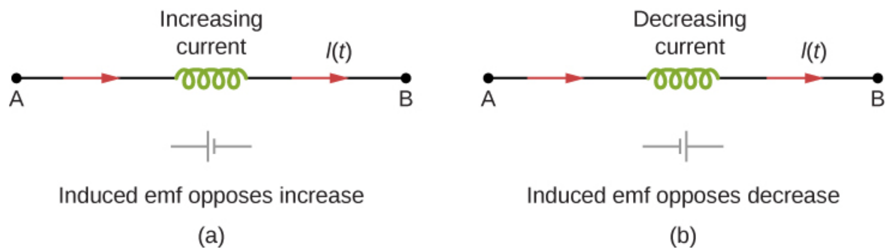


Figure $10.3.4$: The induced source voltage across an inductor always acts to oppose the change in the current. This can be visualized as an imaginary battery causing current to flow to oppose the change in (a) and reinforce the change in (b).

One common application of inductance is to allow **traffic signals** to sense when vehicles are waiting at a street intersection. An electrical circuit with an inductor is placed in the road underneath the location where a waiting car will stop. The body of the car increases the inductance and the circuit changes, sending a signal to the traffic lights to change colors. Similarly, **metal detectors** used for airport security employ the same technique. A coil or inductor in the metal detector frame acts as both a transmitter and a receiver. The pulsed signal from the transmitter coil induces a signal in the receiver. The self-inductance of the circuit is affected by any metal object in the path (Figure $10.3.5$). Metal detectors can be adjusted for sensitivity and can also sense the presence of metal on a person.

Figure 10.3.5: The familiar security gate at an airport not only detects metals, but can also indicate their approximate height above the floor. (credit: “Alexbuirds”/Wikimedia Commons)

Large induced voltages are found in **camera flashes**. Camera flashes use a battery, two inductors that function as a transformer, and a switching system or **oscillator** to induce large voltages. Recall from Oscillations on oscillations that “oscillation” is defined as the fluctuation of a quantity, or repeated regular fluctuations of a quantity, between two extreme values around an average value. Also recall (from Electromagnetic Induction on electromagnetic induction) that we need a changing magnetic field, brought about by a changing current, to induce a voltage in another coil. The oscillator system does this many times as the battery voltage is boosted to over 1000 volts. (You may hear the high-pitched whine from the transformer as the capacitor is being charged.) A capacitor stores the high voltage for later use in powering the flash.

Example 10.3.1:Self-Inductance of a Coil

An induced source voltage of 2.0 V is measured across a coil of 50 closely wound turns while the current through it increases uniformly from 0.0 to 5.0 A in 0.10 s. (a) What is the self-inductance of the coil? (b) With the current at 5.0 A, what is the flux through each turn of the coil?

Strategy

Both parts of this problem give all the information needed to solve for the self-inductance in part (a) or the flux through each turn of the coil in part (b). The equations needed are Equation 10.3.5 for part (a) and Equation 10.3.4 for part (b).

Solution

1. Ignoring the negative sign and using magnitudes, we have, from Equation 10.3.5,

$$L=\frac{\epsilon}{dI/dt}=\frac{2.0\,V}{5.0\,A/0.10\,s}=4.0\times10^{-2}H. \tag{10.3.7}$$

2. From Equation 10.3.4, the flux is given in terms of the current by $\Phi_m=LI/N$, so

$$\Phi_m=\frac{(4.0\times10^{-2}H)(5.0\,A)}{50\,turns)}=4.0\times10^{-3}Wb. \tag{10.3.8}$$

Significance

The self-inductance and flux calculated in parts (a) and (b) are typical values for coils found in contemporary devices. If the current is not changing over time, the flux is not changing in time, so no source voltage is induced.

Exercise 10.3.1

Current flows through the inductor in Figure 10.3.4 from **B** to **A** instead of from **A** to **B** as shown. Is the current increasing or decreasing in order to produce the source voltage given in diagram (a)? In diagram (b)?

Answer

a. decreasing; b. increasing; Since the current flows in the opposite direction of the diagram, in order to get a positive source voltage on the left-hand side of diagram (a), we need to decrease the current to the left, which creates a reinforced source voltage where the positive end is on the left-hand side. To get a positive source voltage on the right-hand side of diagram (b), we need to increase the current to the left, which creates a reinforced source voltage where the positive end is on the right-hand side.

Exercise 10.3.2

A changing current induces a source voltage of 10 V across a 0.25-H inductor. What is the rate at which the current is changing?

Answer

40 A/s

A good approach for calculating the self-inductance of an inductor consists of the following steps:

Problem-Solving Strategy: Self-Inductance

1. Assume a current **I** is flowing through the inductor.
2. Determine the magnetic field $\vec{B}$ produced by the current. If there is appropriate symmetry, you may be able to do this with Ampère's law.
3. Obtain the magnetic flux, Φ_m.
4. With the flux known, the self-inductance can be found from Equation 10.3.4, $L = N\Phi_m/I$.

To demonstrate this procedure, we now calculate the self-inductances of two inductors.

Cylindrical Solenoid

Consider a long, cylindrical solenoid with length **l**, cross-sectional area **A**, and **N** turns of wire. We assume that the length of the solenoid is so much larger than its diameter that we can take the magnetic field to be $B = \mu_0 nI$ throughout the interior of the solenoid, that is, we ignore end effects in the solenoid. With a current **I** flowing through the coils, the magnetic field produced within the solenoid is

$$B = \mu_0 \left(\frac{N}{l}\right) I, \tag{10.3.9}$$

so the magnetic flux through one turn is

$$\Phi_m = BA = \frac{\mu_0 NA}{l} I. \tag{10.3.10}$$

Using Equation 10.3.4, we find for the self-inductance of the solenoid,

Note

$$L_{solenoid} = \frac{N\Phi_m}{I} = \frac{\mu_0 N^2 A}{l}. \tag{10.3.11}$$

If $n = N/l$ is the number of turns per unit length of the solenoid, we may write Equation 10.3.11 as

$$L = \mu_0 \left(\frac{N}{l}\right)^2 Al = \mu_0 n^2 Al = \mu_0 n^2 (V), \tag{10.3.12}$$

where $V = Al$ is the volume of the solenoid. Notice that **the self-inductance of a long solenoid depends only on its physical properties** (such as the number of turns of wire per unit length and the volume), and not on the magnetic field or the current. This is true for inductors in general.

Rectangular Toroid

A toroid with a rectangular cross-section is shown in Figure $10.3.6$. The inner and outer radii of the toroid are R_1 and R_2, and h is the height of the toroid. Applying Ampère's law in the same manner as we did in Example 13.5.2 for a toroid with a circular cross-section, we find the magnetic field inside a rectangular toroid is also given by

$$B = \frac{\mu_0 NI}{2\pi r}, \tag{10.3.13}$$

where **r** is the distance from the central axis of the toroid. Because the field changes within the toroid, we must calculate the flux by integrating over the toroid's cross-section. Using the infinitesimal cross-sectional area element $da = h\,dr$ shown in Figure $10.3.6$, we obtain

$$\Phi_m = \int B\,da = \int_{R_1}^{R_2} \left(\frac{\mu_0 NI}{2\pi r}\right)(hdr) = \frac{\mu_0 NhI}{2\pi} ln \frac{R_2}{R_1}. \tag{10.3.14}$$

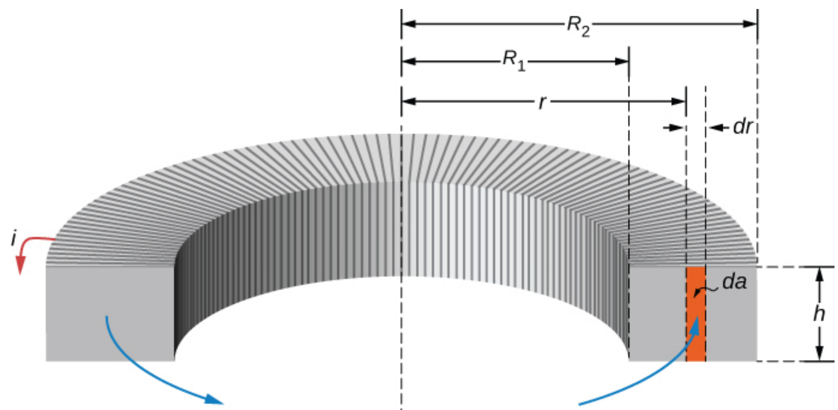


Figure 10.3.6: Calculating the self-inductance of a rectangular toroid.

Now from Equation 10.3.14, we obtain for the self-inductance of a rectangular toroid

Note

$$L = \frac{N\Phi_m}{I} = \frac{\mu_0 N^2 h}{2\pi} ln \frac{R_2}{R_1}. \tag{10.3.15}$$

As expected, the self-inductance is a constant determined by only the physical properties of the toroid.

Exercise 10.3.3

(a) Calculate the self-inductance of a solenoid that is tightly wound with wire of diameter 0.10 cm, has a cross-sectional area of $0.90\,cm^2$, and is 40 cm long. (b) If the current through the solenoid decreases uniformly from 10 to 0 A in 0.10 s, what is the source voltage induced between the ends of the solenoid?

Answer

a. $4.5 \times 10^{-5} H$; b. $4.5 \times 10^{-3} V$

Exercise 10.3.4

(a) What is the magnetic flux through one turn of a solenoid of self-inductance $8.0 \times 10^{-5} H$ when a current of 3.0 A flows through it? Assume that the solenoid has 1000 turns and is wound from wire of diameter 1.0 mm. (b) What is the cross-sectional area of the solenoid?

Answer

a. $2.4 \times 10^{-7} Wb$; b. $6.4 \times 10^{-5} m^2$

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).

10.4: Energy in a Magnetic Field

Learning Objectives

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

- Explain how energy can be stored in a magnetic field
- Derive the equation for energy stored in a coaxial cable given the magnetic energy density

The energy of a capacitor is stored in the electric field between its plates. Similarly, an inductor has the capability to store energy, but in its magnetic field. This energy can be found by integrating the **magnetic energy density**,

$$u_m = \frac{B^2}{2\mu_0}$$

over the appropriate volume. To understand where this formula comes from, let's consider the long, cylindrical solenoid of the previous section. Again using the infinite solenoid approximation, we can assume that the magnetic field is essentially constant and given by $B = \mu_0 nI$ everywhere inside the solenoid. Thus, the energy stored in a solenoid or the magnetic energy density times volume is equivalent to

$$U = u_m(V) = \frac{(\mu_0 nI)^2}{2\mu_0}(Al) = \frac{1}{2}(\mu_0 n^2 Al)I^2. \tag{10.4.1}$$

With the substitution of Equation 14.3.12, this becomes

$$\boxed{U = \frac{1}{2}LI^2.}$$

Although derived for a special case, this equation gives the energy stored in the magnetic field of **any** inductor. We can see this by considering an arbitrary inductor through which a changing current is passing. At any instant, the magnitude of the induced emf is $\epsilon = Ldi/dt$, where i is the induced current at that instance. Therefore, the power absorbed by the inductor is

$$P = \epsilon i = L\frac{di}{dt}i.$$

The total energy stored in the magnetic field when the current increases from 0 to **I** in a time interval from 0 to **t** can be determined by integrating this expression:

$$U = \int_0^t P dt' = \int_0^t L\frac{di}{dt'}i dt' = L\int_0^l i di = \frac{1}{2}LI^2. \tag{10.4.2}$$

Example 10.4.1: Self-Inductance of a Coaxial Cable

Figure $10.4.1$ shows two long, concentric cylindrical shells of radii R_1 and R_2. As discussed in Capacitance on capacitance, this configuration is a simplified representation of a coaxial cable. The capacitance per unit length of the cable has already been calculated. Now (a) determine the magnetic energy stored per unit length of the coaxial cable and (b) use this result to find the self-inductance per unit length of the cable.

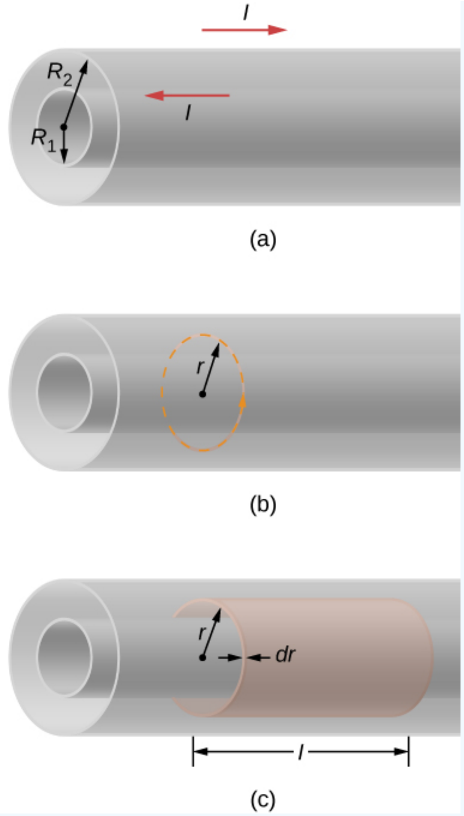


Figure 10.4.1: (a) A coaxial cable is represented here by two hollow, concentric cylindrical conductors along which electric current flows in opposite directions. (b) The magnetic field between the conductors can be found by applying Ampère’s law to the dashed path. (c) The cylindrical shell is used to find the magnetic energy stored in a length **l** of the cable.

Strategy

The magnetic field both inside and outside the coaxial cable is determined by Ampère’s law. Based on this magnetic field, we can use Equation 10.4.2 to calculate the energy density of the magnetic field. The magnetic energy is calculated by an integral of the magnetic energy density times the differential volume over the cylindrical shell. After the integration is carried out, we have a closed-form solution for part (a). The self-inductance per unit length is determined based on this result and Equation 10.4.2.

Solution

1. We determine the magnetic field between the conductors by applying Ampère’s law to the dashed circular path shown in Figure $10.4.1b$. Because of the cylindrical symmetry, $\vec{B}$ is constant along the path, and

$$\oint \vec{B} \cdot d\vec{l} = B(2\pi r) = \mu_0 I.$$

This gives us

$$B = \frac{\mu_0 I}{2\pi r}.$$

In the region outside the cable, a similar application of Ampère’s law shows that $B = 0$, since no net current crosses the area bounded by a circular path where $r > R_2$. This argument also holds when $r < R_1$; that is, in the region within the inner cylinder. All the magnetic energy of the cable is therefore stored between the two conductors. Since the energy density of the magnetic field is

$$u_m = \frac{B^2}{2\mu_0}$$

the energy stored in a cylindrical shell of inner radius **r**, outer radius $r + dr$ and length **l** (see part (c) of the figure) is

$$u_m = \frac{\mu_0 I^2}{8\pi^2 r^2}.$$

Thus, the total energy of the magnetic field in a length **l** of the cable is

$$U = \int_{R_1}^{R_2} dU = \int_{R_1}^{R_2} \frac{\mu_0 I^2}{8\pi^2 r^2}(2\pi r l)dr = \frac{\mu_0 I^2 l}{4\pi} ln \frac{R_2}{R_1},$$

and the energy per unit length is $(\mu_0 I^2/4\pi) ln(R_2/R_1)$.

2. From Equation 10.4.2,

$$U = \frac{1}{2} L I^2,$$

where **L** is the self-inductance of a length **l** of the coaxial cable. Equating the previous two equations, we find that the self-inductance per unit length of the cable is

$$\frac{L}{l} = \frac{\mu_0}{2\pi} \ln \frac{R_2}{R_1}.$$

Significance

The inductance per unit length depends only on the inner and outer radii as seen in the result. To increase the inductance, we could either increase the outer radius (R_2) or decrease the inner radius (R_1). In the limit as the two radii become equal, the inductance goes to zero. In this limit, there is no coaxial cable. Also, the magnetic energy per unit length from part (a) is proportional to the square of the current.

Exercise 10.4.1

How much energy is stored in the inductor of Example 14.3.1 after the current reaches its maximum value?

Solution

0.50 J

10.5: RL Circuits

Learning Objectives

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

- Analyze circuits that have an inductor and resistor in series
- Describe how current and voltage exponentially grow or decay based on the initial conditions

A circuit with resistance and self-inductance is known as an **RL circuit**. Figure $10.5.1a$ shows an **RL** circuit consisting of a resistor, an inductor, a constant source of emf, and switches S_1 and S_2. When S_1 is closed, the circuit is equivalent to a single-loop circuit consisting of a resistor and an inductor connected across a source of emf (Figure $10.5.1b$). When S_1 is opened and S_2 is closed, the circuit becomes a single-loop circuit with only a resistor and an inductor (Figure $10.5.1c$).

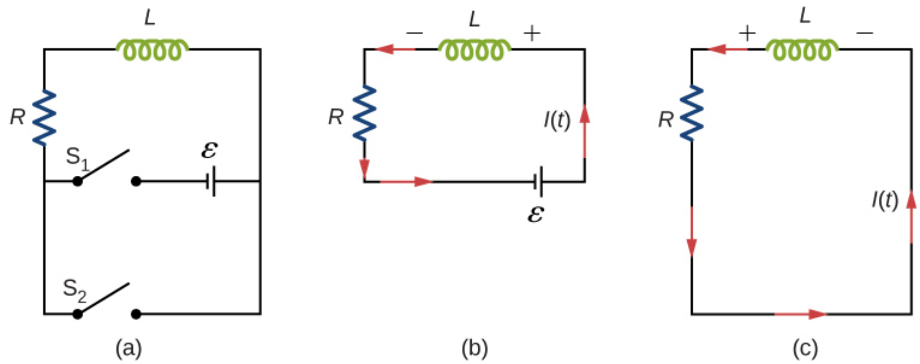


Figure 10.5.1: (a) An RL circuit with switches S_1 and S_2. (b) The equivalent circuit with S_1 closed and S_2 open. (c) The equivalent circuit after S_1 is opened and S_2 is closed.

We first consider the **RL** circuit of Figure $10.5.1b$. Once S_1 is closed and S_2 is open, the source of emf produces a current in the circuit. If there were no self-inductance in the circuit, the current would rise immediately to a steady value of ϵ/R. However, from Faraday's law, the increasing current produces an emf $V_L = -L(dI/dt)$ across the inductor. In accordance with Lenz's law, the induced emf counteracts the increase in the current and is directed as shown in the figure. As a result, $I(t)$ starts at zero and increases asymptotically to its final value.

Applying Kirchhoff's loop rule to this circuit, we obtain

$$\epsilon - L\frac{dI}{dt} - IR = 0, \tag{10.5.1}$$

which is a first-order differential equation for $I(t)$. Notice its similarity to the equation for a capacitor and resistor in series (see **RC Circuits**). Similarly, the solution to Equation 10.5.1 can be found by making substitutions in the equations relating the capacitor to the inductor. This gives

$$I(t) = \frac{\epsilon}{R}(1 - e^{-Rt/L}) \tag{10.5.2}$$

$$= \frac{\epsilon}{R}(1 - e^{-t/\tau_L}), \tag{10.5.3}$$

where

$$\tau_L = L/R \tag{10.5.4}$$

is the **inductive time constant** of the circuit.

The current $I(t)$ is plotted in Figure $10.5.2a$. It starts at zero, and as $t \to \infty$, $I(t)$ approaches ϵ/R asymptotically. The induced emf $V_L(t)$ is directly proportional to dI/dt, or the slope of the curve. Hence, while at its greatest immediately after the switches are thrown, the induced emf decreases to zero with time as the current approaches its final value of ϵ/R. The circuit then becomes equivalent to a resistor connected across a source of emf.

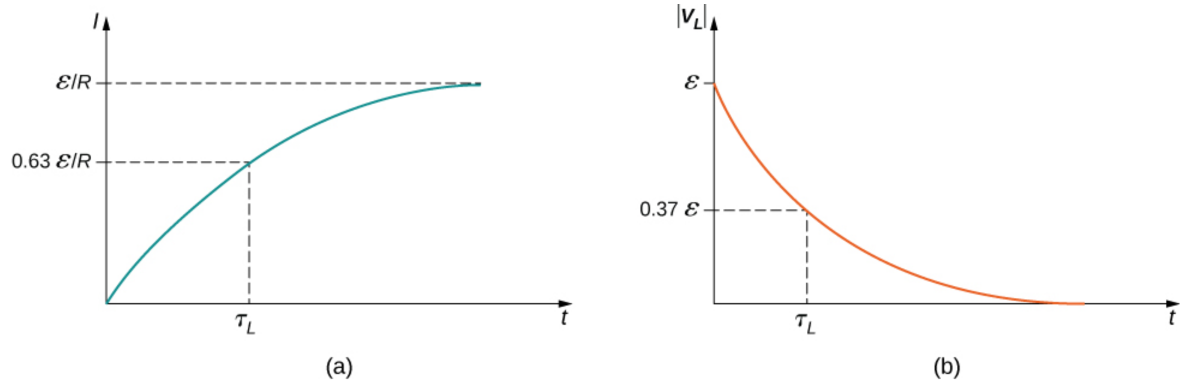


Figure 10.5.2: Time variation of (a) the electric current and (b) the magnitude of the induced voltage across the coil in the circuit of Figure 10.5.1b.

The energy stored in the magnetic field of an inductor is

$$U_L = \frac{1}{2} L I^2.$$

Thus, as the current approaches the maximum current ϵ/R, the stored energy in the inductor increases from zero and asymptotically approaches a maximum of $L(\epsilon/R)^2/2$.

The time constant τ_L tells us how rapidly the current increases to its final value. At $t = \tau_L$, the current in the circuit is, from Equation 10.5.3,

$$I(\tau_L) = \frac{\epsilon}{R}(1 - e^{-1}) = 0.63\frac{\epsilon}{R},$$

which is 63% of the final ϵ/R. The smaller the inductive time constant $\tau_L = L/R$, the more rapidly the current approaches ϵ/R.

We can find the time dependence of the induced voltage across the inductor in this circuit by using $V_L(t) = -L(dI/dt)$ and Equation 10.5.3:

$$V_L(t) = -L\frac{dI}{dt} = -\epsilon e^{-t/\tau_L}.$$

The magnitude of this function is plotted in Figure $10.5.2b$. The greatest value of $L(dI/dt)$ is ϵ; it occurs when dI/dt is greatest, which is immediately after S_1 is closed and S_2 is opened. In the approach to steady state, dI/dt decreases to zero. As a result, the voltage across the inductor also vanishes as $t \to \infty$.

The time constant τ_L also tells us how quickly the induced voltage decays. At $t = \tau_L$ the magnitude of the induced voltage is

$$|V_L(\tau_L)| = \epsilon e^{-1} = 0.37\epsilon = 0.37V(0).$$

The voltage across the inductor therefore drops to about 37% of its initial value after one time constant. The shorter the time constant τ_L, the more rapidly the voltage decreases.

After enough time has elapsed so that the current has essentially reached its final value, the positions of the switches in Figure $10.5.1a$ are reversed, giving us the circuit in part (c). At $t = 0$, the current in the circuit is $I(0) = \epsilon/R$. With Kirchhoff's loop rule, we obtain

$$IR + L\frac{dI}{dt} = 0.$$

The solution to this equation is similar to the solution of the equation for a discharging capacitor, with similar substitutions. The current at time **t** is then

$$I(t) = \frac{\epsilon}{R} e^{-t/\tau_L}.$$

The current starts at $I(0) = \epsilon/R$ and decreases with time as the energy stored in the inductor is depleted (Figure $10.5.3$).

The time dependence of the voltage across the inductor can be determined from $V_L = -L(dI/dt)$:

$$V_L(0) = \epsilon e^{-t/\tau_L}. \tag{10.5.5}$$

This voltage is initially $V_L(0) = \epsilon$, and it decays to zero like the current. The energy stored in the magnetic field of the inductor, $LI^2/2$, also decreases exponentially with time, as it is dissipated by Joule heating in the resistance of the circuit.

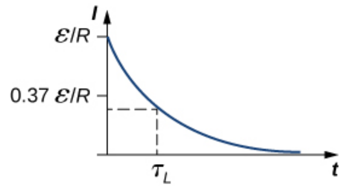


Figure 10.5.3: Time variation of electric current in the *RL* circuit of Figure 10.5.1c. The induced voltage across the coil also decays exponentially.

Example 10.5.1: An RL Circuit with a Source of emf

In the circuit of Figure 10.5.1a, let $\epsilon = 2.0\,V$, $R = 4.0\,\Omega$, and $L = 4.0\,H$. With S_1 closed and S_2 open (Figure 10.5.1b), (a) what is the time constant of the circuit? (b) What are the current in the circuit and the magnitude of the induced emf across the inductor at $t = 0$, at $t = 2.0\tau_L$, and as $t \to \infty$?

Strategy

The time constant for an inductor and resistor in a series circuit is calculated using Equation 10.5.4. The current through and voltage across the inductor are calculated by the scenarios detailed from Equation 10.5.3 and Equation 10.5.5.

Solution

1. The inductive time constant is

$$\tau_L = \frac{L}{R} = \frac{4.0\,H}{4.0\,\Omega} = 1.0\,s.$$

2. The current in the circuit of Figure 10.5.1b increases according to Equation 10.5.3:

$$I(t) = \frac{\epsilon}{R}(1 - e^{-t/\tau_L}).$$

At $t = 0$,

$$(1 - e^{-t/\tau_L}) = (1 - 1) = 0;\ so\, I(0) = 0.$$

At $t = 2.0\tau_L$ and $t \to \infty$, we have, respectively,

$$I(2.0\tau_L) = \frac{\epsilon}{R}(1 - e^{-2.0}) = (0.50\,A)(0.86) = 0.43\,A$$

and

$$I(\infty) = \frac{\epsilon}{R} = 0.50\,A.$$

From Equation 10.5.5, the magnitude of the induced emf decays as

$$|V_L(t)| = \epsilon e^{-t/\tau_L}.$$

At $t = 0$, $t = 2.0\tau_L$, and as $t \to \infty$, we obtain

$$|V_L(0)| = \epsilon = V,$$

$$|V_L(2.0\tau_L)| = (2.0\,V)e^{-2.0} = 0.27\,V$$

and

$$|V_L(\infty)| = 0.$$

Significance

If the time of the measurement were much larger than the time constant, we would not see the decay or growth of the voltage across the inductor or resistor. The circuit would quickly reach the asymptotic values for both of these (Figure $10.5.4$).

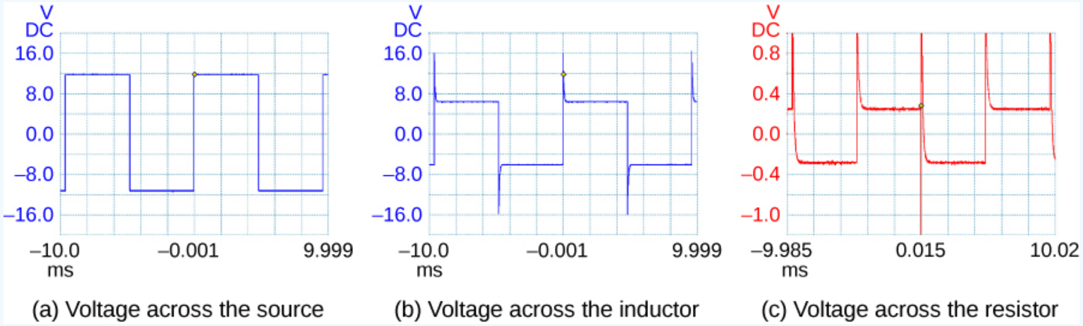


Figure 10.5.4: A generator in an **RL** circuit produces a square-pulse output in which the voltage oscillates between zero and some set value. These oscilloscope traces show (a) the voltage across the source; (b) the voltage across the inductor; (c) the voltage across the resistor.

Example $10.5.1$: An RL Circuit without a Source of emf

After the current in the **RL** circuit of Example $10.5.1$ has reached its final value, the positions of the switches are reversed so that the circuit becomes the one shown in Figure $10.5.1c$.

a. How long does it take the current to drop to half its initial value?
b. How long does it take before the energy stored in the inductor is reduced to 1.0% of its maximum value?

Strategy

The current in the inductor will now decrease as the resistor dissipates this energy. Therefore, the current falls as an exponential decay. We can also use that same relationship as a substitution for the energy in an inductor formula to find how the energy decreases at different time intervals.

Solution

1. With the switches reversed, the current decreases according to

$$I(t)=\frac{\epsilon}{R}e^{-t/\tau_L}=I(0)e^{-t/\tau_L}.$$

At a time **t** when the current is one-half its initial value, we have

$$I(t)=0.50I(0)\ so\ e^{-t/\tau_L}=0.50,$$

and

$$t=-[\ln(0.50)]\tau_L=0.69(1.0\ s)=0.69\ s$$

where we have used the inductive time constant found in Example $10.5.1$.

2. The energy stored in the inductor is given by

$$U_L(t)=\frac{1}{2}L[I(t)]^2=\frac{1}{2}L\left(\frac{\epsilon}{R}e^{-t/\tau_L}\right)^2=\frac{L\epsilon^2}{2R^2}e^{-2t/\tau_L}.$$

If the energy drops to 1.0% of its initial value at a time **t**, we have

$$U_L(t)=(0.010)U_L(0)\ or\ \frac{L\epsilon^2}{2R^2}e^{-2t/\tau_L}=(0.010)\frac{L\epsilon^2}{2R^2}.$$

Upon canceling terms and taking the natural logarithm of both sides, we obtain

$$-\frac{2t}{\tau_L}=\ln(0.010),$$

so

$$t = -\frac{1}{2}\tau_L \ln(0.010).$$

Since $\tau_L = 1.0\ s$, the time it takes for the energy stored in the inductor to decrease to 1.0% of its initial value is

$$t = -\frac{1}{2}(1.0\ s)\ln(0.010) = 2.3\ s.$$

Significance

This calculation only works if the circuit is at maximum current in situation (b) prior to this new situation. Otherwise, we start with a lower initial current, which will decay by the same relationship.

? Exercise 10.5.1

Verify that **RC** and **L/R** have the dimensions of time

? Exercise 10.5.2

a. If the current in the circuit of in Figure $10.5.1b$ increases to 90% of its final value after 5.0 s, what is the inductive time constant?
b. If $R = 20\ \Omega$, what is the value of the self-inductance?
c. If the $R = 20\ \Omega$ resistor is replaced with a $R = 100\ \Omega$ resistor, what is the time taken for the current to reach 90% of its final value?

Answer

a. 2.2 s; b. 43 H; c. 1.0 s

? Exercise 10.5.3

For the circuit of in Figure $10.5.1b$, show that when steady state is reached, the difference in the total energies produced by the battery and dissipated in the resistor is equal to the energy stored in the magnetic field of the coil

10.6: Oscillations in an LC Circuit

Learning Objectives

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

- Explain why charge or current oscillates between a capacitor and inductor, respectively, when wired in series
- Describe the relationship between the charge and current oscillating between a capacitor and inductor wired in series

It is worth noting that both capacitors and inductors store energy, in their electric and magnetic fields, respectively. A circuit containing both an inductor (**L**) and a capacitor (**C**) can oscillate without a source of emf by shifting the energy stored in the circuit between the electric and magnetic fields. Thus, the concepts we develop in this section are directly applicable to the exchange of energy between the electric and magnetic fields in electromagnetic waves, or light. We start with an idealized circuit of zero resistance that contains an inductor and a capacitor, an **LC circuit**.

An **LC** circuit is shown in Figure $10.6.1$. If the capacitor contains a charge q_0 before the switch is closed, then all the energy of the circuit is initially stored in the electric field of the capacitor (Figure $10.6.1a$). This energy is

$$U_C = \frac{1}{2}\frac{q_0^2}{C}.$$

When the switch is closed, the capacitor begins to discharge, producing a current in the circuit. The current, in turn, creates a magnetic field in the inductor. The net effect of this process is a transfer of energy from the capacitor, with its diminishing electric field, to the inductor, with its increasing magnetic field.

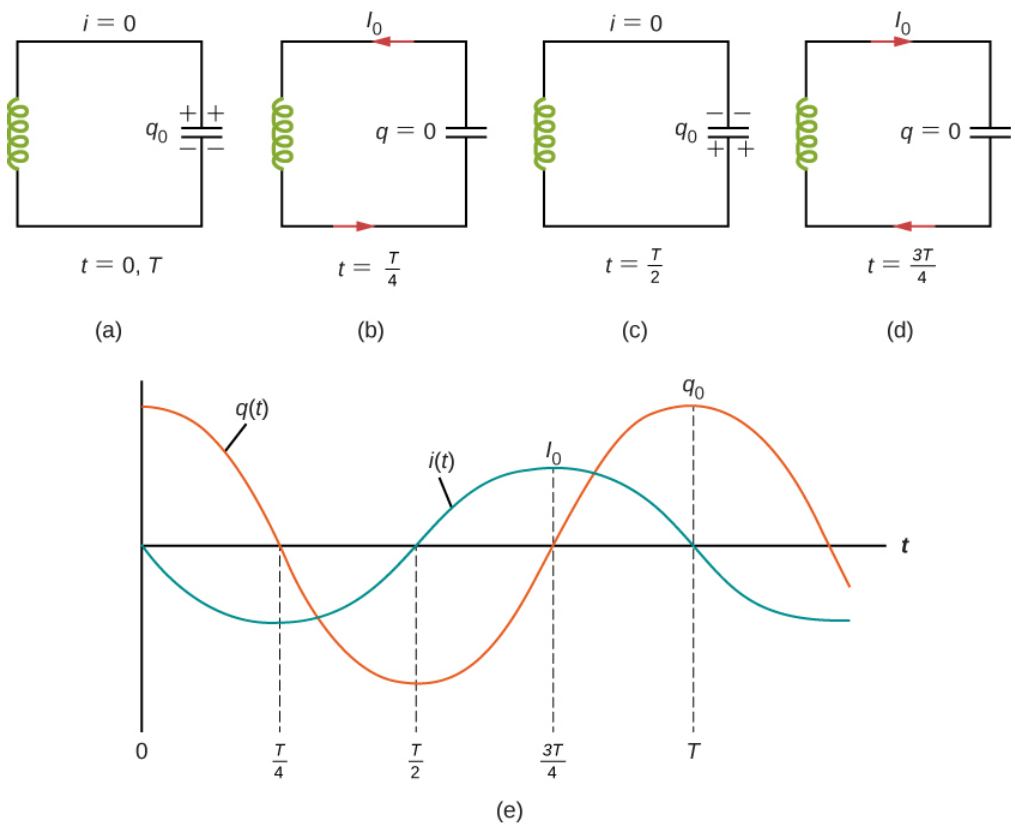


Figure 10.6.1: (a–d) The oscillation of charge storage with changing directions of current in an **LC** circuit. (e) The graphs show the distribution of charge and current between the capacitor and inductor.

In Figure $10.6.1b$, the capacitor is completely discharged and all the energy is stored in the magnetic field of the inductor. At this instant, the current is at its maximum value I_0 and the energy in the inductor is

$$U_L = \frac{1}{2}LI_0^2.$$

Since there is no resistance in the circuit, no energy is lost through Joule heating; thus, the maximum energy stored in the capacitor is equal to the maximum energy stored at a later time in the inductor:

$$\frac{1}{2}\frac{q_0^2}{C} = \frac{1}{2}LI_0^2.$$

At an arbitrary time when the capacitor charge is **q(t)** and the current is **i(t)**, the total energy **U** in the circuit is given by

$$\frac{q^2(t)}{2C} + \frac{Li^2}{2}.$$

Because there is no energy dissipation,

$$U = \frac{1}{2}\frac{q^2}{C} + \frac{1}{2}Li^2 = \frac{1}{2}\frac{q_0^2}{C} = \frac{1}{2}LI_0^2.$$

After reaching its maximum I_0, the current **i(t)** continues to transport charge between the capacitor plates, thereby recharging the capacitor. Since the inductor resists a change in current, current continues to flow, even though the capacitor is discharged. This continued current causes the capacitor to charge with opposite polarity. The electric field of the capacitor increases while the magnetic field of the inductor diminishes, and the overall effect is a transfer of energy from the inductor **back** to the capacitor. From the law of energy conservation, the maximum charge that the capacitor re-acquires is q_0. However, as Figure $10.6.1c$ shows, the capacitor plates are charged **opposite** to what they were initially.

When fully charged, the capacitor once again transfers its energy to the inductor until it is again completely discharged, as shown in Figure $10.6.1d$. Then, in the last part of this cyclic process, energy flows back to the capacitor, and the initial state of the circuit is restored.

We have followed the circuit through one complete cycle. Its electromagnetic oscillations are analogous to the mechanical oscillations of a mass at the end of a spring. In this latter case, energy is transferred back and forth between the mass, which has kinetic energy $mv^2/2$, and the spring, which has potential energy $kx^2/2$. With the absence of friction in the mass-spring system, the oscillations would continue indefinitely. Similarly, the oscillations of an **LC** circuit with no resistance would continue forever if undisturbed; however, this ideal zero-resistance **LC** circuit is not practical, and any **LC** circuit will have at least a small resistance, which will radiate and lose energy over time.

The frequency of the oscillations in a resistance-free **LC** circuit may be found by analogy with the mass-spring system. For the circuit, $i(t) = dq(t)/dt$, the total electromagnetic energy **U** is

$$U = \frac{1}{2}Li^2 + \frac{1}{2}\frac{q^2}{C}.$$

For the mass-spring system, $v(t) = dx(t)/dt$, the total mechanical energy **E** is

$$E = \frac{1}{2}mv^2 + \frac{1}{2}kx^2.$$

The equivalence of the two systems is clear. To go from the mechanical to the electromagnetic system, we simply replace **m** by **L**, **v** by **i**, **k** by 1/**C**, and **x** by **q**. Now **x(t)** is given by

$$x(t) = A\,cos(\omega t + \phi)$$

where $\omega = \sqrt{k/m}$. Hence, the charge on the capacitor in an **LC** circuit is given by

$$q(t) = q_0\,cos(\omega t + \phi) \tag{10.6.1}$$

where the angular frequency of the oscillations in the circuit is

$$\omega = \sqrt{\frac{1}{LC}}. \tag{10.6.2}$$

Finally, the current in the **LC** circuit is found by taking the time derivative of **q(t)**:

$$i(t) = \frac{dq(t)}{dt} = -\omega q_0\,sin(\omega t + \phi).$$

The time variations of **q** and **I** are shown in Figure 10.6.1e for $\phi = 0$.

✔ An LC Circuit

In an **LC** circuit, the self-inductance is 2.0×10^{-2} H and the capacitance is 8.0×10^{-6} F. At $t = 0$ all of the energy is stored in the capacitor, which has charge 1.2×10^{-5} C. (a) What is the angular frequency of the oscillations in the circuit? (b) What is the maximum current flowing through circuit? (c) How long does it take the capacitor to become completely discharged? (d) Find an equation that represents **q(t)**.

Strategy

The angular frequency of the **LC** circuit is given by Equation 10.6.2. To find the maximum current, the maximum energy in the capacitor is set equal to the maximum energy in the inductor. The time for the capacitor to become discharged if it is initially charged is a quarter of the period of the cycle, so if we calculate the period of the oscillation, we can find out what a quarter of that is to find this time. Lastly, knowing the initial charge and angular frequency, we can set up a cosine equation to find **q(t)**.

Solution

1. From Equation 10.6.2, the angular frequency of the oscillations is

$$\omega = \sqrt{\frac{1}{LC}} = \sqrt{\frac{1}{(2.0 \times 10^{-2}\ H)(8.0 \times 10^{-6}\ F)}} = 2.5 \times 10^{3}\ rad/s.$$

2. The current is at its maximum I_0 when all the energy is stored in the inductor. From the law of energy conservation,

$$\frac{1}{2}LI_0^2 = \frac{1}{2}\frac{q_0^2}{C},$$

so

$$I_0 = \sqrt{\frac{1}{LC}}q_0 = (2.5 \times 10^{3}\ rad/s)(1.2 \times 10^{-5} C) = 3.0 \times 10^{-2} A.$$

This result can also be found by an analogy to simple harmonic motion, where current and charge are the velocity and position of an oscillator.

3. The capacitor becomes completely discharged in one-fourth of a cycle, or during a time **T**/4, where **T** is the period of the oscillations. Since

$$T = \frac{2\pi}{\omega} = \frac{2\pi}{2.5 \times 10^{3}\ rad/s} = 2.5 \times 10^{-3} s,$$

the time taken for the capacitor to become fully discharged is $(2.5 \times 10^{-3} s)/4 = 6.3 \times 10^{-4} s$.

4. The capacitor is completely charged at $t = 0$, so $q(0) = q_0$. Using 10.6.1, we obtain

$$q(0) = q_0 = q_0 \cos\phi.$$

Thus, $\phi = 0$, and

$$q(t) = (1.2 \times 10^{-5} C)\cos(2.5 \times 10^{3} t).$$

Significance

The energy relationship set up in part (b) is not the only way we can equate energies. At most times, some energy is stored in the capacitor and some energy is stored in the inductor. We can put both terms on each side of the equation. By examining the circuit only when there is no charge on the capacitor or no current in the inductor, we simplify the energy equation.

? Exercise 10.6.1

The angular frequency of the oscillations in an **LC** circuit is 2.0×10^{3} rad/s. (a) If $L = 0.10\ H$, what is **C**? (b) Suppose that at $t = 0$ all the energy is stored in the inductor. What is the value of ϕ? (c) A second identical capacitor is connected in parallel with the original capacitor. What is the angular frequency of this circuit?

Solution

a. $2.5\ \mu F$; b. $\pi/2$ rad or $3\pi/2$ rad; c. 1.4×10^3 rad/s

10.7: RLC Series Circuits

Learning Objectives

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

- Determine the angular frequency of oscillation for a resistor, inductor, capacitor (**RLC**) series circuit
- Relate the **RLC** circuit to a damped spring oscillation

When the switch is closed in the **RLC** circuit of Figure $10.7.1a$, the capacitor begins to discharge and electromagnetic energy is dissipated by the resistor at a rate i^2R. With **U** given by Equation 14.4.2, we have

$$\frac{dU}{dt} = \frac{q}{C}\frac{dq}{dt} + Li\frac{di}{dt} = -i^2R$$

where **i** and **q** are time-dependent functions. This reduces to

$$L\frac{d^2q}{dt^2} + R\frac{dq}{dt} + \frac{1}{C}q = 0. \tag{10.7.1}$$

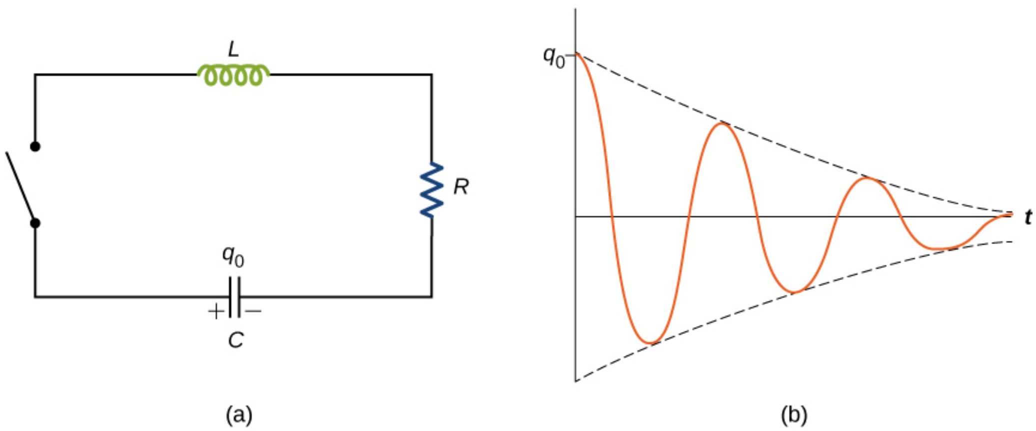


Figure 10.7.1: (a) An **RLC** circuit. Electromagnetic oscillations begin when the switch is closed. The capacitor is fully charged initially. (b) Damped oscillations of the capacitor charge are shown in this curve of charge versus time, or **q** versus **t**. The capacitor contains a charge q_0 before the switch is closed.

This equation is analogous to

$$m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = 0,$$

which is the equation of motion for a **damped mass-spring system** (you first encountered this equation in Oscillations). As we saw in that chapter, it can be shown that the solution to this differential equation takes three forms, depending on whether the angular frequency of the undamped spring is greater than, equal to, or less than **b**/2**m**. Therefore, the result can be underdamped ($\sqrt{k/m} > b/2m$), critically damped ($\sqrt{k/m} = b/2m$), or overdamped ($\sqrt{k/m} < b/2m$). By analogy, the solution **q(t)** to the **RLC** differential equation has the same feature. Here we look only at the case of under-damping. By replacing **m** by **L**, **b** by **R**, **k** by 1/**C**, and **x** by **q** in Equation 10.7.1, and assuming $\sqrt{1/LC} > R/2L$, we obtain

Note

$$q(t) = q_0 e^{-Rt/2L} cos(\omega' t + \phi) \tag{10.7.2}$$

where the angular frequency of the oscillations is given by

Note

$$\omega' = \sqrt{\frac{1}{LC} - \left(\frac{R}{2L}\right)^2} \tag{10.7.3}$$

This underdamped solution is shown in Figure $10.7.1b$. Notice that the amplitude of the oscillations decreases as energy is dissipated in the resistor. Equation 10.7.2 can be confirmed experimentally by measuring the voltage across the capacitor as a function of time. This voltage, multiplied by the capacitance of the capacitor, then gives **q(t)**.

Note

Try an interactive circuit construction kit that allows you to graph current and voltage as a function of time. You can add inductors and capacitors to work with any combination of **R**, **L**, and **C** circuits with both dc and ac sources.

Note

Try out a circuit-based java applet website that has many problems with both dc and ac sources that will help you practice circuit problems.

Exercise 10.7.1

In an **RLC** circuit, $L = 5.0\, mH$, $C = 6.0\, \mu F$, and $R = 200\, \Omega$. (a) Is the circuit underdamped, critically damped, or overdamped? (b) If the circuit starts oscillating with a charge of $3.0 \times 10^{-3} C$ on the capacitor, how much energy has been dissipated in the resistor by the time the oscillations cease?

Answer

a. overdamped; b. 0.75 J

10.8: Basic Radio Circuits

10.9: Application - RL Circuits with AC

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 $10.9.1$. This sine function assumes we start recording the voltage when it is $v = 0 \, V$ at a time of $t = 0 \, s$. 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.

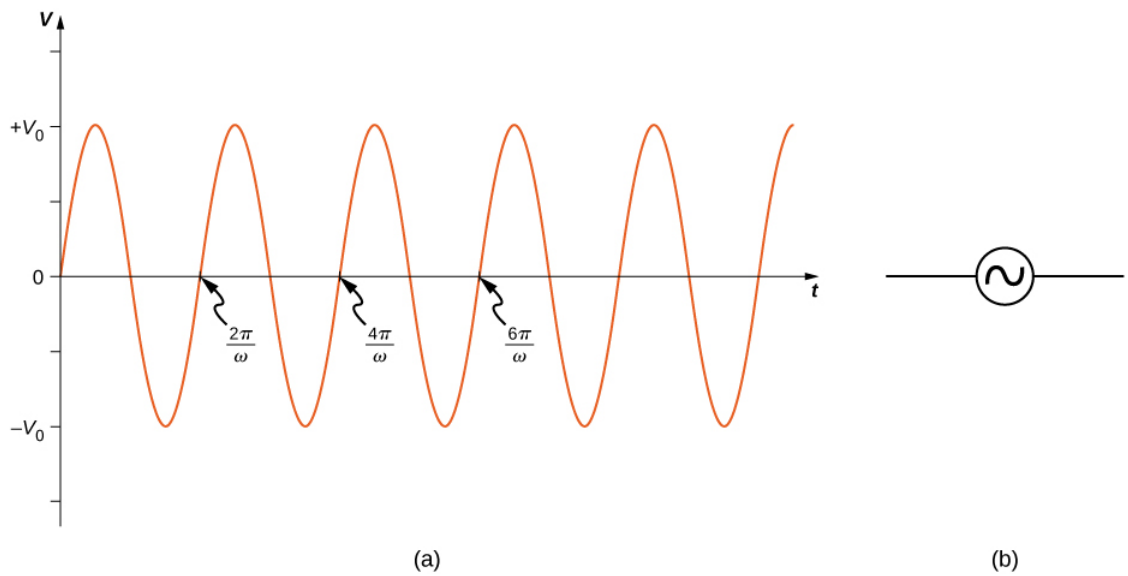


Figure 10.9.1: (a) The output $v(t) = V_0 \, sin \, \omega t$ of an ac generator. (b) Symbol used to represent an ac voltage source in a circuit diagram.

Resistor

First, consider a **resistor** connected across an ac voltage source. From Kirchhoff's loop rule, the instantaneous voltage across the resistor of Figure $10.9.2a$ is

$$v_R(t) = V_0 \, sin \, \omega t$$

and the instantaneous current through the resistor is

$$i_R(t) = \frac{v_R(t)}{R} = \frac{V_0}{R} \, sin \, \omega t = I_0 \, sin \, \omega t.$$

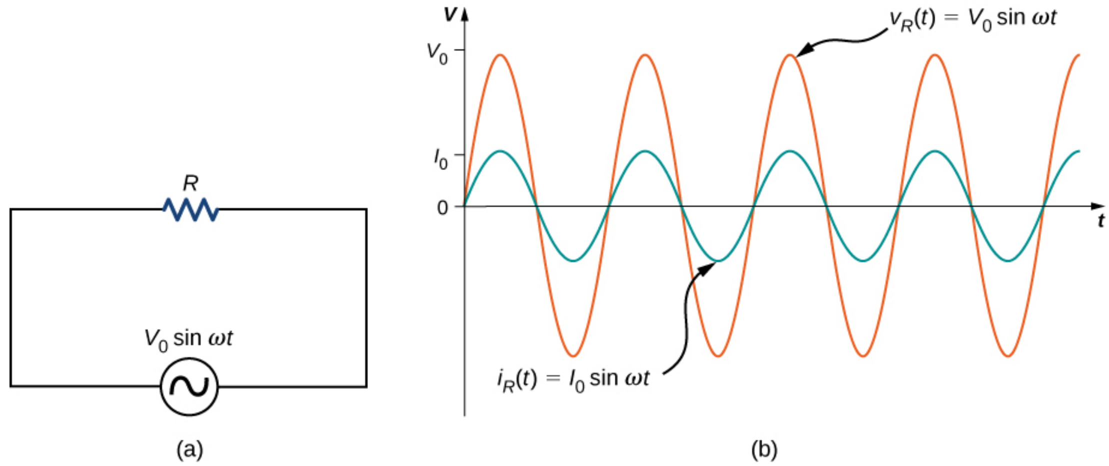


Figure 10.9.2: (a) A resistor connected across an ac voltage source. (b) The current $i_R(t)$ through the resistor and the voltage $v_R(t)$ across the resistor. The two quantities are in phase.

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 $10.9.2b$ 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 $10.9.3a$, 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.

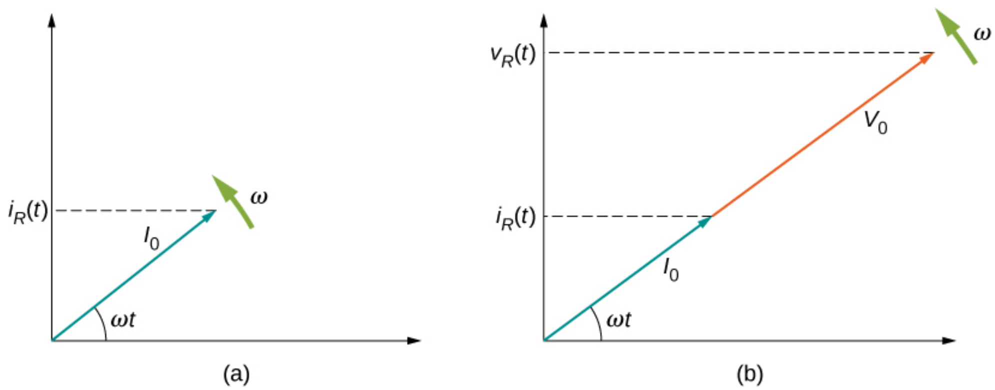


Figure 10.9.3: (a) The phasor diagram representing the current through the resistor of Figure 10.9.2 (b) The phasor diagram representing both $i_R(t)$ and $v_R(t)$.

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 $10.9.3b$ 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 $10.9.4a$ 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,

$$i_C(t) = \frac{dq(t)}{dt} = \omega C V_0 \cos \omega t = I_0 \cos \omega t,$$

where $I_0 = \omega C V_0$ 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:

$$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C. \tag{10.9.1}$$

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.

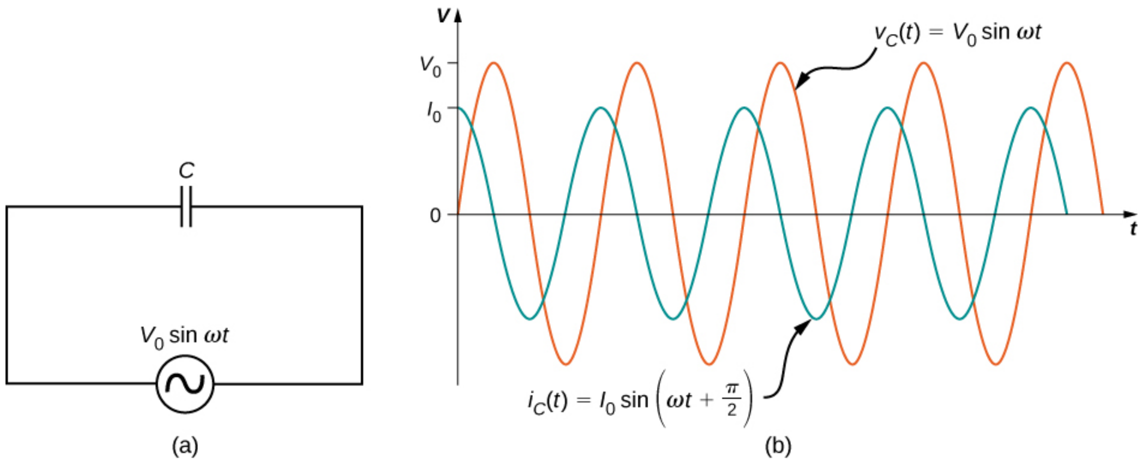


Figure 10.9.4: (a) A capacitor connected across an ac generator. (b) The current $i_C(t)$ through the capacitor and the voltage $v_C(t)$ across the capacitor. Notice that $i_C(t)$ leads $v_C(t)$ by $\pi/2$ rad.

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 $10.9.4b$. 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 $10.9.5$. Here, the relationship between $i_C(t)$ and $v_C(t)$ is represented by having their phasors rotate at the same angular frequency, with the current phasor leading by $\pi/2$ rad.

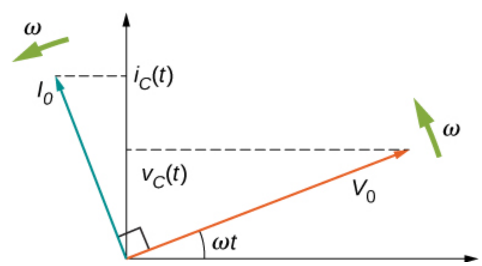


Figure 10.9.5: The current phasor leads the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

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 $10.9.6a$ is

$$v_L(t) = V_0\, sin\, \omega t. \tag{10.9.2}$$

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

$$\frac{di_L(t)}{dt} = \frac{V_0}{L} sin\, \omega t.$$

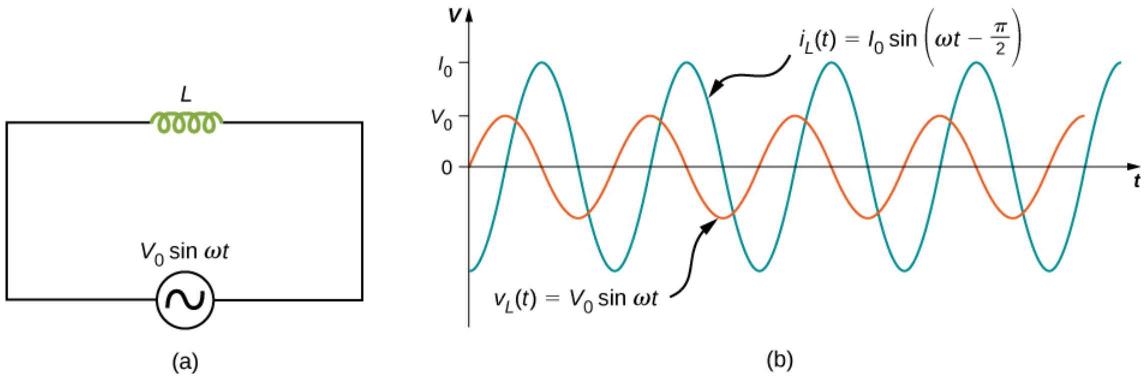


Figure 10.9.6: (a) An inductor connected across an ac generator. (b) The current $i_L(t)$ through the inductor and the voltage $v_L(t)$ across the inductor. Here $i_L(t)$ lags $v_L(t)$ by $\pi/2$ rad.

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

$$i_L(t) = -\frac{V_0}{\omega L} cos\, \omega t = \frac{V_0}{\omega L} sin\left(\omega t - \frac{\pi}{2}\right) = I_0\, sin\left(\omega t - \frac{\pi}{2}\right), \tag{10.9.3}$$

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

$$\frac{V_0}{I_0} = \omega L = X_L. \tag{10.9.4}$$

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 10.9.2 and Equation 10.9.3, 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 10.9.7.

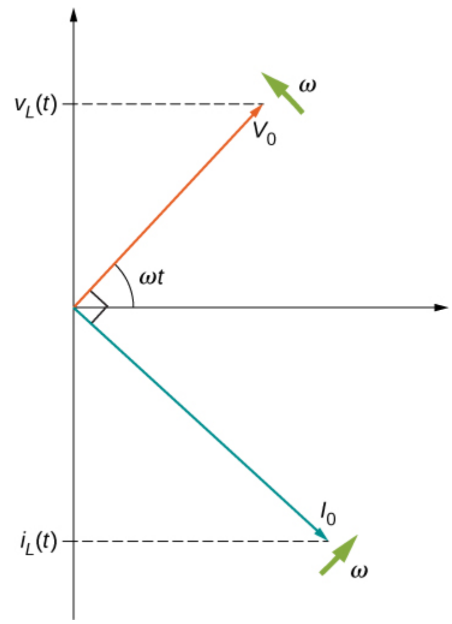


Figure 10.9.7: The current phasor lags the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

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 10.9.1: Simple AC CIrcuits

An ac generator produces an emf of amplitude 10 V at a frequency $f = 60\,Hz$. Determine the voltages across and the currents through the circuit elements when the generator is connected to (a) a $100\,\Omega$ resistor, (b) a $10\,\mu F$ 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

$$v(t) = V_0\,sin\,\omega t = (10\,V)sin\,120\pi t,$$

where $\omega = 2\pi f = 120\pi\,rad/s$ is the angular frequency. Since **v(t)** is also the voltage across each of the elements, we have

$$v(t)=v_R(t)=v_C(t)=v_L(t)=(10\,V)sin\,120\pi t.$$

a. When $R=100\,\Omega$, the amplitude of the current through the resistor is

$$I_0=V_0/R=10\,V/100\,\Omega=0.10\,A,$$

so

$$i_R(t)=(0.10\,A)sin\,120\pi t.$$

b. From Equation 10.9.1, the capacitive reactance is

$$X_C=\frac{1}{\omega C}=\frac{1}{(120\pi\,rad/s)(10\times10^{-6}F)}=265\,\Omega,$$

so the maximum value of the current is

$$I_0=\frac{V_0}{X_C}=\frac{10\,V}{265\,\Omega}=3.8\times10^{-2}\,A$$

and the instantaneous current is given by

$$i_C(t)=(3.8\times10^{-2}\,A)\,sin\,\left(120\pi t+\frac{\pi}{2}\right).$$

c. From Equation 10.9.4, the inductive reactance is

$$X_L=\omega L=(120\pi\,rad/s)(15\times10^{-3}H)=5.7\,\Omega.$$

The maximum current is therefore

$$I_0=\frac{10\,V}{5.7\,\Omega}=1.8\,A$$

and the instantaneous current is

$$i_L(t)=(1.8\,A)\,sin\,\left(120\pi t-\frac{\pi}{2}\right).$$

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 10.9.1

Repeat Example 10.9.1 for an ac source of amplitude 20 V and frequency 100 Hz.

Answer

a. $(20\,V)\,sin\,200\pi t$; $(0.20\,A)\,sin\,200\pi t$

b. $(20\,V)\,sin\,200\pi t$; $(0.13\,A)\,sin\,(200\pi t+\pi/2)$

c. $(20\,V)\,sin\,200\pi t$; $(2.1\,A)\,sin\,(200\pi t-\pi/2)$

10.10: Application - 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 10.10.1, 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

$$v(t) = V_0 \sin \omega t.$$

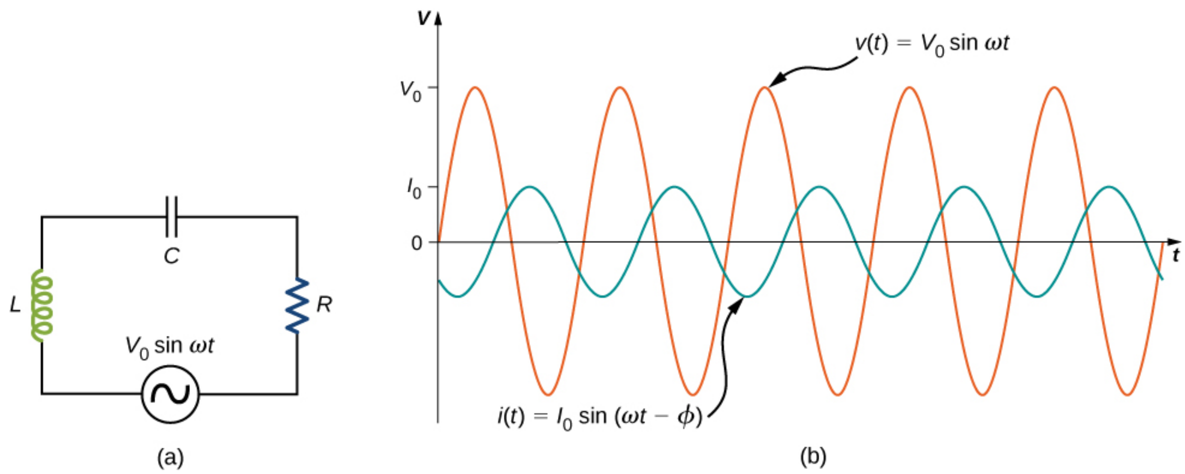


Figure 10.10.1: (a) An RLC series circuit. (b) A comparison of the generator output voltage and the current. The value of the phase difference ϕ depends on the values of R, C, and L.

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 10.10.2, 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 $V_L = I_0 X_L$.

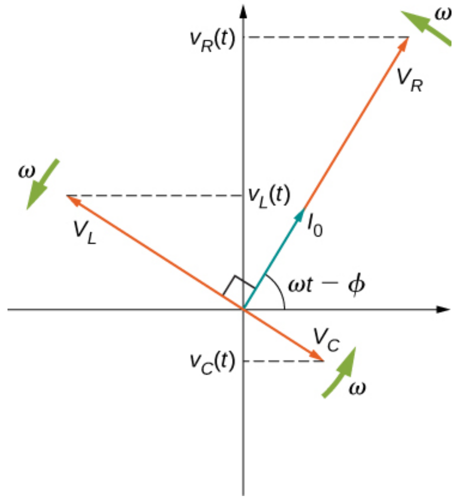


Figure 10.10.1.

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 $10.10.3$ The resultant phasor has an amplitudeV_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,

$$\phi=tan^{-1}\frac{V_L-V_C}{V_R}=tan^{-1}\frac{I_0X_L-I_0X_C}{I_0R},$$

and after cancellation of I_0, this becomes

$$\phi=tan^{-1}\frac{X_L-X_C}{R}. \tag{10.10.1}$$

Furthermore, from the Pythagorean theorem,

$$V_0=\sqrt{V_R^2+(V_L-V_C)^2}=\sqrt{(I_0R)^2+(I_0X_L-I_)X_C)^2}=I_0\sqrt{R^2+(X_L-X_C)^2}.$$

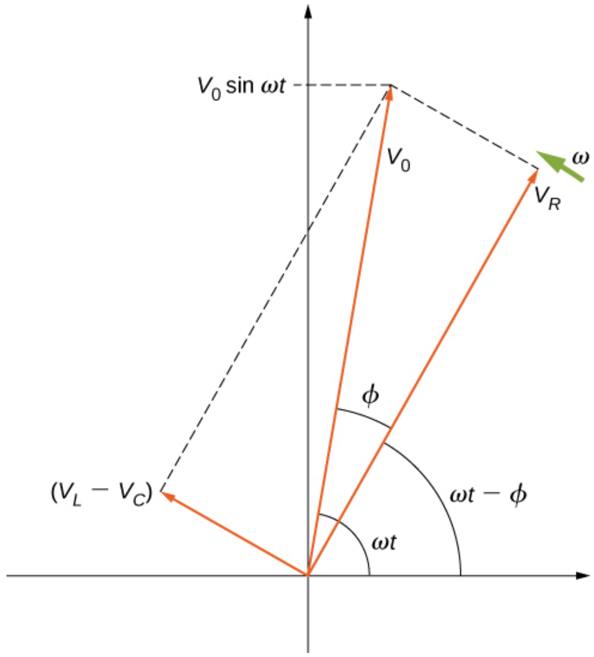


Figure 10.10.3: The resultant of the phasors for $v_L(t)$, $v_C(t)$, and $v_R(t)$ is equal to the phasor for $v_R(t) = V_0\, sin\, \omega t$. The **i(t)** phasor (not shown) is aligned with the $v_R(t)$ phasor.

The current amplitude is therefore the ac version of Ohm's law:

$$I_0 = \frac{V_0}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{V_0}{Z}, \tag{10.10.2}$$

where

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \tag{10.10.3}$$

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 10.10.4).

Figure 10.10.4: 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 10.10.5). 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.

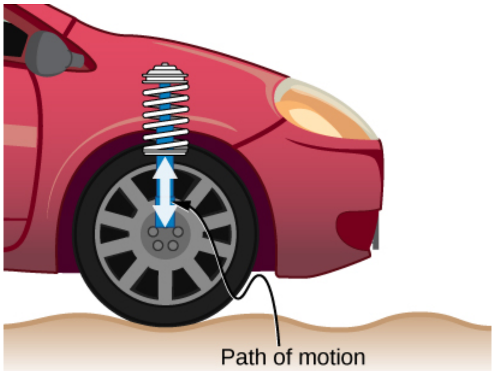


Figure 10.10.5: On a car, the shock absorber damps motion and dissipates energy. This is much like the resistance in an RLC circuit. The mass and spring determine the resonant frequency.

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 10.10.1: 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 $R=4.00\,\Omega$, $L=3.00\times10^{-3}H$, and $C=8.00\times10^{-4}F$, 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 10.10.2, 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

$$X_C=\frac{1}{\omega C}=\frac{1}{2\pi(200\,Hz)(8.00\times10^{-4}F)}=0.995\,\Omega.$$

2. From Equation 15.3.14, the inductive reactance is

$$X_L=\omega L=2\pi(200\,Hz)(3.00\times10^{-3}H)=3.77\,\Omega.$$

3. Substituting the values of **R**, X_C, and X_L into Equation 10.10.2, 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

$$I_0=\frac{V_0}{Z}=\frac{0.100\,V}{4.87\,\Omega}=2.05\times10^{-2}A.$$

5. From Equation 10.10.1, the phase difference between the current and the emf is

$$\phi=tan^{-1}\frac{X_L-X_C}{R}=tan^{-1}\frac{2.77\,\Omega}{4.00\,\Omega}=0.607\,rad.$$

Significance

The phase angle is positive because the reactance of the inductor is larger than the reactance of the capacitor.

Exercise 10.10.1

Find the voltages across the resistor, the capacitor, and the inductor in the circuit of Figure $10.10.1$ using $v(t)=V_0\,sin\,\omega t$ as the output of the ac generator.

Solution

$v_R=(V_0R/Z)\,sin\,(\omega t-\phi)$; $v_C=(V_0X_C/Z)\,sin\,(\omega t-\phi+\pi 2/)=-(V_0X_C/Z)\,cos\,(\omega t-\phi)$;
$v_L=(V_0X_L/Z)\,sin\,(\omega t-\phi-\pi/2)=(V_0X_L/Z)\,cos\,(\omega t-\phi)$

- **15.4: RLC Series Circuits with AC** by OpenStax is licensed CC BY 4.0. Original source: https://openstax.org/details/books/university-physics-volume-2.

10.11: Inductance (Summary)

Key Terms

henry (H)	unit of inductance, $1H = 1\Omega \cdot s$; it is also expressed as a volt second per ampere
inductance	property of a device that tells how effectively it induces an emf in another device
inductive time constant	denoted by τ, the characteristic time given by quantity L/R of a particular series RL circuit
inductor	part of an electrical circuit to provide self-inductance, which is symbolized by a coil of wire
LC circuit	circuit composed of an ac source, inductor, and capacitor
magnetic energy density	energy stored per volume in a magnetic field
mutual inductance	geometric quantity that expresses how effective two devices are at inducing emfs in one another
RLC circuit	circuit with an ac source, resistor, inductor, and capacitor all in series.
self-inductance	effect of the device inducing emf in itself

Key Equations

Mutual inductance by flux	$M = \frac{N_2\Phi_2}{1_I} = \frac{N_1\Phi_{12}}{I_2}$
Mutual inductance in circuits	$\varepsilon_1 = -M\frac{dI_2}{dt}$
Self-inductance in terms of magnetic flux	$N\Phi_m = LI$
Self-inductance in terms of emf	$\varepsilon = -L\frac{dI}{dt}$
Self-inductance of a solenoid	$L_{solenoid} = \frac{\mu_0 N^2 A}{l}$
Self-inductance of a toroid	$L_{toroid} = \frac{\mu_0 N^2 h}{2\pi} ln\frac{R_2}{R_1}.$
Energy stored in an inductor	$U = \frac{1}{2}LI^2$
Current as a function of time for a **RL** circuit	$I(t) = \frac{\varepsilon}{R}(1 - e^{-t/\tau_L})$
Time constant for a **RL** circuitτ	$\tau_L = L/R$
Charge oscillation in **LC** circuits	$q(t) = q_0 cos(\omega t + \phi)$
Angular frequency in **LC** circuits	$\omega = \sqrt{\frac{1}{LC}}$
Current oscillations in **LC** circuits	$i(t) = -\omega q_0 sin(\omega t + \phi)$
Charge as a function of time in **RLC** circuit	$q(t) = q_0 e^{-Rt/2L} cos(\omega' t + \phi)$
Angular frequency in **RLC** circuit	$\omega' = \sqrt{\frac{1}{LC} - (\frac{R}{2L})^2}$

Summary

14.2 Mutual Inductance

- Inductance is the property of a device that expresses how effectively it induces an emf in another device.
- Mutual inductance is the effect of two devices inducing emfs in each other.
- A change in current dI_1/dt in one circuit induces an emf (ε_2) in the second:

$$\varepsilon_2 = -M\frac{dI_1}{dt},$$

where **M** is defined to be the mutual inductance between the two circuits and the minus sign is due to Lenz's law.

- Symmetrically, a change in current dI_2/dt through the second circuit induces an emf (ε_1) in the first:

$$\varepsilon_1 = -M\frac{dI_2}{dt},$$

where **M** is the same mutual inductance as in the reverse process.

14.3 Self-Inductance and Inductors

- Current changes in a device induce an emf in the device itself, called self-inductance,

$$\varepsilon = -L\frac{dI}{dt},$$

where **L** is the self-inductance of the inductor and dI/dt is the rate of change of current through it. The minus sign indicates that emf opposes the change in current, as required by Lenz's law. The unit of self-inductance and inductance is the henry (H), where $1H = 1\Omega \cdot s$.

- The self-inductance of a solenoid is

$$L = \frac{\mu_0 N^2 A}{l},$$

where **N** is its number of turns in the solenoid, **A** is its cross-sectional area, **l** is its length, and $\mu_0 = 4\pi \times 10^{-7} T \cdot m/A$ is the permeability of free space.

- The self-inductance of a toroid is

$$L = \frac{\mu_0 N^2 h}{2\pi} ln\frac{R_2}{R_1},$$

where **N** is its number of turns in the toroid, R_1 and R_2 are the inner and outer radii of the toroid, **h** is the height of the toroid, and $\mu_0 = 4\pi \times 10^{-7} T \cdot m/A$ is the permeability of free space.

14.4 Energy in a Magnetic Field

- The energy stored in an inductor **U** is

$$U = \frac{1}{2}LI^2.$$

- The self-inductance per unit length of coaxial cable is

$$\frac{L}{l} = \frac{\mu_0}{2\pi} ln\frac{R_2}{R_1}.$$

14.5 RL Circuits

- When a series connection of a resistor and an inductor—an **RL** circuit—is connected to a voltage source, the time variation of the current is

$$I(t) = \frac{\varepsilon}{R}(1 - e^{-Rt/L}) = \frac{\varepsilon}{R}(1 - e^{-t/\tau_L}) \text{ (turning on)},$$

where the initial current is $I_0 = \varepsilon/R$..

- The characteristic time constant τ is $\tau_L = L/R$, where **L** is the inductance and **R** is the resistance.
- In the first time constant τ, the current rises from zero to $0.632I_0$, and to 0.632 of the remainder in every subsequent time interval τ.
- When the inductor is shorted through a resistor, current decreases as

$$I(t) = \frac{\varepsilon}{R}e^{-t/\tau_L} \text{ (turning off)}.$$

Current falls to $0.368I_0$ in the first time interval τ, and to 0.368 of the remainder toward zero in each subsequent time τ.

14.6 Oscillations in an LC Circuit

- The energy transferred in an oscillatory manner between the capacitor and inductor in an **LC** circuit occurs at an angular frequency $\omega = \sqrt{\frac{1}{LC}}$.
- The charge and current in the circuit are given by

$$q(t) = q_0 cos(\omega t + \phi)\ ,$$
$$i(t) = -\omega q_0 sin(\omega t + \phi)\ .$$

14.7 RLC Series Circuits

- The underdamped solution for the capacitor charge in an **RLC** circuit is

$$q(t) = q_0 e^{-Rt/2L} cos(\omega' t + \phi).$$

- The angular frequency given in the underdamped solution for the **RLC** circuit is

$$\omega' = \sqrt{\frac{1}{LC} - (\frac{R}{2L})^2}\ .$$

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).

10.12: Inductance (Exercise)

Conceptual Questions

14.2 Mutual Inductance

1. Show that $N\Phi_m/I$ and $\varepsilon/(dI/dt)$, which are both expressions for self-inductance, have the same units.

2. A 10-H inductor carries a current of 20 A. Describe how a 50-V emf can be induced across it.

3. The ignition circuit of an automobile is powered by a 12-V battery. How are we able to generate large voltages with this power source?

4. When the current through a large inductor is interrupted with a switch, an arc appears across the open terminals of the switch. Explain.

14.3 Self-Inductance and Inductors

5. Does self-inductance depend on the value of the magnetic flux? Does it depend on the current through the wire? Correlate your answers with the equation $N\Phi_m = LI$.

6. Would the self-inductance of a 1.0 m long, tightly wound solenoid differ from the self-inductance per meter of an infinite, but otherwise identical, solenoid?

7. Discuss how you might determine the self-inductance per unit length of a long, straight wire.

8. The self-inductance of a coil is zero if there is no current passing through the windings. True or false?

9. How does the self-inductance per unit length near the center of a solenoid (away from the ends) compare with its value near the end of the solenoid?

14.4 Energy in a Magnetic Field

10. Show that $LI^2/2$ has units of energy.

14.5 RL Circuits

11. Use Lenz's law to explain why the initial current in the RL circuit of Figure 14.12(b) is zero.

12. When the current in the **RL** circuit of Figure 14.12(b) reaches its final value ε/R, what is the voltage across the inductor? Across the resistor?

13. Does the time required for the current in an **RL** circuit to reach any fraction of its steady-state value depend on the emf of the battery?

14. An inductor is connected across the terminals of a battery. Does the current that eventually flows through the inductor depend on the internal resistance of the battery? Does the time required for the current to reach its final value depend on this resistance?

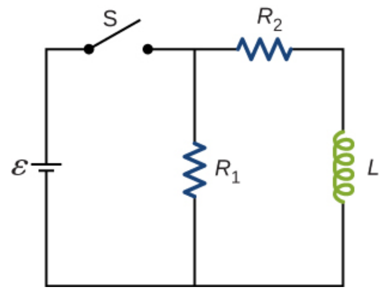
15. At what time is the voltage across the inductor of the **RL** circuit of Figure 14.12(b) a maximum?

16. In the simple **RL** circuit of Figure 14.12(b), can the emf induced across the inductor ever be greater than the emf of the battery used to produce the current?

17. If the emf of the battery of Figure 14.12(b) is reduced by a factor of 2, by how much does the steady-state energy stored in the magnetic field of the inductor change?

18. A steady current flows through a circuit with a large inductive time constant. When a switch in the circuit is opened, a large spark occurs across the terminals of the switch. Explain.

19. Describe how the currents through R_1 and R_2 shown below vary with time after switch S is closed.



20. Discuss possible practical applications of **RL** circuits.

14.6 Oscillations in an LC Circuit

21. Do Kirchhoff's rules apply to circuits that contain inductors and capacitors?

22. Can a circuit element have both capacitance and inductance?

23. In an LC circuit, what determines the frequency and the amplitude of the energy oscillations in either the inductor or capacitor?

14.7 RLC Series Circuits

24. When a wire is connected between the two ends of a solenoid, the resulting circuit can oscillate like an **RLC** circuit. Describe what causes the capacitance in this circuit.

25. Describe what effect the resistance of the connecting wires has on an oscillating **LC** circuit.

26. Suppose you wanted to design an **LC** circuit with a frequency of 0.01 Hz. What problems might you encounter?

27. A radio receiver uses an **RLC** circuit to pick out particular frequencies to listen to in your house or car without hearing other unwanted frequencies. How would someone design such a circuit?

Problems

14.2 Mutual Inductance

28. When the current in one coil changes at a rate of 5.6 A/s, an emf of $6.3 \times 10^{-3} V$. is induced in a second, nearby coil. What is the mutual inductance of the two coils?

29. An **emf** of $9.7 \times 10^{-3} V$ is induced in a coil while the current in a nearby coil is decreasing at a rate of 2.7 A/s. What is the mutual inductance of the two coils?

30. Two coils close to each other have a mutual inductance of 32 mH. If the current in one coil decays according to $I = I_0 e^{-\alpha t}$, where $I_0 = 5.0 A$ and $\alpha = 2.0 \times 10^3 s^{-1}$, what is the emf induced in the second coil immediately after the current starts to decay? At $t = 1.0 \times 10^{-3} s$?

31. A coil of 40 turns is wrapped around a long solenoid of cross-sectional area $7.5 \times 10^{-3} m^2$. The solenoid is 0.50 m long and has 500 turns.

(a) What is the mutual inductance of this system?

(b) The outer coil is replaced by a coil of 40 turns whose radius is three times that of the solenoid. What is the mutual inductance of this configuration?

32. A 600-turn solenoid is 0.55 m long and 4.2 cm in diameter. Inside the solenoid, a small (1.1cm×1.4cm), single-turn rectangular coil is fixed in place with its face perpendicular to the long axis of the solenoid. What is the mutual inductance of this system?

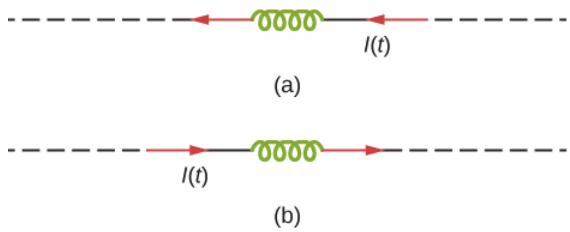
33. A toroidal coil has a mean radius of 16 cm and a cross-sectional area of $0.25 cm^2$; it is wound uniformly with 1000 turns. A second toroidal coil of 750 turns is wound uniformly over the first coil. Ignoring the variation of the magnetic field within a toroid, determine the mutual inductance of the two coils.

34. A solenoid of N_1 turns has length l_1 and radius R_1, and a second smaller solenoid of N_2 turns has length l_2 and radius R_2. The smaller solenoid is placed completely inside the larger solenoid so that their long axes coincide. What is the mutual inductance of the two solenoids?

14.3 Self-Inductance and Inductors

35. An **emf** of 0.40 V is induced across a coil when the current through it changes uniformly from 0.10 to 0.60 A in 0.30 s. What is the self-inductance of the coil?

36. The current shown in part (a) below is increasing, whereas that shown in part (b) is decreasing. In each case, determine which end of the inductor is at the higher potential.



37. What is the rate at which the current though a 0.30-H coil is changing if an emf of 0.12 V is induced across the coil?

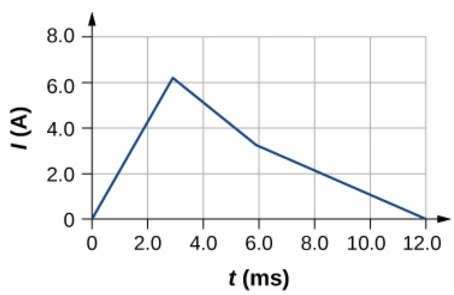
38. When a camera uses a flash, a fully charged capacitor discharges through an inductor. In what time must the 0.100-A current through a 2.00-mH inductor be switched on or off to induce a 500-V emf?

39. A coil with a self-inductance of 2.0 H carries a current that varies with time according to $I(t) = (2.0A)sin120\pi t$. Find an expression for the emf induced in the coil.

40. A solenoid 50 cm long is wound with 500 turns of wire. The cross-sectional area of the coil is $2.0cm^2$ What is the self-inductance of the solenoid?

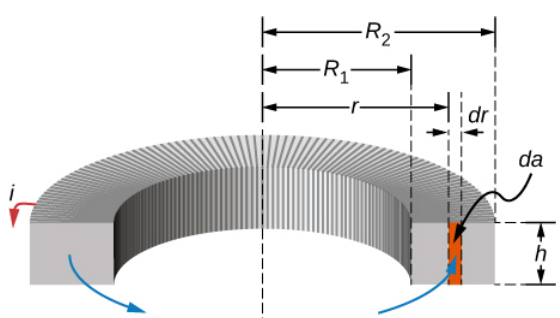
41. A coil with a self-inductance of 3.0 H carries a current that decreases at a uniform rate $dI/dt = -0.050A/s$. What is the **emf** induced in the coil? Describe the polarity of the induced **emf.**

42. The current $I(t)$ through a 5.0-mH inductor varies with time, as shown below. The resistance of the inductor is **5.0Ω**. Calculate the voltage across the inductor at $t = 2.0ms, t = 4.0ms$, and $t = 8.0ms$.



43. A long, cylindrical solenoid with 100 turns per centimeter has a radius of 1.5 cm. (a) Neglecting end effects, what is the self-inductance per unit length of the solenoid? (b) If the current through the solenoid changes at the rate 5.0 A/s, what is the emf induced per unit length?

44. Suppose that a rectangular toroid has 2000 windings and a self-inductance of 0.040 H. If h=0.10mh=0.10m, what is the ratio of its outer radius to its inner radius?



45. What is the self-inductance per meter of a coaxial cable whose inner radius is 0.50 mm and whose outer radius is 4.00 mm?

14.4 Energy in a Magnetic Field

46. At the instant a current of 0.20 A is flowing through a coil of wire, the energy stored in its magnetic field is $6.0 \times 10^{-3} J$. What is the self-inductance of the coil?

47. Suppose that a rectangular toroid has 2000 windings and a self-inductance of 0.040 H. If $h = 0.10m$, what is the current flowing through a rectangular toroid when the energy in its magnetic field is $2.0 \times 10^{-6} J$?

48. Solenoid **A** is tightly wound while solenoid **B** has windings that are evenly spaced with a gap equal to the diameter of the wire. The solenoids are otherwise identical. Determine the ratio of the energies stored per unit length of these solenoids when the same current flows through each.

49. A 10-H inductor carries a current of 20 A. How much ice at **0°C** could be melted by the energy stored in the magnetic field of the inductor? (Hint: Use the value $L_f = 334 J/g$ for ice.)

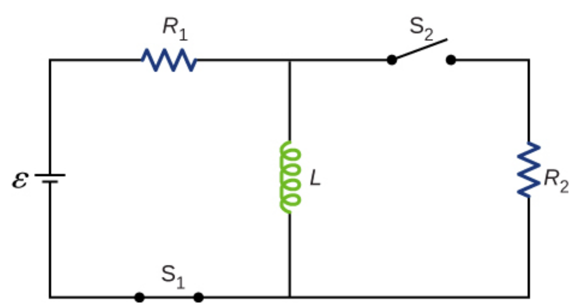
50. A coil with a self-inductance of 3.0 H and a resistance of **100Ω** carries a steady current of 2.0 A. (a) What is the energy stored in the magnetic field of the coil? (b) What is the energy per second dissipated in the resistance of the coil?

51. A current of 1.2 A is flowing in a coaxial cable whose outer radius is five times its inner radius. What is the magnetic field energy stored in a 3.0-m length of the cable?

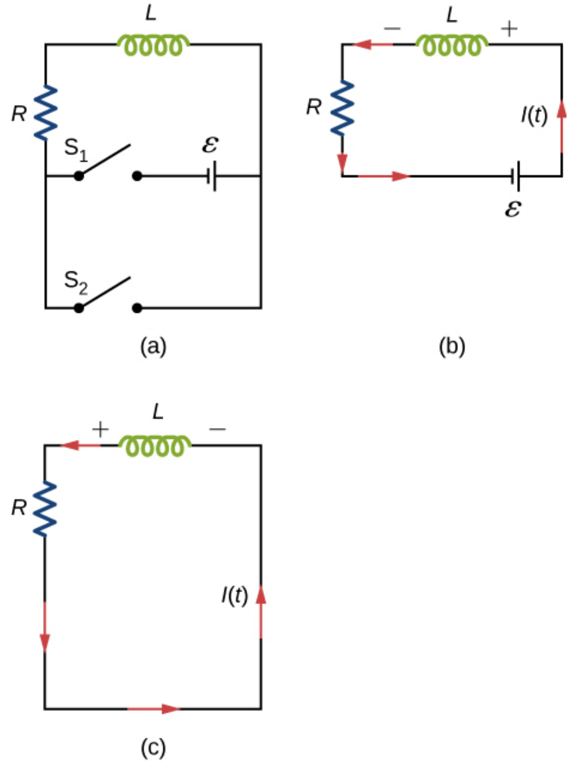
14.5 RL Circuits

52. In Figure 14.12, $\varepsilon = 12V$, $L = 20mH$, and $R = 5.0\Omega$. Determine (a) the time constant of the circuit, (b) the initial current through the resistor, (c) the final current through the resistor, (d) the current through the resistor when $t = 2\tau_L$, and (e) the voltages across the inductor and the resistor when $t = 2\tau_L$.

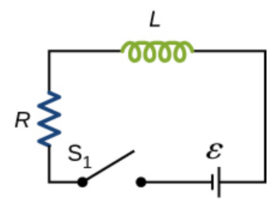
53. For the circuit shown below, $\varepsilon = 20V$, $L = 4.0mH$, and $R = 5.0\Omega$. After steady state is reached with S_1 closed and S_2 open, S_2 is closed and immediately thereafter (at $t = 0$) S_1 is opened. Determine (a) the current through L at $t = 0$, (b) the current through L at $t = 4.0 \times 10^{-4} s$, and (c) the voltages across L and R at $t = 4.0 \times 10^{-4} s$. $R_1 = R_2 = R$.



54. The current in the **RL** circuit shown here increases to **40%** of its steady-state value in 2.0 s. What is the time constant of the circuit?



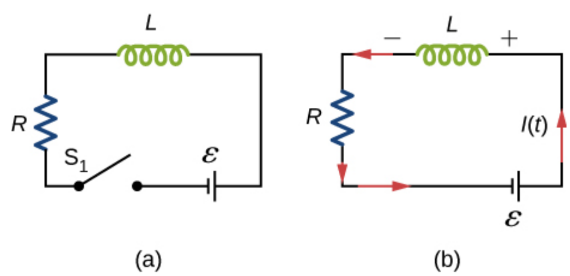
55. How long after switch S_1 is thrown does it take the current in the circuit shown to reach half its maximum value? Express your answer in terms of the time constant of the circuit.



56. Examine the circuit shown below in part (a). Determine dI/dt at the instant after the switch is thrown in the circuit of

(a), thereby producing the circuit of

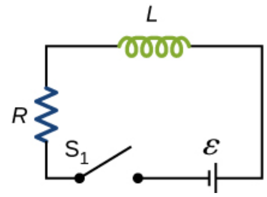
(b). Show that if I were to continue to increase at this initial rate, it would reach its maximum ε/R in one time constant.



57. The current in the **RL** circuit shown below reaches half its maximum value in 1.75 ms after the switch S_1 is thrown. Determine

(a) the time constant of the circuit and

(b) the resistance of the circuit if $L = 250 mH$.

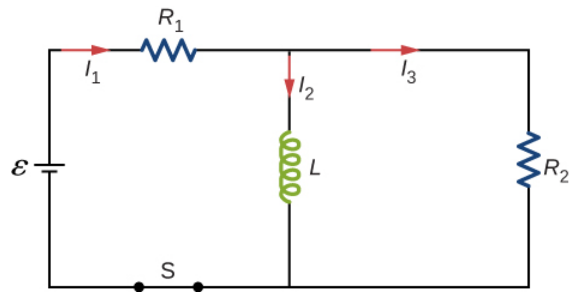


58. Consider the circuit shown below. Find I_1, I_2, and I_3 when

(a) the switch S is first closed,

(b) after the currents have reached steady-state values, and

(c) at the instant the switch is reopened (after being closed for a long time).



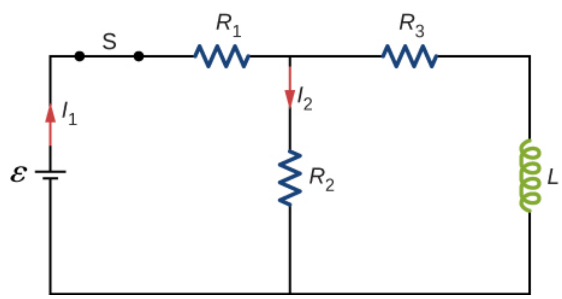
59. For the circuit shown below, $\varepsilon = 50V$, $R_1 = 10\Omega$, $R_2 = R_3 = 19.4\Omega$ and $L = 2.0mH$. Find the values of I_1 and I_2

(a) immediately after switch S is closed,

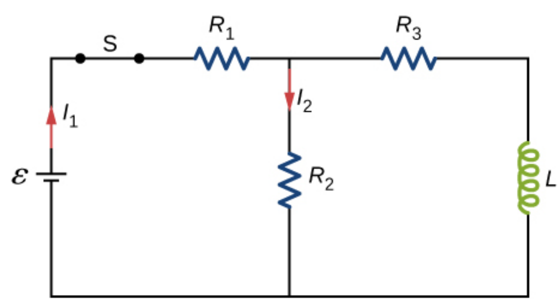
(b) a long time after S is closed,

(c) immediately after S is reopened, and

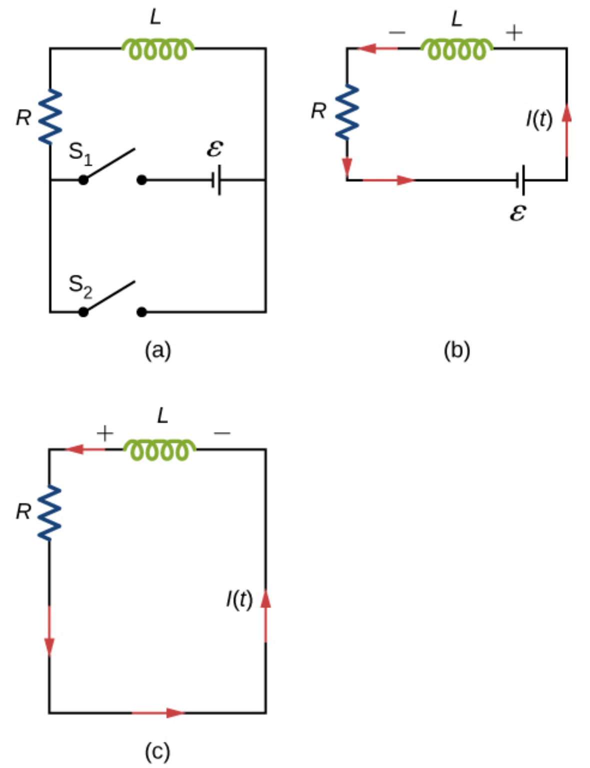
(d) a long time after S is reopened.



60. For the circuit shown below, find the current through the inductor $2.0 \times 10^{-5}s$ after the switch is reopened.



61. Show that for the circuit shown below, the initial energy stored in the inductor, $LI^2(0)/2$, is equal to the total energy eventually dissipated in the resistor, $\int_0^\infty I^2(t)Rdt$.



14.6 Oscillations in an LC Circuit

62. A 5000-pF capacitor is charged to 100 V and then quickly connected to an 80-mH inductor. Determine

(a) the maximum energy stored in the magnetic field of the inductor,

(b) the peak value of the current, and

(c) the frequency of oscillation of the circuit.

63. The self-inductance and capacitance of an **LC** circuit are 0.20 mH and 5.0 pF. What is the angular frequency at which the circuit oscillates?

64. What is the self-inductance of an **LC** circuit that oscillates at 60 Hz when the capacitance is $10\mu F$?

65. In an oscillating **LC** circuit, the maximum charge on the capacitor is $2.0 \times 10^{-6} C$ and the maximum current through the inductor is 8.0 mA.

(a) What is the period of the oscillations?

(b) How much time elapses between an instant when the capacitor is uncharged and the next instant when it is fully charged?

66. The self-inductance and capacitance of an oscillating **LC** circuit are $L = 20 mH$ and $C = 1.0 \mu F$, respectively.

(a) What is the frequency of the oscillations?

(b) If the maximum potential difference between the plates of the capacitor is 50 V, what is the maximum current in the circuit?

67. In an oscillating **LC** circuit, the maximum charge on the capacitor is q_m. Determine the charge on the capacitor and the current through the inductor when energy is shared equally between the electric and magnetic fields. Express your answer in terms of q_m, L, and C.

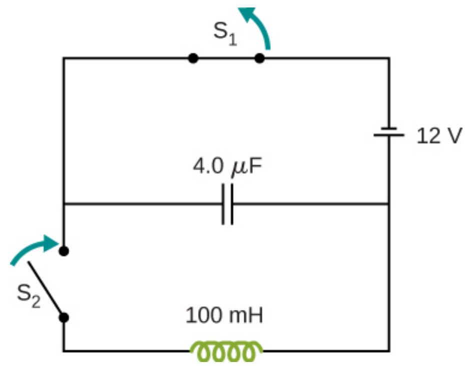
68. In the circuit shown below, S_1 is opened and S_2 is closed simultaneously. Determine

(a) the frequency of the resulting oscillations,

(b) the maximum charge on the capacitor,

(c) the maximum current through the inductor, and

(d) the electromagnetic energy of the oscillating circuit.



69. An **LC** circuit in an AM tuner (in a car stereo) uses a coil with an inductance of 2.5 mH and a variable capacitor. If the natural frequency of the circuit is to be adjustable over the range 540 to 1600 kHz (the AM broadcast band), what range of capacitance is required?

14.7 RLC Series Circuits

70. In an oscillating **RLC** circuit, $R = 5.0\Omega, L = 5.0mH$, and $C = 500\mu F$. What is the angular frequency of the oscillations?

71. In an oscillating **RLC** circuit with $L = 10mH, C = 1.5\mu F$, and $R = 2.0\Omega$, how much time elapses before the amplitude of the oscillations drops to half its initial value?

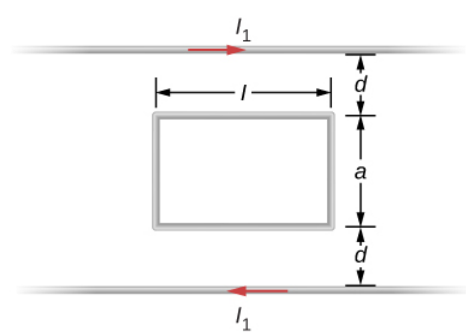
72. What resistance **R** must be connected in series with a 200-mH inductor of the resulting **RLC** oscillating circuit is to decay to **50%** of its initial value of charge in 50 cycles? To **0.10%** of its initial value in 50 cycles?

Additional Problems

73. Show that the self-inductance per unit length of an infinite, straight, thin wire is infinite.

74. Two long, parallel wires carry equal currents in opposite directions. The radius of each wire is a, and the distance between the centers of the wires is d. Show that if the magnetic flux within the wires themselves can be ignored, the self-inductance of a length l of such a pair of wires is $L = \frac{\mu_0 l}{\pi} ln \frac{d-a}{a}$. (Hint: Calculate the magnetic flux through a rectangle of length l between the wires and then use $L = N\Phi/I$.)

75. A small, rectangular single loop of wire with dimensions l, and a is placed, as shown below, in the plane of a much larger, rectangular single loop of wire. The two short sides of the larger loop are so far from the smaller loop that their magnetic fields over the smaller fields over the smaller loop can be ignored. What is the mutual inductance of the two loops?



76. Suppose that a cylindrical solenoid is wrapped around a core of iron whose magnetic susceptibility is x. Using Equation 14.9, show that the self-inductance of the solenoid is given by $L = \frac{(1+x)\mu_0 N^2 A}{l}$, where **I** is its length, **A** its cross-sectional area, and **N** its total number of turns.

77. The solenoid of the preceding problem is wrapped around an iron core whose magnetic susceptibility is 4.0×10^3.

(a) If a current of 2.0 A flows through the solenoid, what is the magnetic field in the iron core?

(b) What is the effective surface current formed by the aligned atomic current loops in the iron core?

(c) What is the self-inductance of the filled solenoid?

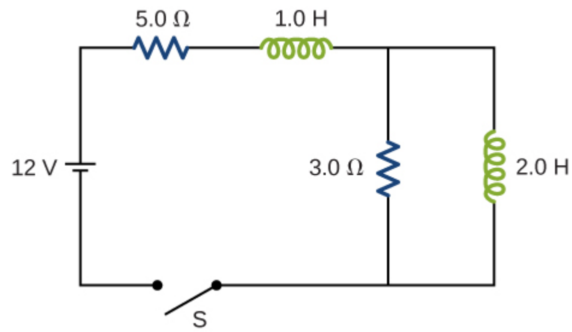
78. A rectangular toroid with inner radius $R_1 = 7.0 cm$, outer radius $R_2 = 9.0 cm$, height $h = 3.0$, and $N = 3000$ turns is filled with an iron core of magnetic susceptibility 5.2×10^3.

(a) What is the self-inductance of the toroid?

(b) If the current through the toroid is 2.0 A, what is the magnetic field at the center of the core?

(c) For this same 2.0-A current, what is the effective surface current formed by the aligned atomic current loops in the iron core?

79. The switch S of the circuit shown below is closed at $t = 0$. Determine (a) the initial current through the battery and (b) the steady-state current through the battery.



80. In an oscillating **RLC** circuit, $R = 7.0 \Omega, L = 10 mH$, and $C = 3.0 \mu F$. Initially, the capacitor has a charge of $8.0 \mu C$ and the current is zero. Calculate the charge on the capacitor

(a) five cycles later and

(b) 50 cycles later.

81. A 25.0-H inductor has 100 A of current turned off in 1.00 ms.

(a) What voltage is induced to oppose this?

(b) What is unreasonable about this result?

(c) Which assumption or premise is responsible?

Challenge Problems

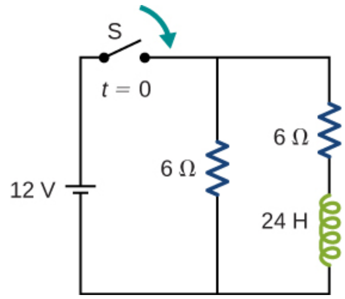
82. A coaxial cable has an inner conductor of radius a, and outer thin cylindrical shell of radius b. A current **I** flows in the inner conductor and returns in the outer conductor. The self-inductance of the structure will depend on how the current in the inner cylinder tends to be distributed. Investigate the following two extreme cases.

(a) Let current in the inner conductor be distributed only on the surface and find the self-inductance.

(b) Let current in the inner cylinder be distributed uniformly over its cross-section and find the self-inductance. Compare with your results in (a).

83. In a damped oscillating circuit the energy is dissipated in the resistor. The **Q**-factor is a measure of the persistence of the oscillator against the dissipative loss. (a) Prove that for a lightly damped circuit the energy, **U**, in the circuit decreases according to the following equation. $\frac{dU}{dt} = -2\beta U$, where $\beta = \frac{R}{2L}$. (b) Using the definition of the **Q**-factor as energy divided by the loss over the next cycle, prove that **Q**-factor of a lightly damped oscillator as defined in this problem is $Q \equiv \frac{U_{begin}}{\Delta U_{one\ cycle}} = \frac{1}{2\pi R}\sqrt{\frac{L}{C}}$. (Hint: For (b), to obtain **Q**, divide **E** at the beginning of one cycle by the change ΔE over the next cycle.)

84. The switch in the circuit shown below is closed at $t = 0 s$. Find currents through (a) R_1, (b) R_2, and (c) the battery as function of time.



85. A square loop of side 2 cm is placed 1 cm from a long wire carrying a current that varies with time at a constant rate of 3 A/s as shown below.

(a) Use Ampère’s law and find the magnetic field as a function of time from the current in the wire.

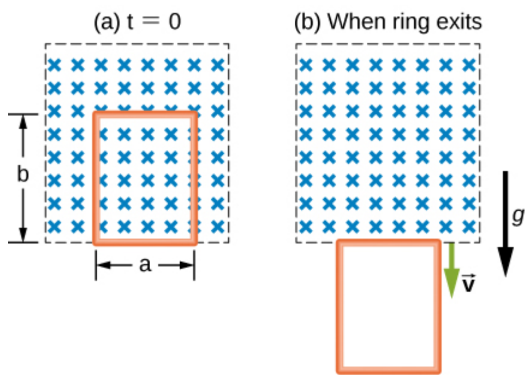
(b) Determine the magnetic flux through the loop.

(c) If the loop has a resistance of **3Ω**, how much induced current flows in the loop?

86. A rectangular copper ring, of mass 100 g and resistance **0.2Ω**, is in a region of uniform magnetic field that is perpendicular to the area enclosed by the ring and horizontal to Earth’s surface. The ring is let go from rest when it is at the edge of the nonzero magnetic field region (see below).

(a) Find its speed when the ring just exits the region of uniform magnetic field.

(b) If it was let go at t=0t=0, what is the time when it exits the region of magnetic field for the following values: $a=25cm, b=50cm, B=3T,$ and $g=9.8m/s^2$? Assume the magnetic field of the induced current is negligible compared to 3 T.



10.13: Inductance (Answers)

Check Your Understanding

14.1. $4.77 \times 10^{-2} V$

14.2. a. decreasing;

b. increasing; Since the current flows in the opposite direction of the diagram, in order to get a positive emf on the left-hand side of diagram (a), we need to decrease the current to the left, which creates a reinforced emf where the positive end is on the left-hand side. To get a positive emf on the right-hand side of diagram (b), we need to increase the current to the left, which creates a reinforced emf where the positive end is on the right-hand side.

14.3. 40 A/s

14.4. a. $4.5 \times 10^{-5} H$;

b. $4.5 \times 10^{-3} V$

14.5. a. $2.4 \times 10^{-7} Wb$;

b. $6.4 \times 10^{-5} m^2$

14.6. 0.50 J

14.8. a. 2.2 s;

b. 43 H;

c. 1.0 s

14.10. a. $2.5 \mu F$;

b. $\pi/2 rad$ or $3\pi/2 rad$;

c. 1.4×10^3 rad/s

14.11. a. overdamped;

b. 0.75 J

Conceptual Questions

1. $\dfrac{Wb}{A} = \dfrac{T \cdot m^2}{A} = \dfrac{V \cdot s}{A} = \dfrac{V}{A/s}$

3. The induced current from the 12-V battery goes through an inductor, generating a large voltage.

5. Self-inductance is proportional to the magnetic flux and inversely proportional to the current. However, since the magnetic flux depends on the current I, these effects cancel out. This means that the self-inductance does not depend on the current. If the emf is induced across an element, it does depend on how the current changes with time.

7. Consider the ends of a wire a part of an **RL** circuit and determine the self-inductance from this circuit.

9. The magnetic field will flare out at the end of the solenoid so there is less flux through the last turn than through the middle of the solenoid.

11. As current flows through the inductor, there is a back current by Lenz's law that is created to keep the net current at zero amps, the initial current.

13. no

15. At $t = 0$, or when the switch is first thrown.

17. 1/4

19. Initially, $I_{R1} = \dfrac{\varepsilon}{R_1}$ and $I_{R2} = 0$, and after a long time has passed, $I_{R1} = \dfrac{\varepsilon}{R_1}$ and $I_{R2} = \dfrac{\varepsilon}{R_2}$.

21. yes

23. The amplitude of energy oscillations depend on the initial energy of the system. The frequency in a **LC** circuit depends on the values of inductance and capacitance.

25. This creates an **RLC** circuit that dissipates energy, causing oscillations to decrease in amplitude slowly or quickly depending on the value of resistance.

27. You would have to pick out a resistance that is small enough so that only one station at a time is picked up, but big enough so that the tuner doesn't have to be set at exactly the correct frequency. The inductance or capacitance would have to be varied to tune into the station however practically speaking, variable capacitors are a lot easier to build in a circuit.

Problems

29. $M = 3.6 \times 10^{-3} H$

31. a. $3.8 \times 10^{-4} H$;

b. $3.8 \times 10^{-4} H$

33. $M_{21} = 2.3 \times 10^{-5} H$

35. 0.24 H

37. 0.4 A/s

39. $\varepsilon = 480\pi sin(120\pi t - \pi/2)V$

41. 0.15 V. This is the same polarity as the emf driving the current.

43. a. 0.089 H/m;

b. 0.44 V/m

45. $\frac{L}{l} = 4.16 \times 10^{-7} H/m$

47. 0.01 A

49. 6.0 g

51.$U_m = 7.0 \times 10^{-7} J$

53. a. 4.0 A;

b. 2.4 A;

c. on R: $V = 12V$; on L: $V = 7.9V$

55. 0.69τ

57. a. 2.52 ms;

b. 99.2Ω

59. a. $I_1 = I_2 = 1.7A$;

b. $I_1 = 2.73A, I_2 = 1.36A$;

c. $I_1 = 0, I_2 = 0.54$;

d. $I_1 = I_2 = 0$

61. proof

63. $\omega = 3.2 \times 10^7 rad/s$

65. a. $7.9 \times 10^{-4} s$;

b. $4.0 \times 10^{-4} s$

67. $q = \frac{qm}{\sqrt{2}}, I = \frac{q_m}{\sqrt{2LC}}$

69. $C=\frac{1}{4\pi^2 f^2 L}$

$f_1 = 540Hz$; $C_1 = 3.5\times10^{-11}F$

$f_2 =$; $1600Hz$; $C_2 = 4.0\times10^{-12}F$

71. 6.9 ms

Additional Problems

73. proof

Outside, $B=\frac{\mu_0 I}{2\pi r}$ Inside, $B=\frac{\mu_0 Ir}{2\pi a^2}$

$U=\frac{\mu_0 I^2 l}{4\pi}(\frac{1}{4}+ln\frac{R}{a})$

So, $\frac{2U}{I^2}=\frac{\mu_0 l}{P2\pi}(\frac{1}{4}+ln\frac{R}{a})$ and $L=\infty$

75. $M=\frac{\mu_0 l}{\pi}ln\frac{d+a}{d}$

77. a. 100 T;

b. 2 A;

c. 0.50 H

79. a. 0 A;

b. 2.4 A

81. $a. 2.50\times10^6 V;$

(b) The voltage is so extremely high that arcing would occur and the current would not be reduced so rapidly.

(c) It is not reasonable to shut off such a large current in such a large inductor in such an extremely short time.

Challenge Problems

83. proof

85. a. $\frac{dB}{dt}=6\times10^{-6}T/s;$

b. $\Phi=\frac{\mu_0 aI}{2\pi}ln(\frac{a+b}{b});$

c. 4.0 nA

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).

CHAPTER OVERVIEW

11: Electromagnetic Waves

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.

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.

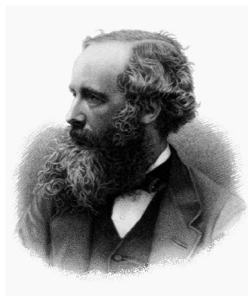


Figure 11.2.1: James Clerk Maxwell, a 19th-century physicist, developed a theory that explained the relationship between electricity and magnetism and correctly predicted that visible light is caused by electromagnetic waves. (credit: G. J. Stodart)

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 , 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

$$c=\frac{1}{\sqrt{\mu_0\epsilon_0}}. \tag{11.2.1}$$

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

$$c=\frac{1}{\sqrt{\left(8.85\times10^{-12}\frac{C^2}{N\cdot m^2}\right)\left(4\pi\times10^{-7}\frac{T\cdot m}{A}\right)}}=3.00\times10^{8}m/s, \tag{11.2.2}$$

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.

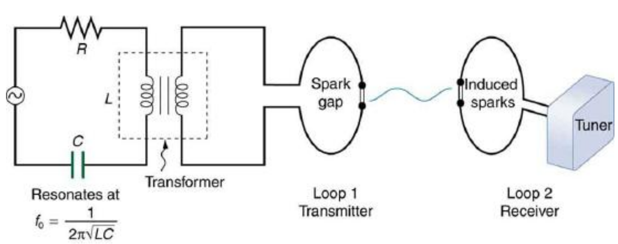


Figure 11.2.2: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves. An RLC circuit connected to the first loop caused sparks across a gap in the wire loop and generated electromagnetic waves. Sparks across a gap in the second loop located across the laboratory gave evidence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

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}}, \tag{11.2.3}$$

  where mu_0 is the permeability of free space and ϵ_0 is the permitivity 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 x 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

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.

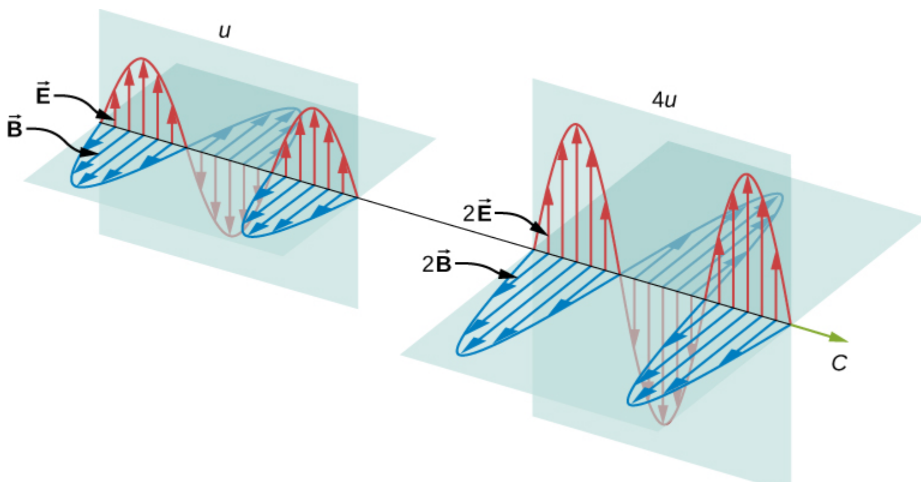


Figure $11.3.1$: Energy carried by a wave depends on its amplitude. With electromagnetic waves, doubling the **E** fields and **B** fields quadruples the energy density **u** and the energy flux **uc**.

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 **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 $E = cB = \frac{1}{\sqrt{\epsilon_0 \mu_0}} B$ 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

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

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.

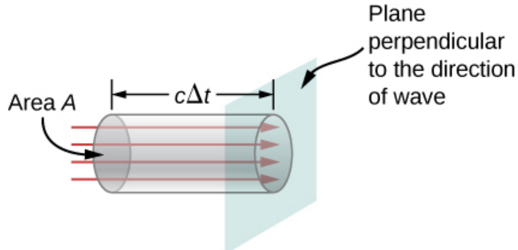


Figure 11.3.2: The energy $uAc\Delta t$ contained in the electric and magnetic fields of the electromagnetic wave in the volume $Ac\Delta t$ passes through the area A in time Δt.

The energy passing through area A in time Δt is

$$u \times volume = uAc\Delta t.$$

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.

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

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):

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

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

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

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

$$I = \frac{cB_0^2}{2\mu_0} \tag{11.3.5}$$

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

$$I = \frac{E_0 B_0}{2\mu_0}. \tag{11.3.6}$$

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:

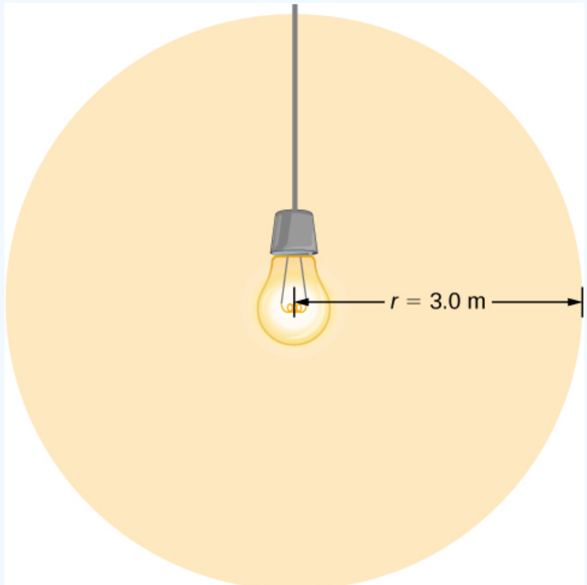
$$B_0 = \frac{E_0}{c} = 2.9\times 10^{-9}\, T.$$

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

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

$$E_0 = \sqrt{2\frac{P}{4\pi r^2 c\epsilon_0}} = \sqrt{2\frac{5.00\ W}{4\pi(3.0\ m)^2(3.00\times10^8\ m/s)(8.85\times10^{-12} C^2/N\cdot m^2)}} = 5.77\ N/C,$$

$$B_0 = E_0/c = 1.92\times10^{-8}\ T\ .$$

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?

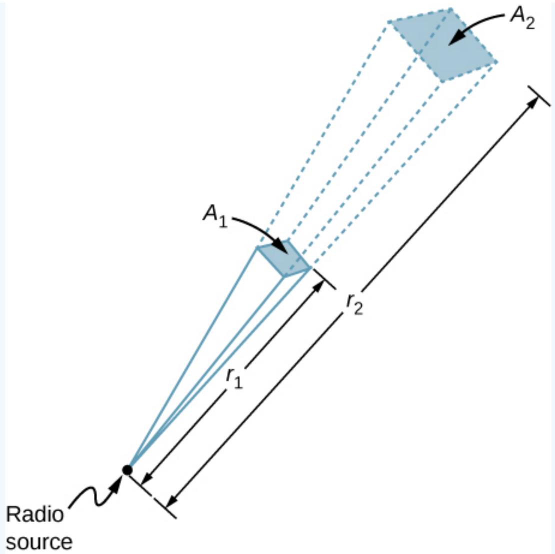


Figure 11.3.3: In three dimensions, a signal spreads over a solid angle as it travels outward from its source.

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 **P** by a factor of (90 kW/60 kW) and change the area by the same factor to keep $I = \frac{P}{A} = \frac{c\epsilon_0 E_0^2}{2}$ the same. Then use the proportion of area **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

$$\frac{r_2^2}{r_1^2} = \frac{A_2}{A_1} = \frac{90\,W}{60\,W},$$

$$r_2 = \sqrt{\frac{90}{60}}(100\,km)$$

$$= 122\,km.$$

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.

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 medicineSecurity	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,

$$c = f\lambda. \tag{11.4.1}$$

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.

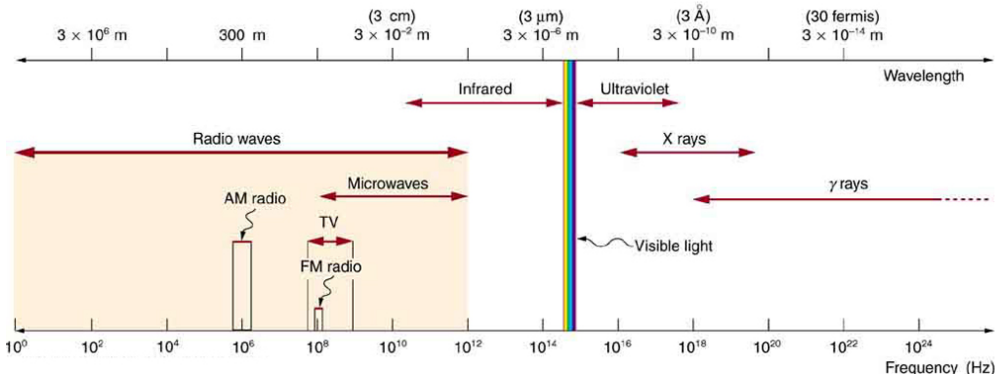


Figure 11.4.1: The electromagnetic spectrum, showing the major categories of electromagnetic waves. The range of frequencies and wavelengths is remarkable. The dividing line between some categories is distinct, whereas other categories overlap.

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).

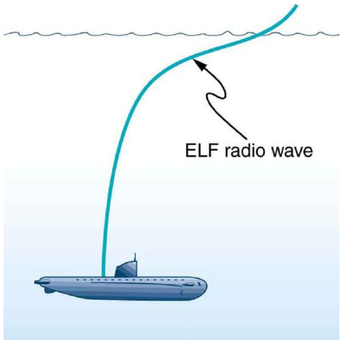


Figure 11.4.3: Very long wavelength radio waves are needed to reach this submarine, requiring extremely low frequency signals (ELF). Shorter wavelengths do not penetrate to any significant depth.

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.

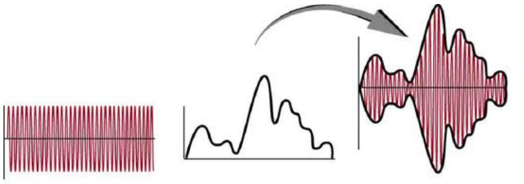


Figure 11.4.4: Amplitude modulation for AM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The amplitude of the carrier is modulated by the audio signal without changing its basic frequency.

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.

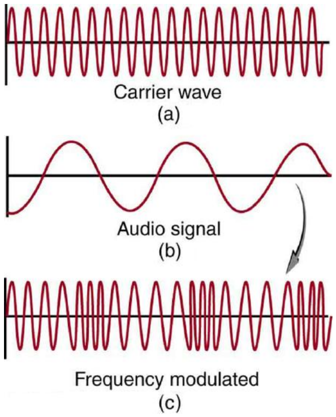


Figure 11.4.5: Frequency modulation for FM radio. (a) A carrier wave at the station's basic frequency. (b) An audio signal at much lower audible frequencies. (c) The frequency of the carrier is modulated by the audio signal without changing its amplitude.

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 $c = 3.00 \times 10^8 m/s$ 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

$$\lambda = \frac{c}{f}. \tag{11.4.2}$$

a. For the $f = 1530 kHz$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 m/s}{1530 \times 10^3 cycles/s} \tag{11.4.3}$$

$$= 196 m. \tag{11.4.4}$$

b. For the $f = 105.1MHz$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 m/s}{105.1 \times 10^6 cycles/s} \tag{11.4.5}$$

$$= 2.85m. \tag{11.4.6}$$

c. For the $f = 1.90GHz$ AM radio signal, then,

$$\lambda = \frac{3.00 \times 10^8 m/s}{1.90 \times 10^9 cycles/s} \tag{11.4.7}$$

$$= 0.158m. \tag{11.4.8}$$

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 $11.4.6$).

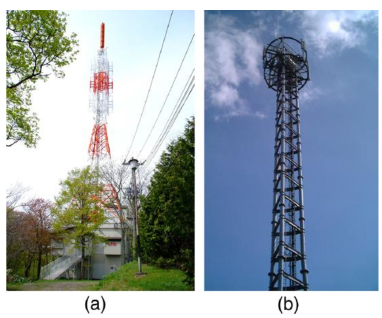


Figure 11.4.6: (a) A large tower is used to broadcast TV signals. The actual antennas are small structures on top of the tower -- they are placed at great heights to have a clear line of sight over a large broadcast area. (credit: Ozizo, Wikimedia Commons) (b) The NTT Dokomo mobile phone tower at Tokorozawa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

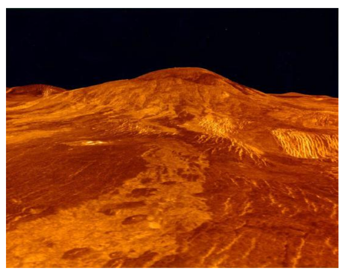


Figure $11.4.7$: An image of Sif Mons with lava flows on Venus, based on Magellan synthetic aperture radar data combined with radar altimetry to produce a three-dimensional map of the surface. The Venusian atmosphere is opaque to visible light, but not to the microwaves that were used to create this 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 $40°C$ 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.

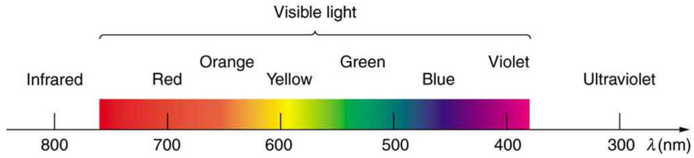


Figure 11.4.8: A small part of the electromagnetic spectrum that includes its visible components. The divisions between infrared, visible, and ultraviolet are not perfectly distinct, nor are those between the seven rainbow colors.

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 $0.30 \mu m$ thick. Calculate the energy absorbed, assuming the corneal tissue has the same properties as water; it is initially at $34^{\circ} C$. Assume the evaporated tissue leaves at a temperature of $100^{\circ} C$.

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 $100^{\circ} C$, we can apply concepts of thermal energy. We know that

$$Q = mc\Delta T, \tag{11.4.9}$$

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

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

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

$$Q_v = mL_v = m\,(2256 kJ/kg)\,.$$

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,

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

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

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

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

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

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 (hv) can be written as:

$$\mathrm{CFCl_3 + hv \rightarrow CFCl_2 + Cl.} \tag{11.4.10}$$

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

$$\mathrm{Cl + O_3 \rightarrow ClO + O_2} \tag{11.4.11}$$

and

$$\mathrm{ClO + O_3 \rightarrow Cl + 2\,O_2.} \tag{11.4.12}$$

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 $11.4.9$).

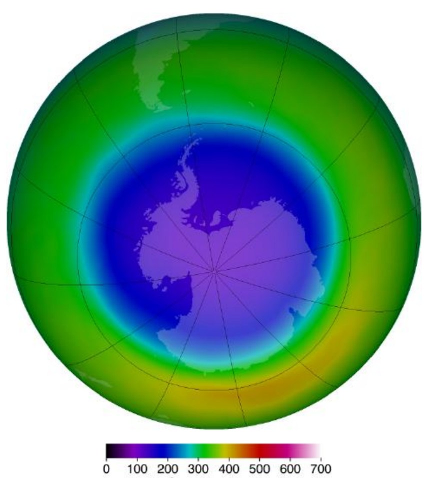


Figure $11.4.9$: This map of ozone concentration over Antarctica in October 2011 shows severe depletion suspected to be caused by CFCs. Less dramatic but more general depletion has been observed over northern latitudes, suggesting the effect is global. With less ozone, more ultraviolet radiation from 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.

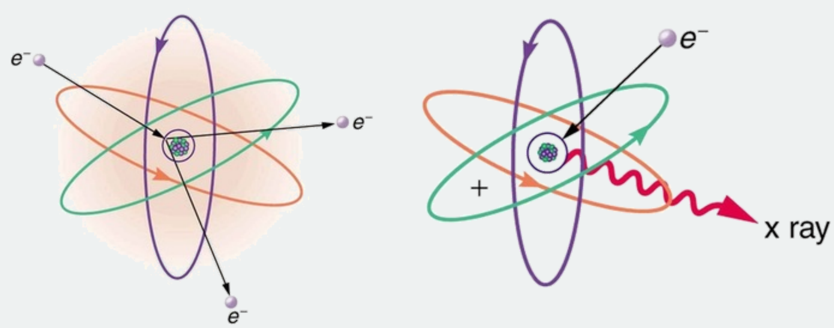


Figure 11.4.10: Artist’s conception of an electron ionizing an atom followed by the recapture of an electron and emission of an X-ray. An energetic electron strikes an atom and knocks an electron out of one of the orbits closest to the nucleus. Later, the atom captures another electron, and the energy 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.

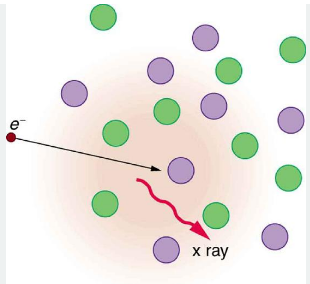


Figure 11.4.11: Artist’s conception of an electron being slowed by collisions in a material and emitting X-ray radiation. This energetic electron makes numerous collisions with electrons and atoms in a material it penetrates. An accelerated charge radiates EM waves, a second method by which X-rays are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

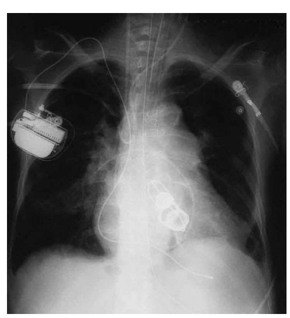


Figure 11.4.12: This shadow X-ray image shows many interesting features, such as artificial heart valves, a pacemaker, and the wires used to close the sternum. (credit: P. P. Urone)

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.

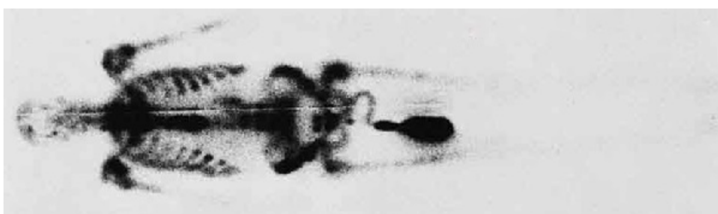


Figure 11.4.13: This is an image of the γ rays emitted by nuclei in a compound that is concentrated in the bones and eliminated through the kidneys. Bone cancer is evidenced by nonuniform concentration in similar structures. For example, some ribs are darker than others. (credit: P. P. Urone)

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.

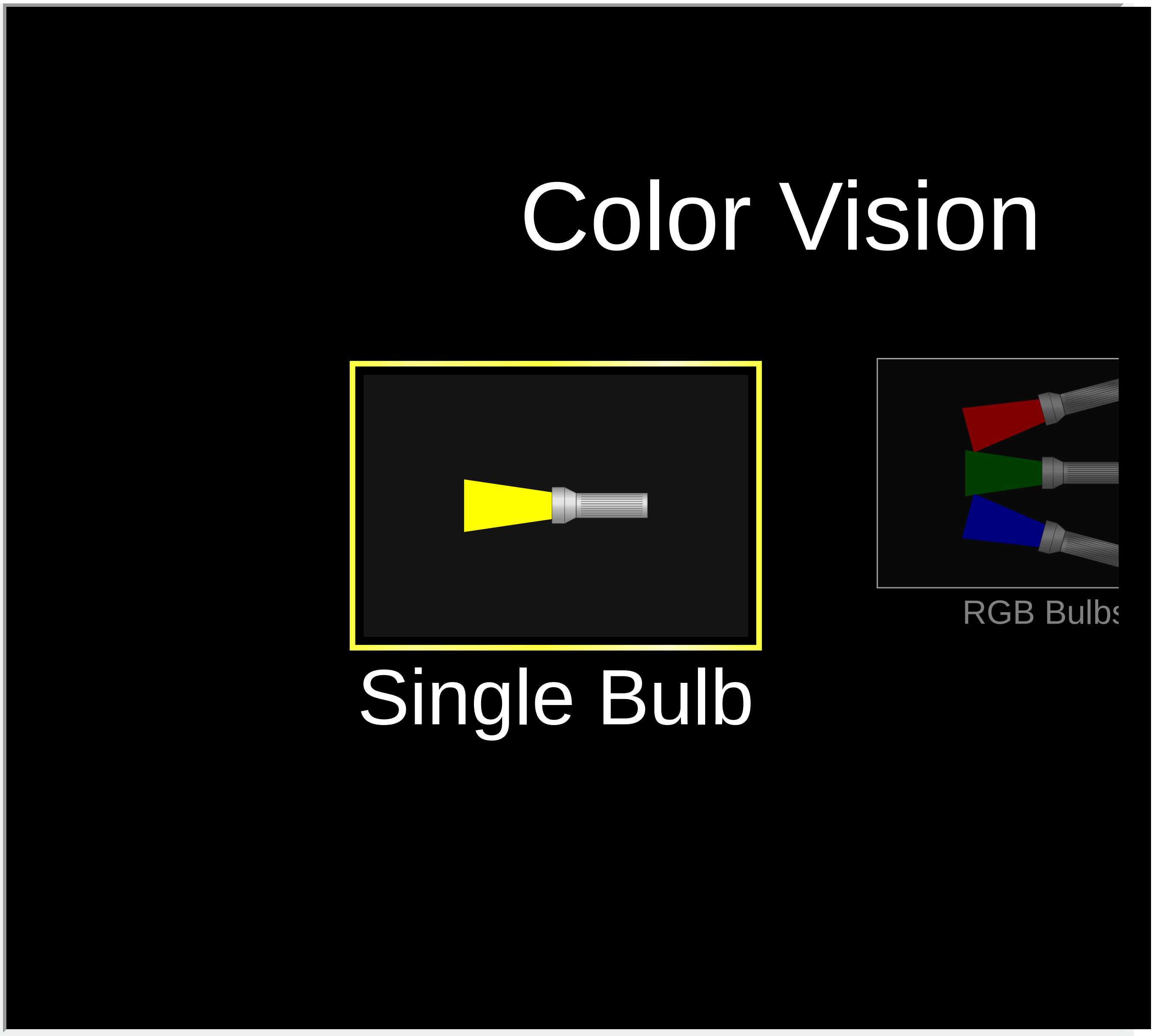
Infrared telescopes need to have their detectors cooled by liquid nitrogen to be able to gather useful signals. Since infrared radiation is predominantly from thermal agitation, if the detectors were not cooled, the vibrations of the molecules in the antenna would be stronger than the signal being collected.

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.



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 $0.74 \mu m$to $300 \mu m$

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

---

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 $11.5.2$.

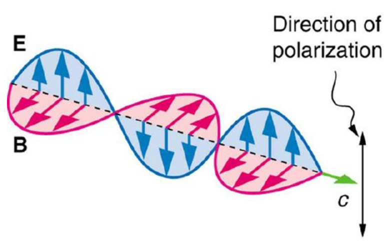


Figure 11.5.2: An EM wave, such as light, is a transverse wave. The electric and magnetic fields are perpendicular to the direction of propagation.

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.

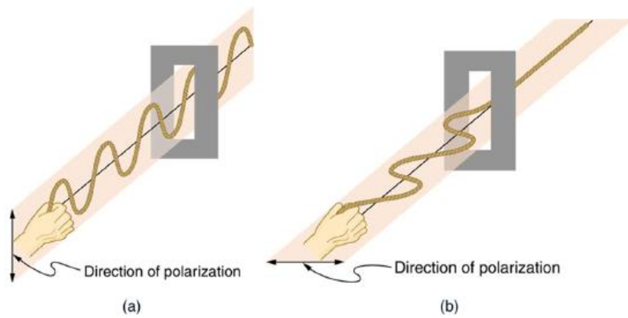


Figure 11.5.3: The transverse oscillations in one rope are in a vertical plane, and those in the other rope are in a horizontal plane. The first is said to be vertically polarized, and the other is said to be horizontally polarized. Vertical slits pass vertically polarized waves and block horizontally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

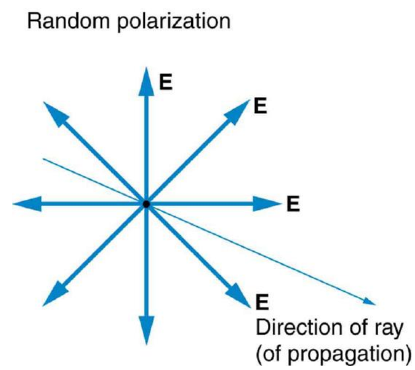


Figure 11.5.4: The slender arrow represents a ray of unpolarized light. The bold arrows represent the direction of polarization of the individual waves composing the ray. Since the light is unpolarized, the arrows point in all directions.

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).

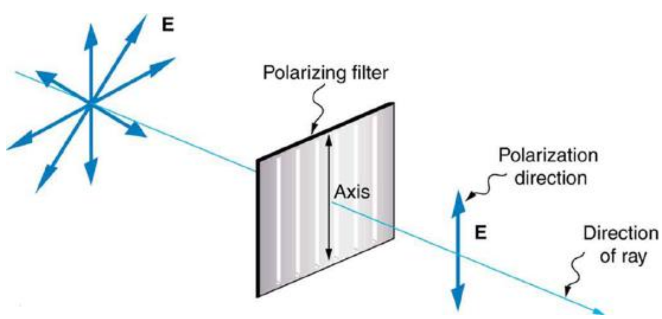


Figure 11.5.5: A polarizing filter has a polarization axis that acts as a slit passing through electric fields parallel to its direction. The direction of polarization of an EM wave is defined to be the direction of its electric field.

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

$$I = I_0 \cos\theta^2, \tag{11.5.1}$$

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.

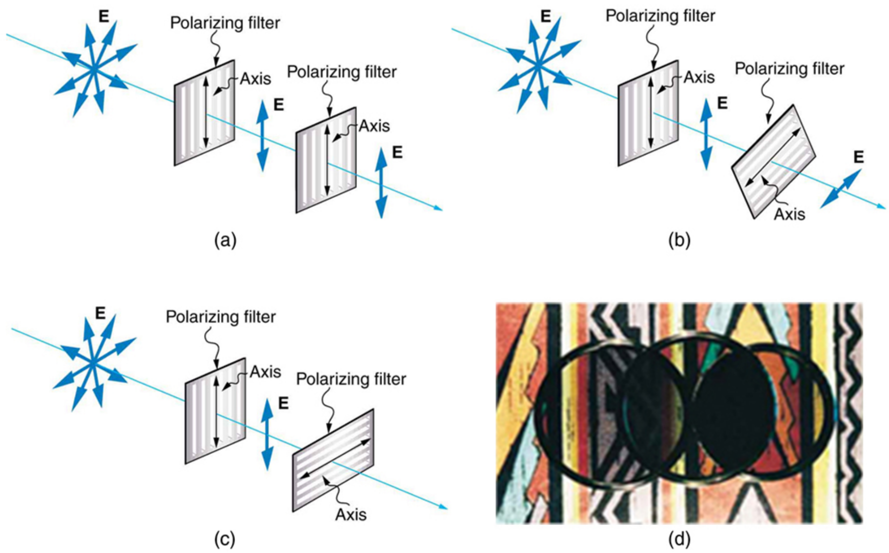


Figure 11.5.6: The effect of rotating two polarizing filters, where the first polarizes the light. (a) All of the polarized light is passed by the second polarizing filter, because its axis is parallel to the first. (b) As the second is rotated, only part of the light is passed. (c) When the second is 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)

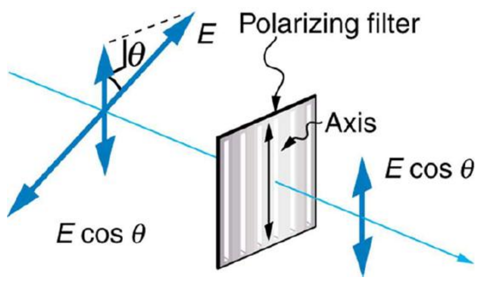


Figure 11.5.7: A polarizing filter transmits only the component of the wave parallel to its axis, $E \cos\theta$, reducing the intensity of any light not polarized parallel to its axis.

✔ 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, $I = 0.100 I_0$. Using this information, the equation $I = I_0 \cos\theta^2$ can be used to solve for the needed angle.

Solution

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

$$\cos\theta = \sqrt{\frac{I}{I_0}} = \sqrt{\frac{0.100 I_0}{I_0}} = 0.3162. \tag{11.5.2}$$

Solving for θ yields

$$\theta = \cos 0.3162^{-2} = 71.6^\circ \tag{11.5.3}$$

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.

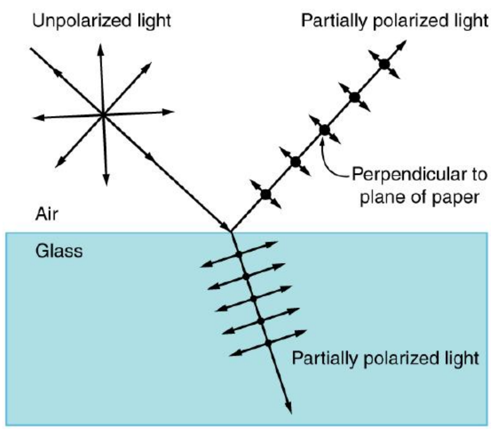


Figure 11.5.8: Polarization by reflection. Unpolarized light has equal amounts of vertical and horizontal polarization. After interaction with a surface, the vertical components are preferentially absorbed or refracted, leaving the reflected light more horizontally polarized. This is akin to arrows striking 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 a angle of reflection θ_b, given by

$$\tan \theta_b = \frac{n_2}{n_1}, \tag{11.5.4}$$

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.

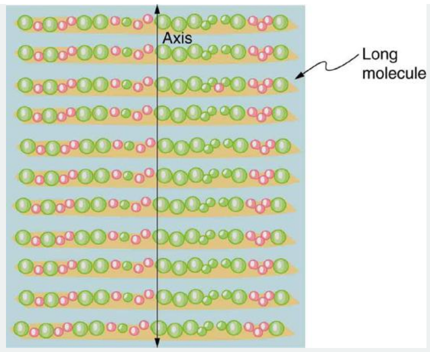


Figure 11.5.9: Long molecules are aligned perpendicular to the axis of a polarizing filter. The component of the electric field in an EM wave perpendicular to these molecules passes through the filter, while the component parallel to the molecules is absorbed.

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.

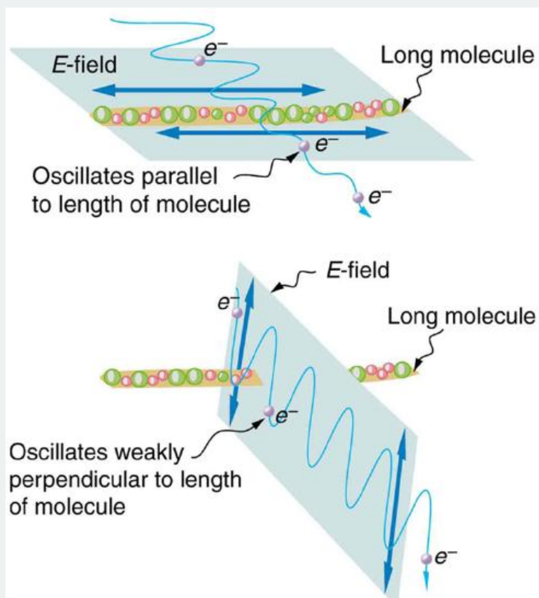


Figure 11.5.10: Artist’s conception of an electron in a long molecule oscillating parallel to the molecule. The oscillation of the electron absorbs energy and reduces the intensity of the component of the EM wave that is parallel to the molecule.

✔ Example 11.5.2: Calculating Polarization by Reflection

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

Strategy:

All we need to solve these problems are the indices of refraction. Air has $n_1 = 100$, 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

\tan{\theta_{b}} = \frac{n_{2}}{n_{1}} \label{27.9.4}

gives

$$\tan\theta_b = \frac{n_2}{n_1} = \frac{1.333}{1.00} = 1.333. \tag{11.5.5}$$

Solving for the angle θ_b yields

$$\theta_b = \tan 1.333^{-1} = 53.1^\circ. \tag{11.5.6}$$

Solution (b):

Similarly, for crown glass and air,

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

Thus,

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

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.

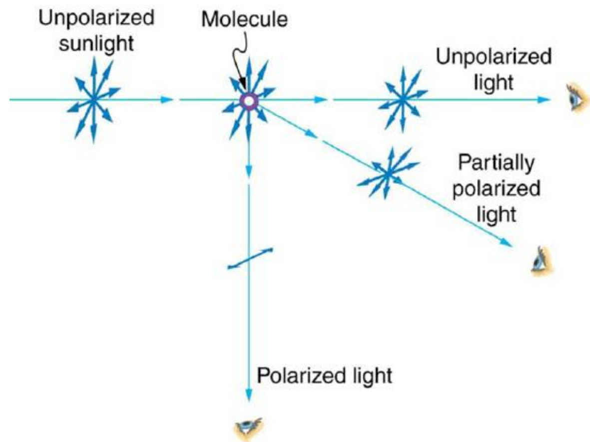


Figure 11.5.11: Polarization by scattering. Unpolarized light scattering from air molecules shakes their electrons perpendicular to the direction of the original ray. The scattered light therefore has a polarization perpendicular to the original direction and none parallel to the original direction.

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.

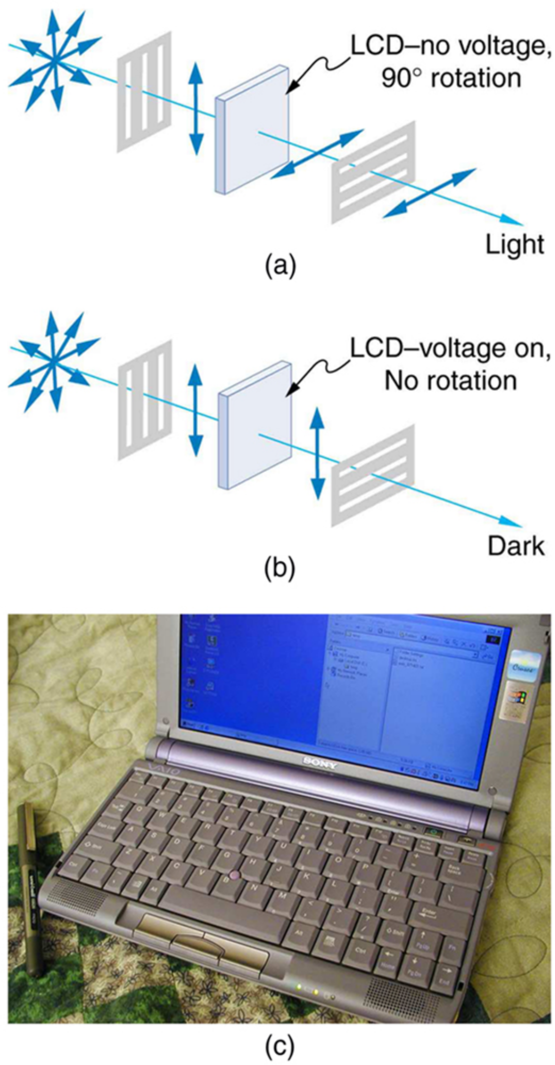


Figure 11.5.12: (a) Polarized light is rotated 90° by a liquid crystal and then passed by a polarizing filter that has its axis perpendicular to the original polarization direction. (b) When a voltage is applied to the liquid crystal, the polarized light is not rotated and is blocked by the filter, making 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.

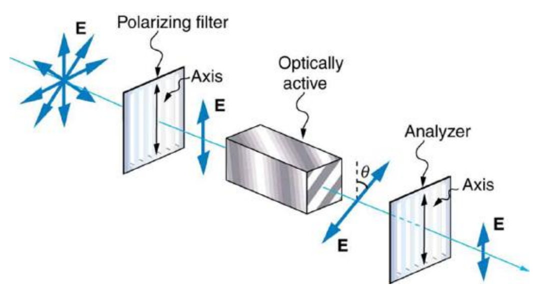


Figure 11.5.13: Optical activity is the ability of some substances to rotate the plane of polarization of light passing through them. The rotation is detected with a polarizing filter or analyzer.

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.

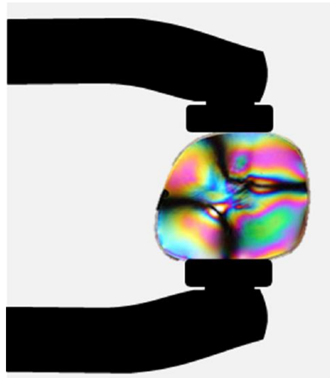


Figure 11.5.14: Optical stress analysis of a plastic lens placed between crossed polarizers. (credit: Infopro, Wikimedia Commons)

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 **berefringent** (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.

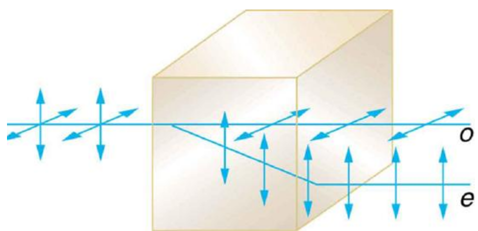


Figure 11.5.15: Birefringent materials, such as the common mineral calcite, split unpolarized beams of light into two. The ordinary ray behaves as expected, but the extraordinary ray does not obey Snell's law.

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

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 $0.74\mu m$ to $300\mu m$
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 range

Key Equations

Speed of EM waves	$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$
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	$I = S_{avg} = \frac{c\varepsilon_0 E_0^2}{2} = \frac{cB_0^2}{2\mu_0} = \frac{E_0 B_0}{2\mu_0}$
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

  $$I = \frac{c\varepsilon_0 E_0^2}{2}$$

  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).

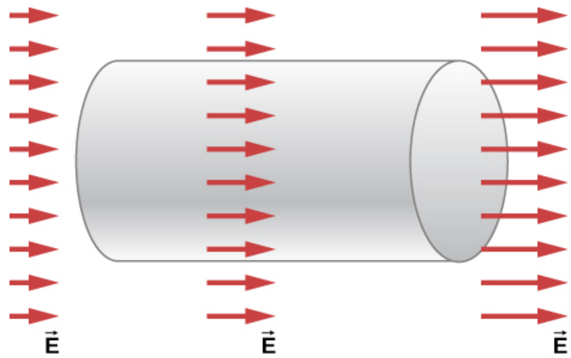
11.7: Electromagnetic Waves (Exercises)

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.

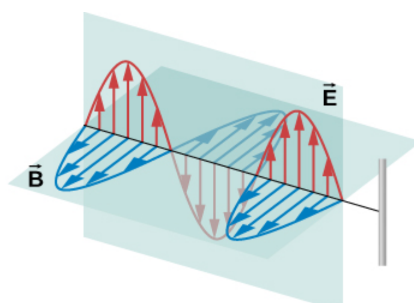


3. Why is it much easier to demonstrate in a student lab that a changing magnetic field induces an electric field than it is to demonstrate that a changing electric field produces a magnetic field?

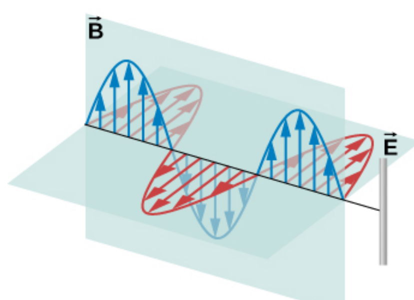
16.3 Plane Electromagnetic Waves

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.

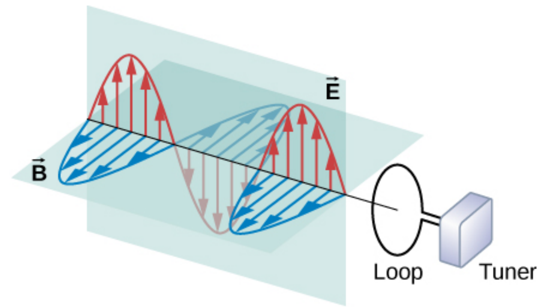


(a)

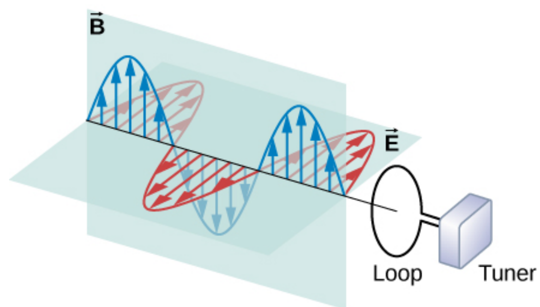


(b)

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



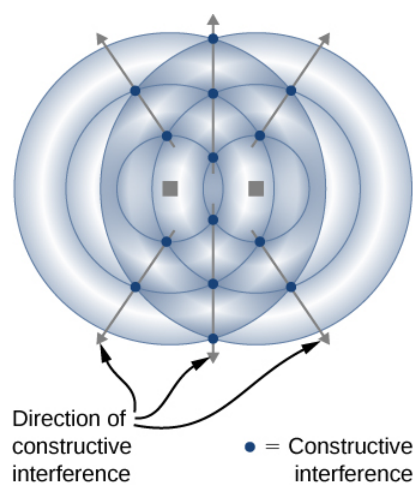
(a)



(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) $f = 10.0kHz$,

(b) $f = \lambda = 750nm$,

(c) $f = 1.25 \times 10^8 Hz$,

(d) $0.30nm$

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 **r** from the axis of two circular parallel plates, produced by placing charge **Q(t)** on the plates is

$$B_{ind} = \frac{\mu_0}{2\pi r}\frac{dQ(t)}{dt}$$

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

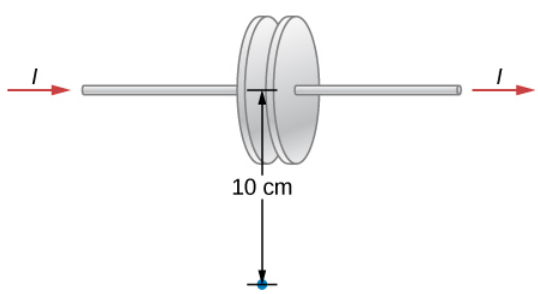
35. A potential difference $V(t) = V_0 sin\omega t$ is maintained across a parallel-plate capacitor with capacitance **C** consisting of two circular parallel plates. A thin wire with resistance **R** connects the centers of the two plates, allowing charge to leak between plates while they are charging.

(a) Obtain expressions for the leakage current $I_{res}(t)$ in the thin wire. Use these results to obtain an expression for the current $I_{real}(t)$ 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 $I_{real}(t)$ with the sum of the displacement current $I_d(t)$ and resistor current $I_{res}(t)$ 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 capacitator?



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 must be the angular frequency ω for a voltage $V(t) = V_0 sin\omega t$ with $V_0 = 100V$ 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 $d = 2mm$ varies sinusoidally as $V = (15mV)cos(150t)$, 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:

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

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

$$E=(6.0\times10^{-3}V/m)sin[2\pi(\frac{x}{18m}-\frac{t}{6.0\times10^{-8}s})]\hat{j}\,.$$

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 $2.15\times10^{-11}T$.

(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 $r_0=3.00cm$ 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 $1.4kW/m^2$.

(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 $1.5\times10^{11}m$ 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

$\vec{B}=B_0(coskz+\omega t)\hat{j}$, where $B_0=5.00\times10^{-10}T$ and $k=3.14\times10^{-2}m^{-1}$.

(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 **7.50μV/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 $1.50 \times 10^{13} m^2$ (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 $1.00 \times 10^{11} V/m$ 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 $1.00 - mm^2$ 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 **2μm** and mass **10μg** 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 **20μg** 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 the ball?

69. Suppose that $\vec{S}_{avg}$ for sunlight at a point on the surface of Earth is $900 W/m^2$.

(a) If sunlight falls perpendicularly on a kite with a reflecting surface of area 0.75m20.75m2, what is the average force on the kite due to radiation pressure?

(b) 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 $1.0 kW/m^2$. A sunbather has a body surface area of 0.8m20.8m2 facing the sun while reclining on a beach chair on a clear day.

(a) how much energy from direct sunlight reaches the sunbather's skin per second?

(b) 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**.

(a) How much work does the radiation pressure do to accelerate the particle from rest in the given time it absorbs the light?

(b) 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.

(a) What is its frequency?

(b) 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.

(b) 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 $1 \times 10^{-12} s$, 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.

(a) Calculate the wavelength of the radio waves from WWVB.

(b) 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 210210rather 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 (**λ/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 $2.0 \times 10^{17} Hz$?

(b) Yellow light of frequency $5.1 \times 10^{14} Hz$?

(c) Gamma rays of frequency $1.0 \times 10^{23} Hz$?

82. For red light of **λ=660nm**, what are f, ω and k?

83. A radio transmitter broadcasts plane electromagnetic waves whose maximum electric field at a particular location is $1.55 \times 10^{-3} V/m$. 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 $f = 2.45 \times 10^9 Hz$ 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 $3.76 \times 10^3 J/kg \cdot {}^\circ C$? 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 $1.00 - mm^2$ 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 $1.00 \times 10^{-12} T$.

(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 **2.50μH**, 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 Centuri, 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 **I** in the wire, with resistance **R** and voltage **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 $1.37 kW/m^2$. 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 $5.972 \times 10^{24} kg$.

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?

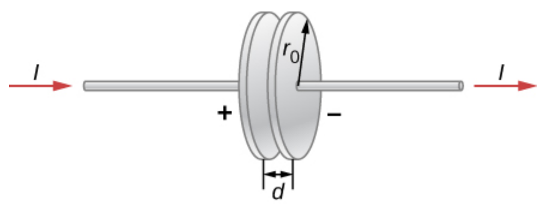
(c) How accurately (in nanoseconds) must you be able to measure the echo time to an airplane 12.0 km away to determine its distance within 10.0 m?

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 **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 **V(t)** and **dV(t)/dt**.

(b) Assuming that **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 **E(t)**, for **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 **V(t)** and **dV(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 $\rho = 2.0 g/cm^3$.

(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.

- **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 $c = 1/\sqrt{\varepsilon_0 \mu_0}$ 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 $100 W/m^2$) 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 is 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. $B_{ind} = \frac{\mu_0}{P2\pi r} I_{ind} = \frac{\mu_0}{2\pi r}\varepsilon_0 \frac{\partial \Phi_E}{\partial t} = \frac{\mu_0}{2\pi r}\varepsilon_0 (A\frac{\partial E}{\partial t}) = \frac{\mu_0}{2\pi r}\varepsilon_0 A(\frac{1}{d}\frac{dV(t)}{dt}) = \frac{\mu_0}{2\pi r}[\frac{\varepsilon_0 A}{d}][\frac{1}{C}\frac{dQ(t)}{dt}] = \frac{\mu_0}{2\pi r}\frac{dQ(t)}{dt}$ because $C = \frac{\varepsilon_0 A}{d}$

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

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

41. 499 s

43. 25 m

45. a. 5.00 V/m;

b. $9.55 \times 10^8 Hz$;

c. 31.4 cm;

d. toward the +**x**-axis;

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

47. $I_d = \pi\varepsilon_0\omega R^2 E_0 sin(kx - \omega t)$

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

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

b. 394 m

53. 11.5 m

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

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

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

59. $20.8 W/m^2$

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

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

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

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

c. 1.12 W

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

67. $F = ma = (p)(\pi r^2), p = \dfrac{ma}{\pi r^2} = \dfrac{\varepsilon_0}{2E_0^2}$

$$E_0 = \sqrt{\frac{2ma}{\varepsilon_0 \pi r^2}} = \sqrt{\frac{2(10^{-8} kg)(0.30 m/s^2)}{(8.854 \times 10^{-12} C^2/N \cdot m^2)(\pi)(2 \times 10^{-6} m)^2}}$$

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

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

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

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

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

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= $\int \vec{S} \cdot d\vec{A} = (\frac{1}{\mu_0} EB)(2\pi rL) = \frac{1}{\mu_0}(\frac{V}{L})(\frac{\mu_0 i}{2\pi r})(2\pi rL) = iV = i^2 R$

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.

CHAPTER OVERVIEW

12: Antenna Systems

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.

12.2: 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 $12.2.1$.

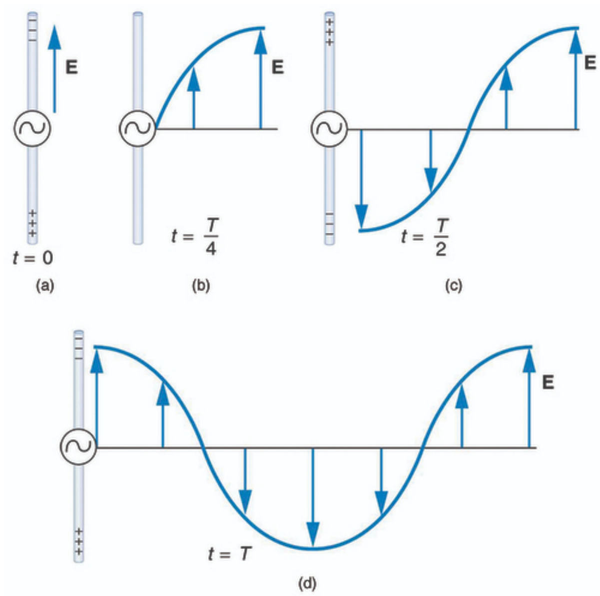


Figure $12.2.1$: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field ($textbf{E}$) propagates away from the antenna at the speed of light, forming part 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 $12.2.2$). 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.

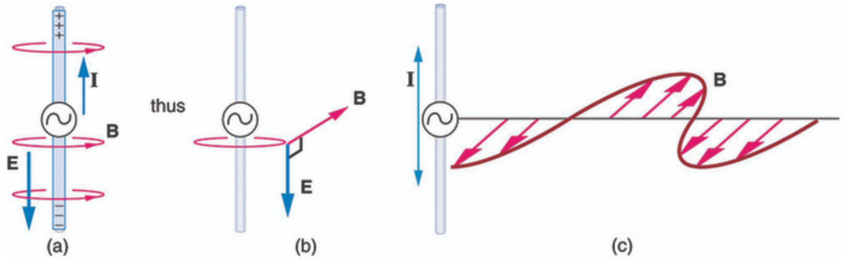


Figure 12.2.2: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (E and B) near the wire are perpendicular; they are shown 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**.

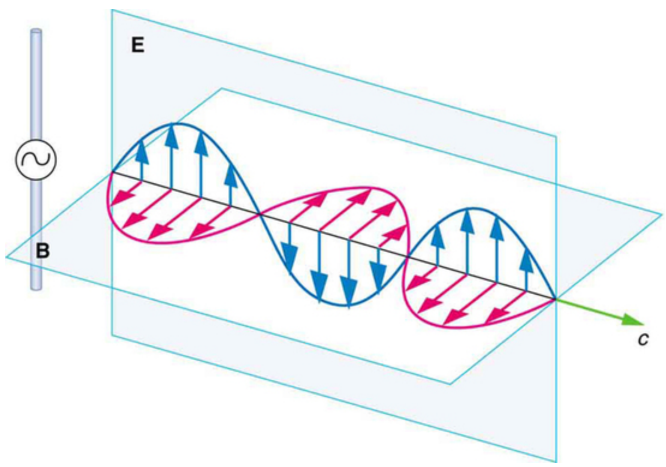


Figure 12.2.3: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields ($textbf{E}$ and $textbf{B}$) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one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

$$\frac{E}{B} = c \tag{12.2.1}$$

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 $1000 V/m$?

Strategy:

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

$$B = \frac{E}{c}. \tag{12.2.2}$$

Solution:

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

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

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.

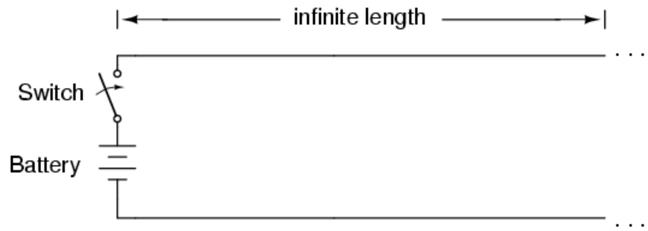
PHET EXPLORATIONS: RADIO WAVES AND ELECTROMAGNETIC FIELDS

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.

12.3: Transmission Lines and Characteristic Impedance

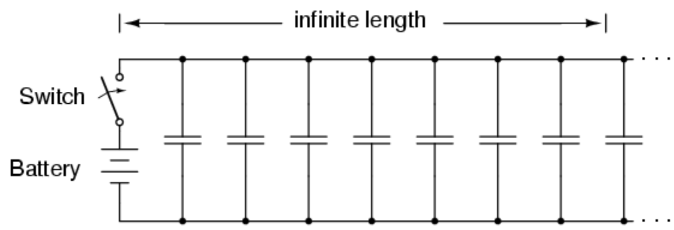
The Parallel Wires of Infinite Length

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)

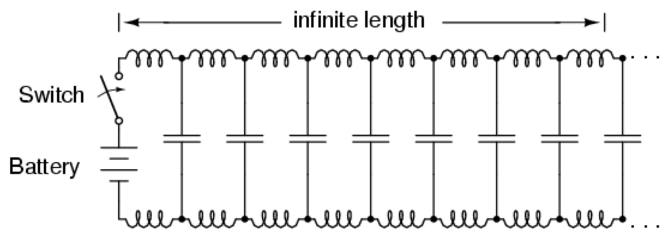


Equivalent circuit showing stray capacitance between conductors.

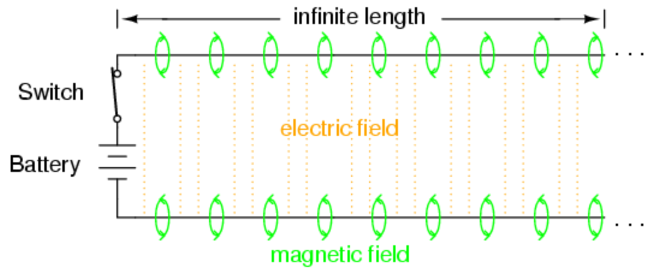
Voltage applied between two conductors creates an electric field between those conductors. Energy is stored in this electric field, and this storage of energy results in an opposition to change in voltage. The reaction of a capacitance against changes in voltage is described by the equation i = C(de/dt), 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:

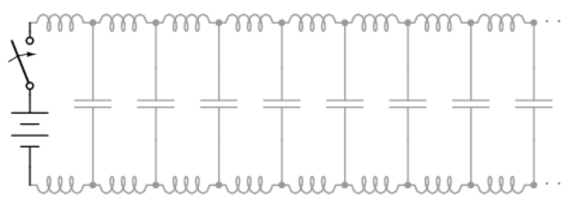


Equivalent circuit showing stray capacitance and inductance.

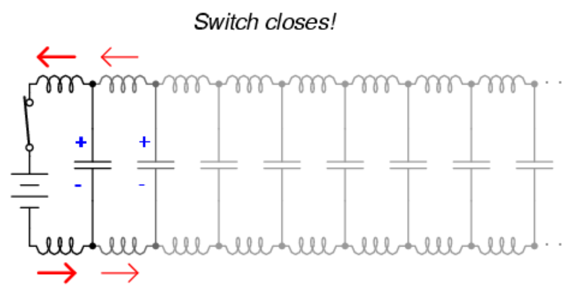


Voltage charges capacitance, current charges inductance.

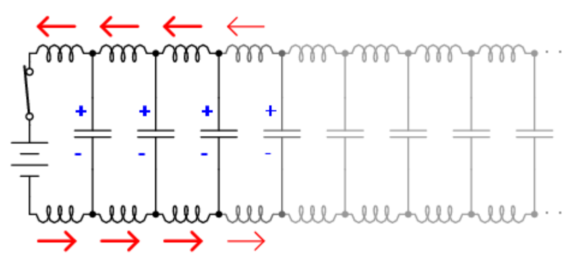
Because the electrons in the two wires transfer motion to and from each other at nearly the speed of light, the “wave front” of voltage and current change will propagate down the length of the wires at that same velocity, resulting in the distributed capacitance and inductance progressively charging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



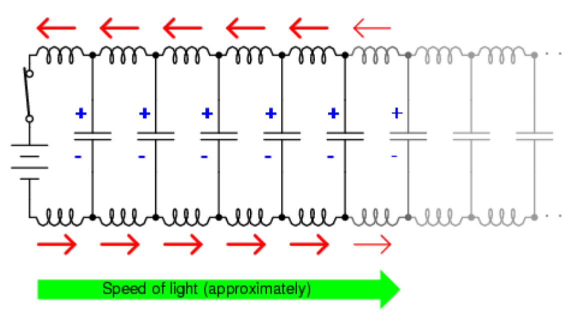
Uncharged transmission line.



Begin wave propagation.



Continue wave propagation.



Propagate at speed of light.

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:

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

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:

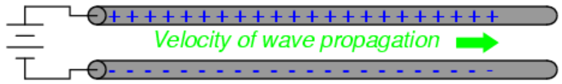
$$\boxed{Z_0 = \frac{138}{\sqrt{k}} \log \frac{d_1}{d_2}}$$

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.

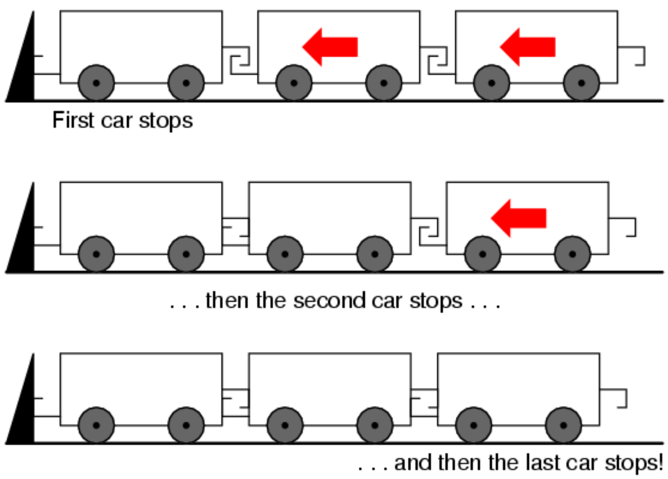
- **14.3: Characteristic Impedance** by Tony R. Kuphaldt is licensed notset. Original source: https://www.allaboutcircuits.com/textbook/alternating-current.

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 Ω "RG-58/U" cable I measured with an ohmmeter years ago had been infinitely long, I actually would have been able to measure 50 Ω 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.

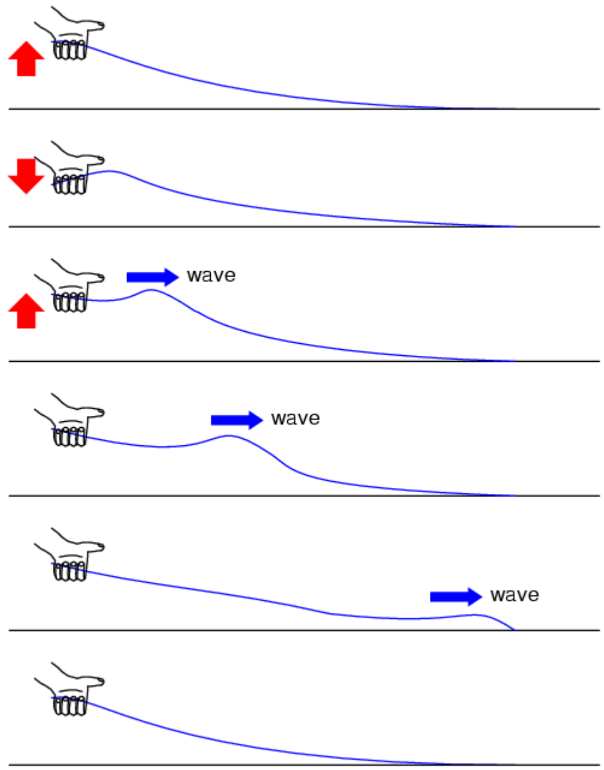
A signal propagating from the source-end of a transmission line to the load-end is called an *incident wave*. The propagation of a signal from load-end to source-end (such as what happened in this example with current encountering the end of an open-circuited transmission line) is called a *reflected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

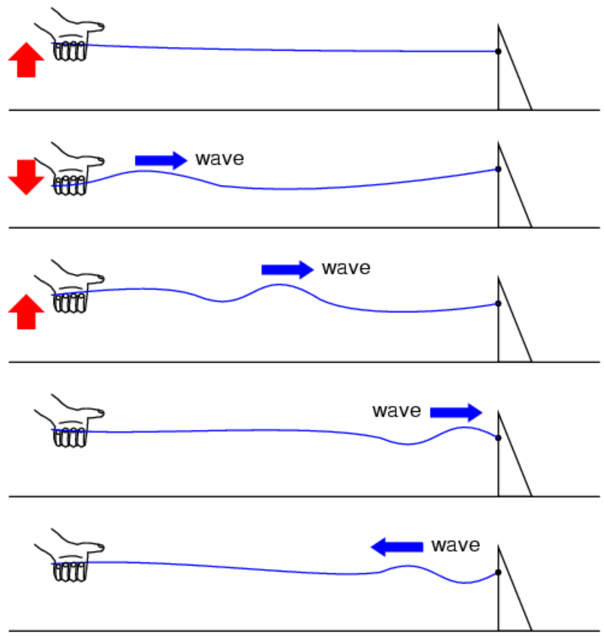
A similar phenomenon takes place if the end of a transmission line is short-circuited: when the voltage wave-front reaches the end of the line, it is reflected back to the source, because voltage cannot exist between two electrically common points. When this reflected wave reaches the source, the source 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.

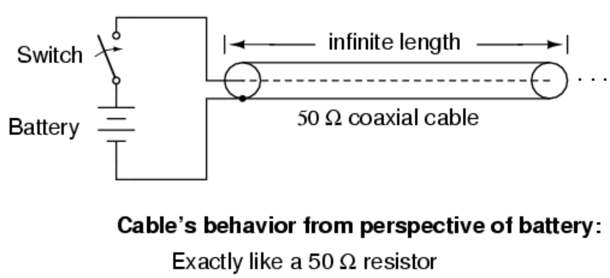
Reflections may be eliminated from the transmission line if the load’s impedance exactly equals the characteristic (“surge”) impedance of the line. For example, a 50 Ω coaxial cable that is either open-circuited or short-circuited will reflect all of the incident energy back to the source. However, if a 50 Ω 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 Ω characteristic impedance and infinite length behaves exactly like a 50 Ω resistance as measured from one end. (Figure below)

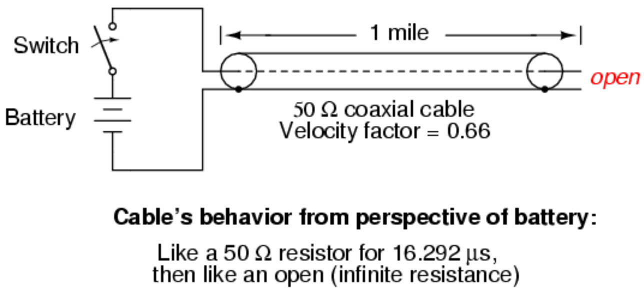
If we cut this line to some finite length, it will behave as a 50 Ω resistor to a constant source of DC voltage for a brief time, but then behave like an open- or a short-circuit, depending on what condition we leave the cut end of the line: open (Figure below)

or shorted. (Figure below)

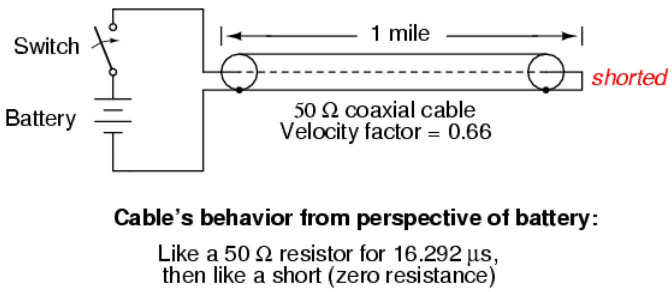
However, if we *terminate* the line with a 50 Ω resistor, the line will once again behave as a 50 Ω resistor, indefinitely: the same as if it were of infinite length again: (Figure below)



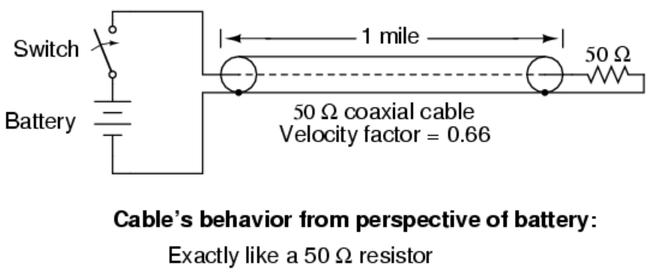
Infinite transmission line looks like resistor.



One mile transmission.



Shorted transmission line.



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.

12.5: “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^o 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 Ω 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 Ω 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 Ω 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 Ω 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.

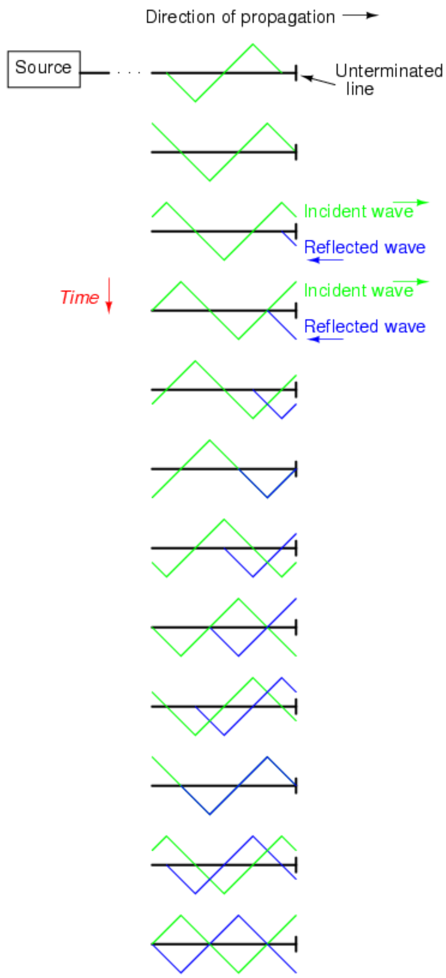
Review

- 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.

12.6: Standing Waves and Resonance

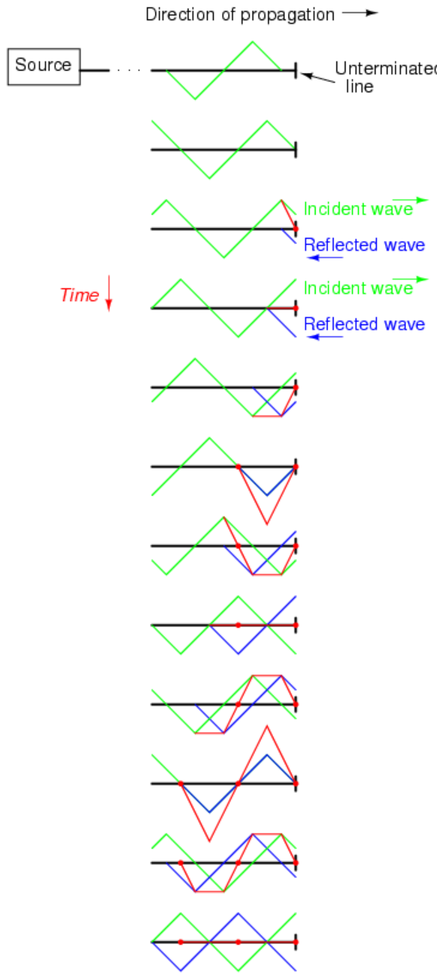
Whenever there is a mismatch of impedance between transmission line and load, reflections will occur. If the incident signal is a continuous AC waveform, these reflections will mix with more of the oncoming incident waveform to produce stationary waveforms called *standing waves*.

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)



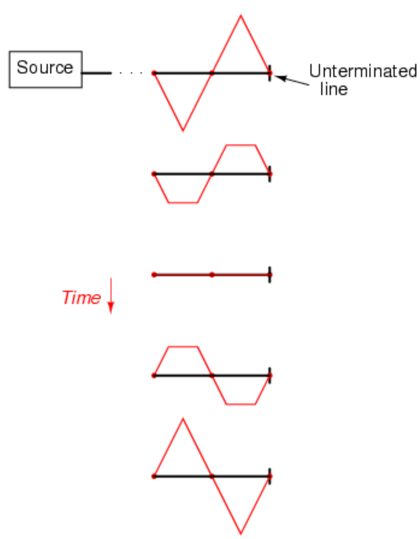
Incident wave reflects off end of unterminated transmission line.

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



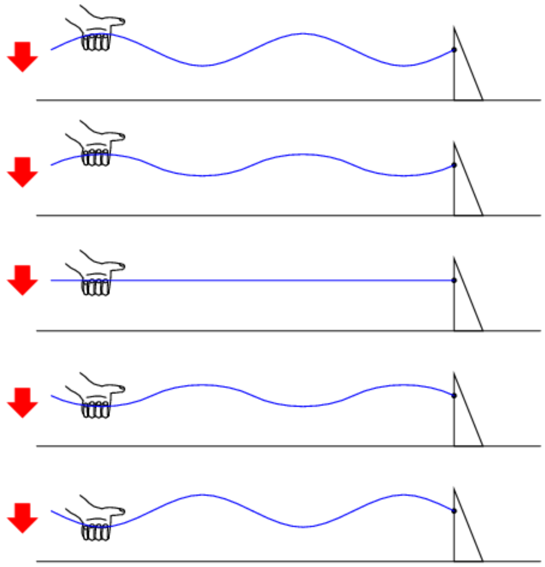
The sum of the incident and reflected waves is a stationary wave.

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)



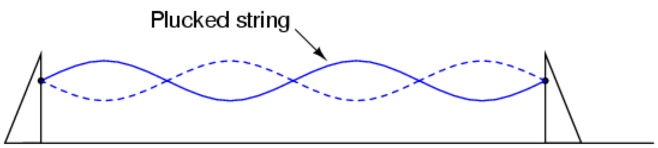
The standing wave does not propgate along the transmission line.

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)



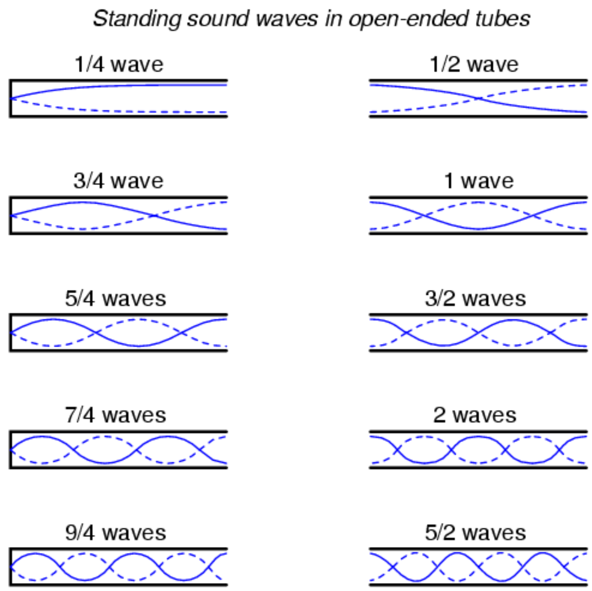
Standing waves on a rope.

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)

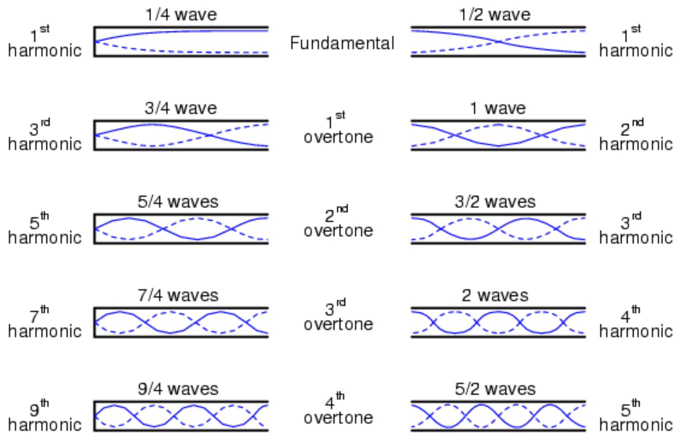


Standing sound waves in open ended tubes.

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 75 Ω: (Figure below)

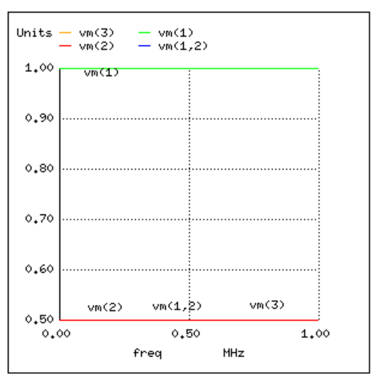
Perfectly matched transmission line.

Using SPICE to simulate the circuit, we'll specify the transmission line (`t1`) with a 75 Ω 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 75 Ω 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
```

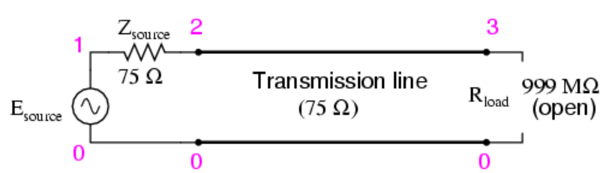
Running this simulation and plotting the source impedance drop (as an indication of current), the source voltage, the line's source-end voltage, and the load voltage, we see that the source voltage—shown as `vm(1)` (voltage magnitude between node 1 and the implied ground point of node 0) on the graphic 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.

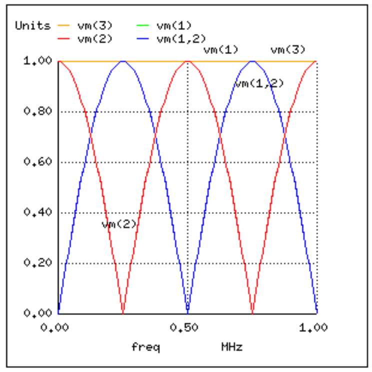
In a system where all impedances are perfectly matched, there can be no standing waves, and therefore no resonant "peaks" or "valleys" in the Bode plot.

Now, let's change the load impedance to 999 MΩ, 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.

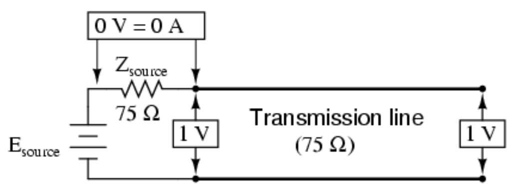
```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 999meg
.ac lin 101 1m 1meg
* Using ''Nutmeg'' program to plot analysis
.end
```



Resonances on open transmission line.

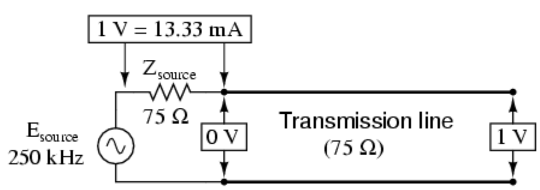
Here, both the supply voltage `vm(1)` and the line's load-end voltage `vm(3)` remain steady at 1 volt. The other voltages dip and peak at different frequencies along the sweep range of 1 mHz to 1 MHz. There are five points of interest along the horizontal axis of the analysis: 0 Hz, 250 kHz, 500 kHz,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



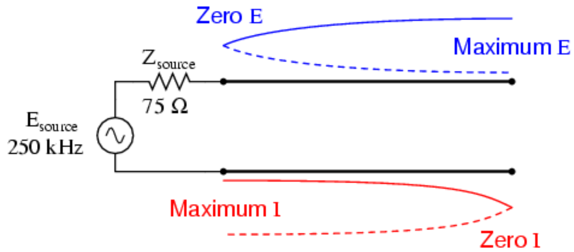
At f=0: input: V=1, I=0; end: V=1, I=0.

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)



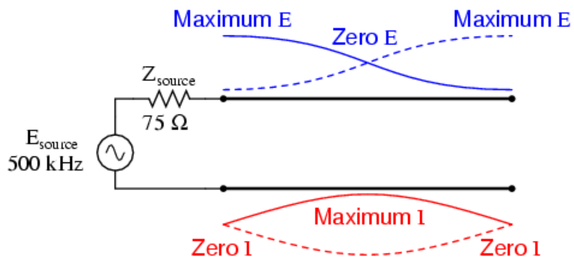
At f=250 KHz: input: V=0, I=13.33 mA; end: V=1 I=0.

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)



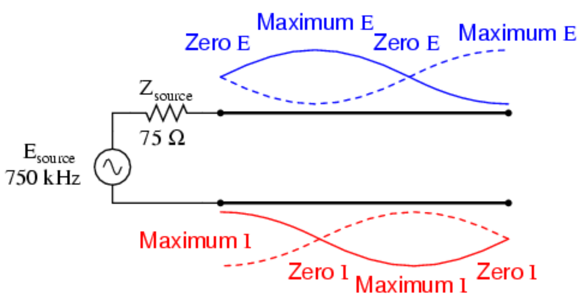
Open end of transmission line shows current node, voltage antinode at open end.

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)



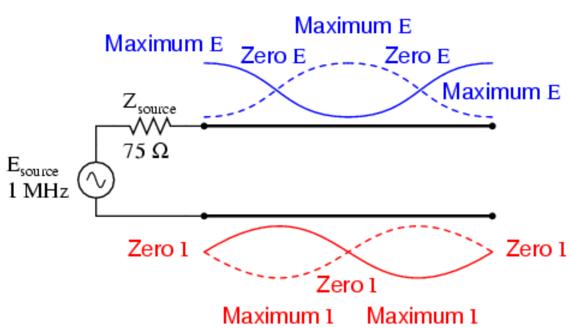
Full standing wave on half wave open transmission line.

At 750 kHz, the plot looks a lot like it was at 250 kHz: zero source-end voltage (`vm(2)`) and maximum current (`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.

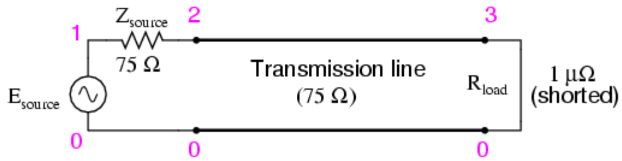
When the supply frequency sweeps up to 1 MHz, a full standing wave exists on the transmission line. At this point, the source-end of the line experiences the same voltage and current amplitudes as the load-end: full voltage and zero current. In essence, the source "sees" an open circuit at the point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



Double standing waves on full wave open transmission line.

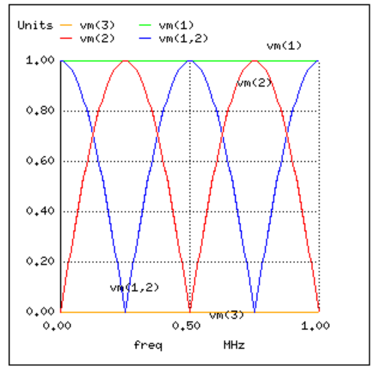
In a similar fashion, a short-circuited transmission line generates standing waves, although the node and antinode assignments for voltage and current are reversed: at the shorted end of the line, there will be zero voltage (node) and maximum current (antinode).

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 μΩ load impedance: (Figure below)

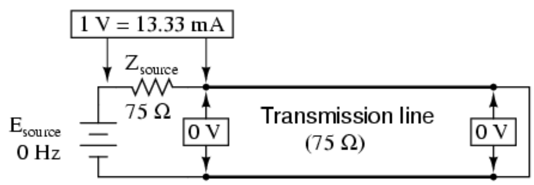


Shorted transmission line.

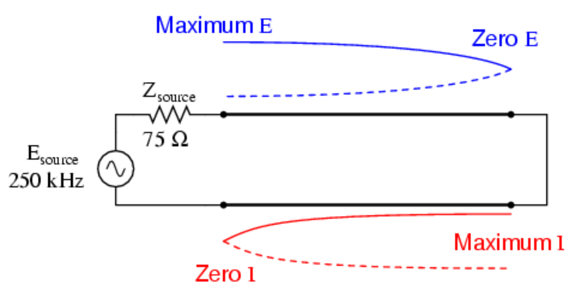
```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 1u
.ac lin 101 1m 1meg
* Using ‘‘Nutmeg’’ program to plot analysis
.end
```



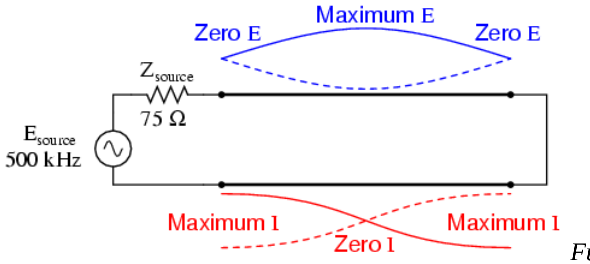
Resonances on shorted transmission line



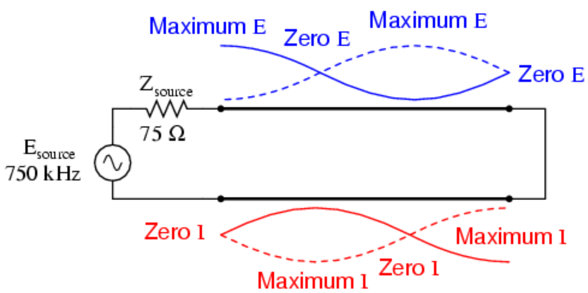
At f=0 Hz: input: V=0, I=13.33 mA; end: V=0, I=13.33 mA.



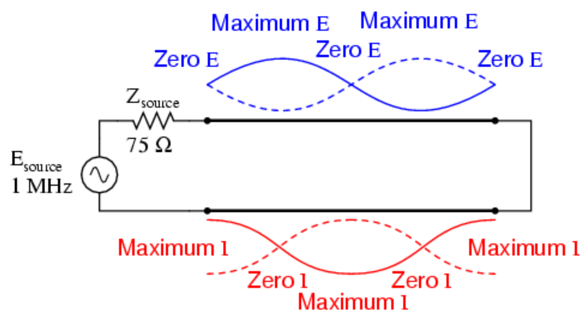
Half wave standing wave pattern on 1/4 wave shorted transmission line.



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



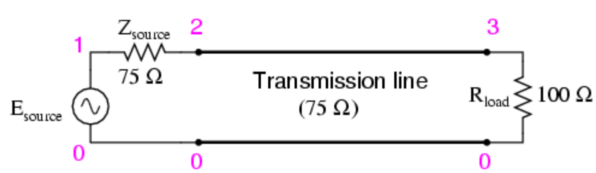
1 1/2 standing wavepattern on 3/4 wave shorted transmission line.



Double standing waves on full wave shorted transmission line.

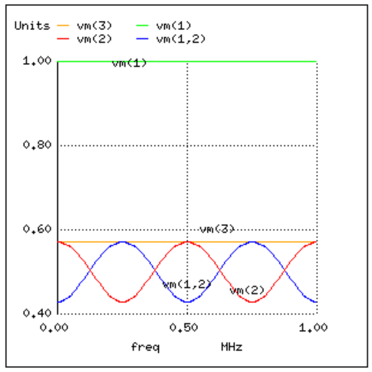
In both these circuit examples, an open-circuited line and a short-circuited line, the energy reflection is total: 100% of the incident wave reaching the line's end gets reflected back toward the source. If, however, the transmission line is terminated in some impedance other than an open or a short, 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 Ω resistor instead of a 75 Ω 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
```



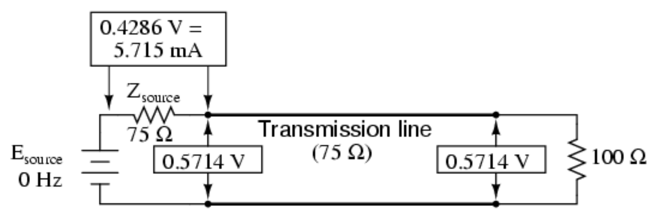
Weak resonances on a mismatched transmission line

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).

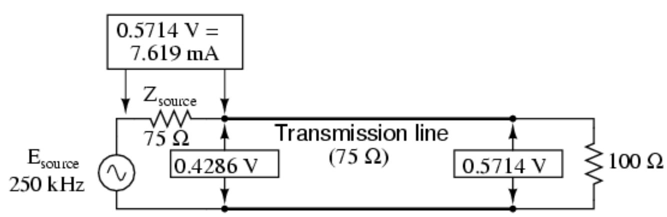
```
Transmission line
v1 1 0 ac 1 sin
rsource 1 2 75
t1 2 0 3 0 z0=75 td=1u
rload 3 0 100
.ac lin 5 1m 1meg
.print ac v(1,2) v(1) v(2) v(3)
.end
```

```
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
```

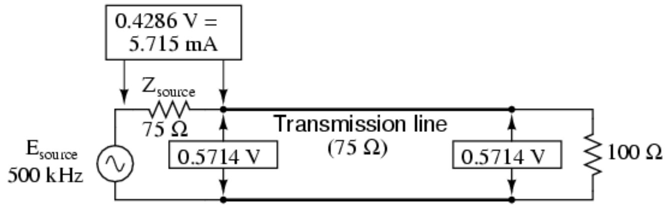
At all frequencies, the source voltage, `v(1)` , remains steady at 1 volt, as it should. The load voltage, `v(3)` , also remains steady, but at a lesser voltage: 0.5714 volts. However, both the line input voltage (`v(2)`) and the voltage dropped across the source's 75 Ω impedance (`v(1,2)` , indicating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



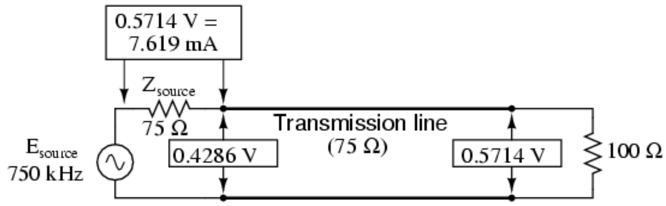
At f=0 Hz: input: V=0.57.14, I=5.715 mA; end: V=0.5714, I=5.715 mA.



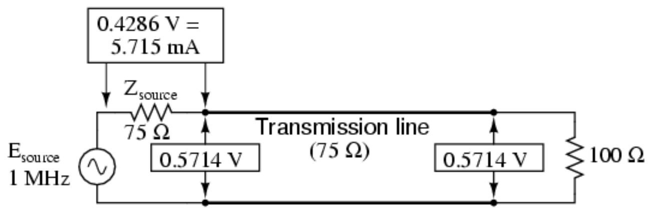
At f=250 KHz: input: V=0.4286, I=7.619 mA; end: V=0.5714, I=7.619 mA.



At f=500 KHz: input: V=0.5714, I=5.715 mA; end: V=5.714, I=5.715 mA.



At f=750 KHz: input: V=0.4286, I=7.619 mA; end: V=0.5714, I=7.619 mA.



At f=1 MHz: input: V=0.5714, I=5.715 mA; end: V=0.5714, I=0.5715 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 75 Ω line terminated by a 100 Ω 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 100 Ω divided by the characteristic impedance of 75 Ω yields a quotient of exactly 1.333, matching the previous calculation very closely.

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

$$\boxed{SWR = \frac{Z_{load}}{Z_0} \text{ or } \frac{Z_0}{Z_{load}} \quad \textit{which ever is greater}}$$

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 50 Ω 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 50 Ω 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, which ever 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.

12.7: 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

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

which implies that the magnetic field B is very weak relative to the electric field E.

CHAPTER OVERVIEW

13: Propagation of Electromagnetic Waves

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.

13.2: 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.

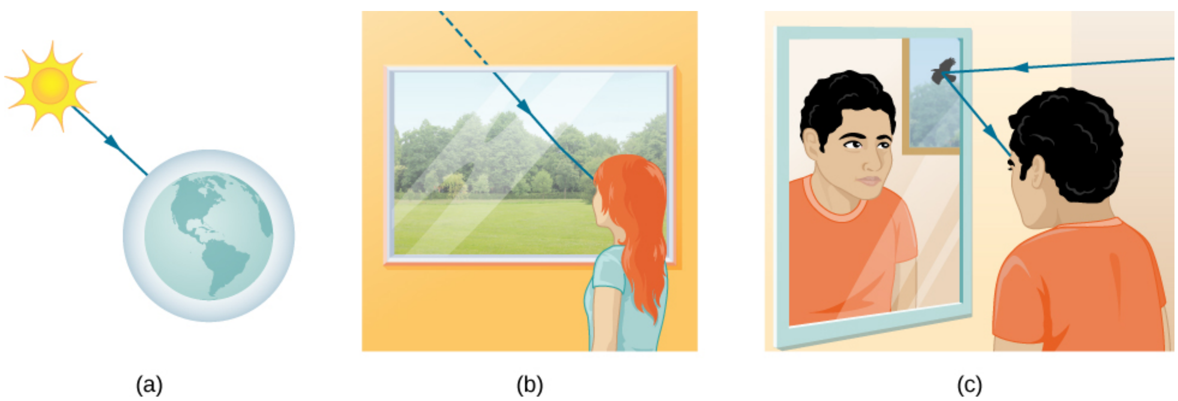


Figure $13.2.1$: Three methods for light to travel from a source to another location. (a) Light reaches the upper atmosphere of Earth, traveling through empty space directly from the source. (b) Light can reach a person by traveling through media like air and glass. (c) Light can also reflect from an 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.

13.3: 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:

$$\theta_r = \theta_i \tag{13.3.1}$$

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.

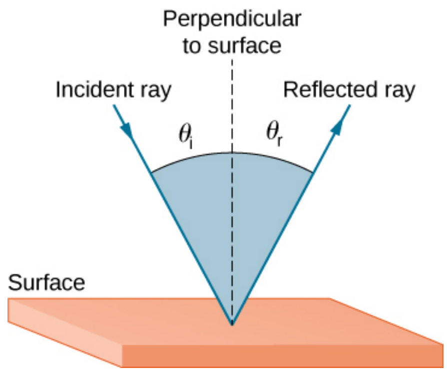


Figure 13.3.1: The law of reflection states that the angle of reflection equals the angle of incidence—$\theta_r=\theta_i$. The angles are measured relative to the perpendicular to the surface at the point where the ray strikes the surface.

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 $13.3.1a$.

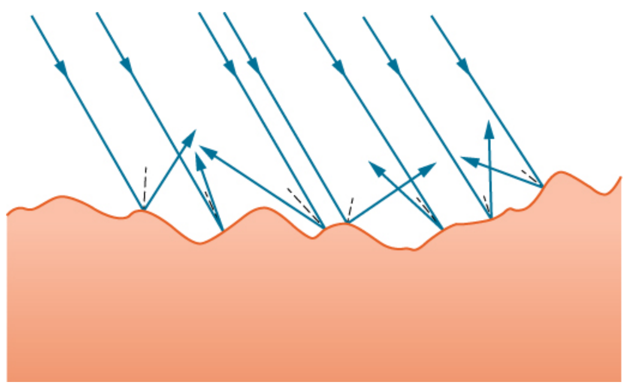


Figure 13.3.2: Light is diffused when it reflects from a rough surface. Here, many parallel rays are incident, but they are reflected at many different angles, because the surface is rough.

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.

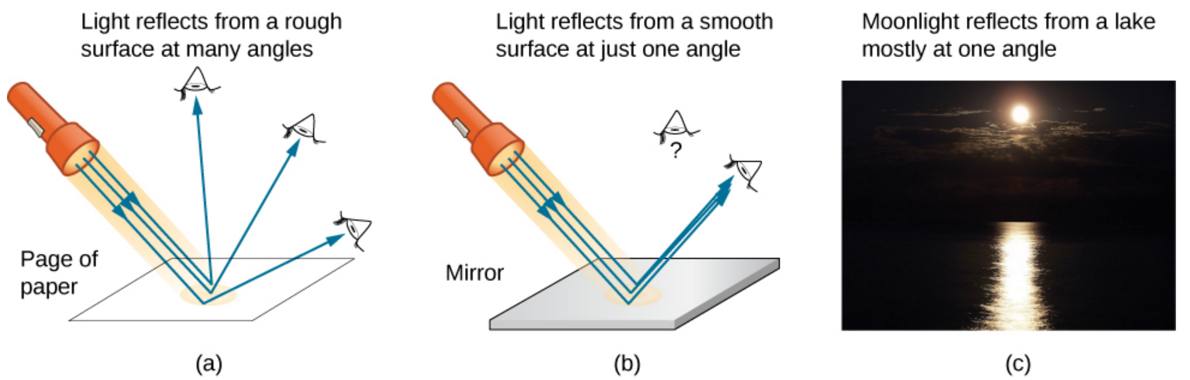


Figure 13.3.3: (a) When a sheet of paper is illuminated with many parallel incident rays, it can be seen at many different angles, because its surface is rough and diffuses the light. (b) A mirror illuminated by many parallel rays reflects them in only one direction, because its surface is very smooth. 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.

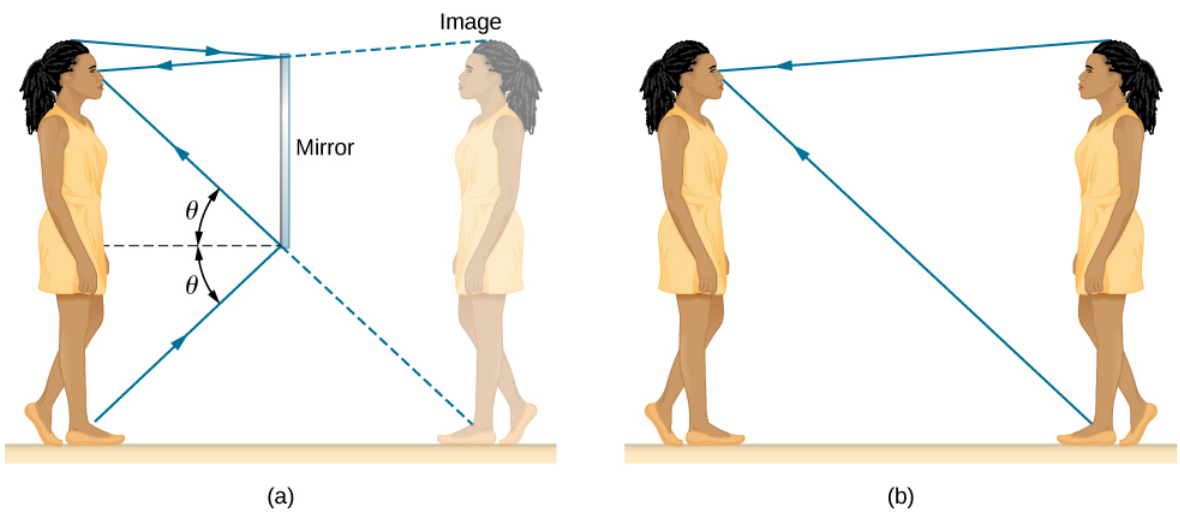


Figure 13.3.4: (a) Your image in a mirror is behind the mirror. The two rays shown are those that strike the mirror at just the correct angles to be reflected into the eyes of the person. The image appears to be behind the mirror at the same distance away as (b) if you were looking at your twin directly, 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.

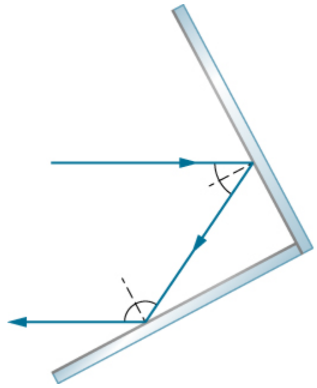


Figure 13.3.5: A light ray that strikes two mutually perpendicular reflecting surfaces is reflected back exactly parallel to the direction from which it came.

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.

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.

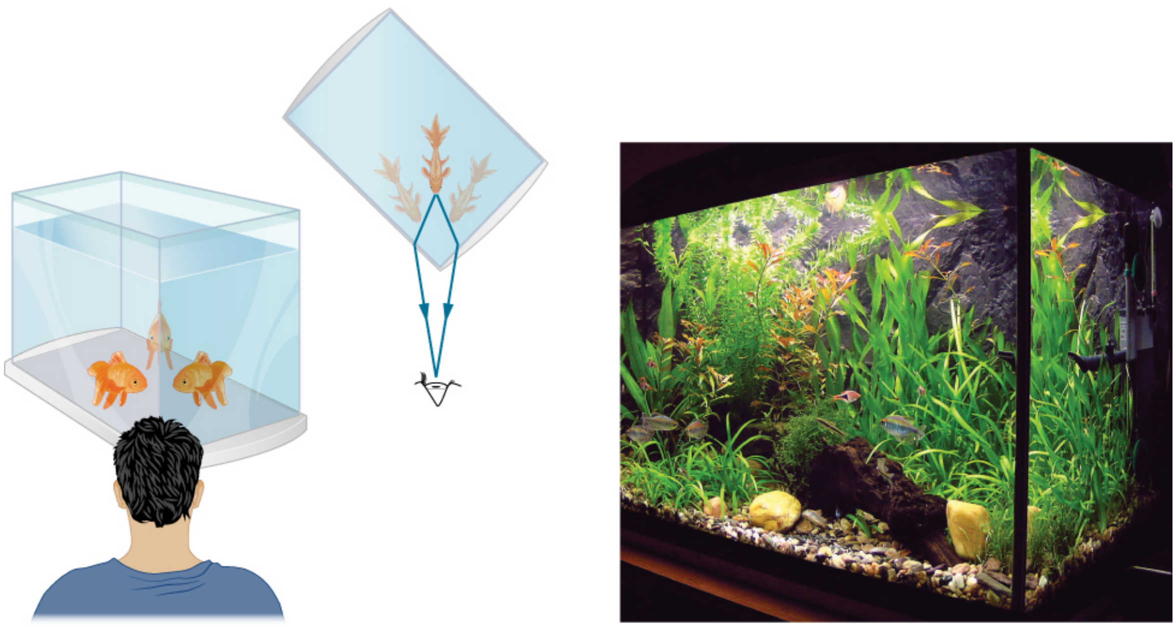
13.4: Refraction of Rays

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.



(a) (b)

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

$$c = 2.99792458 \times 10^8 m/s \equiv 3.00 \times 10^8 m/s \tag{13.4.1}$$

where the approximate value of 3.00×10^8m/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**:

$$n = \frac{c}{v} \tag{13.4.2}$$

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≥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 MediaFor 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

$$\begin{aligned} v &= \frac{3.00 \times 10^8 m/s}{1.923} \\ &= 1.56 \times 10^8 m/s. \end{aligned}$$

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.

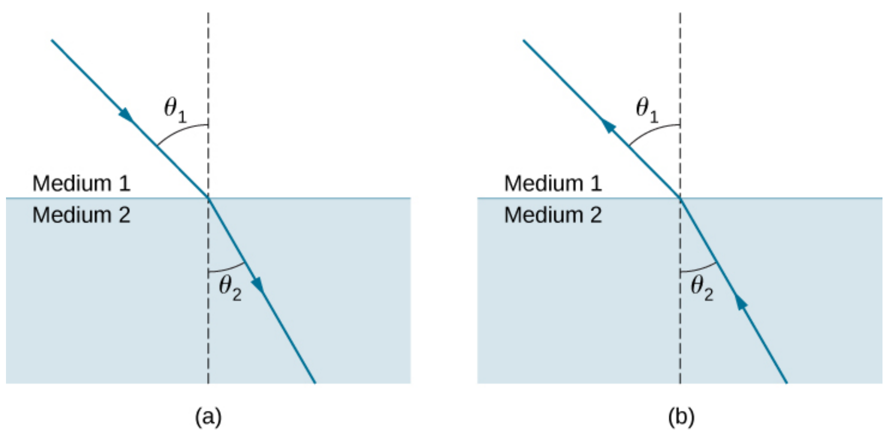


Figure 13.4.2: The change in direction of a light ray depends on how the index of refraction changes when it crosses from one medium to another. In the situations shown here, the index of refraction is greater in medium 2 than in medium 1. (a) A ray of light moves closer to the perpendicular when entering 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

$$n_1 \sin \theta_1 = n_2 \sin \theta_2. \tag{13.4.3}$$

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 θ_1=30.0°. 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°

13.5: Application- Line-of-Sight Transmission

Learning Objectives

- Defeine 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.

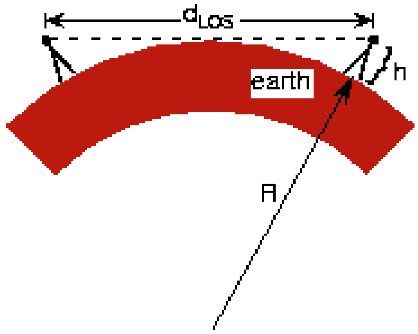


Figure 6.5.1 Two antennae are shown each having the same height. Line-of-sight transmission means the transmitting and receiving antennae can "see" each other as shown. The maximum distance at which they can see each other, $\mathbf{d_{LOS}}$, occurs when the sighting line just grazes the earth's surface.

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

$$d_{LOS} = 2\sqrt{2hR + H^2} \simeq 2\sqrt{2Rh}$$

where **R** is the earth's radius and $\mathbf{R = 6.38 \times 10^6\ m}$.

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

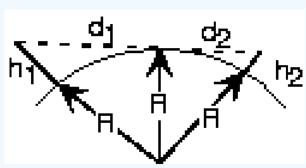


Figure 6.5.2

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 to 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

$$\sqrt{2h_1R}$$

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 **λf = 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**.

13.6: 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.

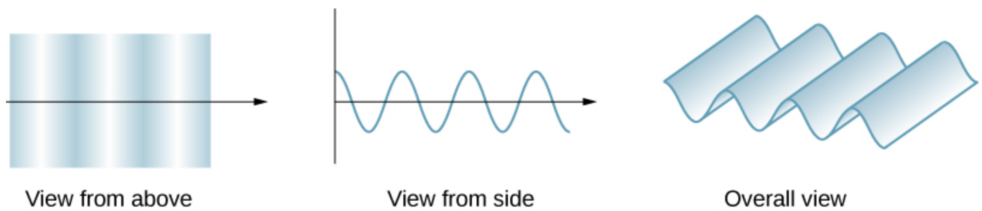


Figure 13.6.1: A transverse wave, such as an electromagnetic light wave, as viewed from above and from the side. The direction of propagation is perpendicular to the wave fronts (or wave crests) and is represented by a ray.

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.

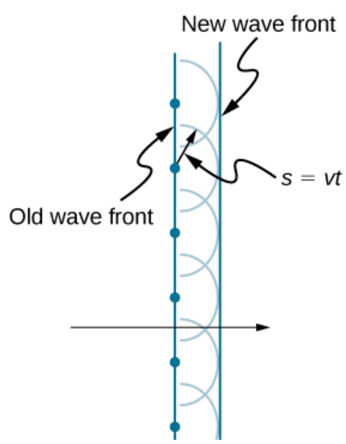


Figure 13.6.2: Huygens's principle applied to a straight wave front. Each point on the wave front emits a semicircular wavelet that moves a distance s=vt. The new wave front is a line tangent to the wavelets.

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.

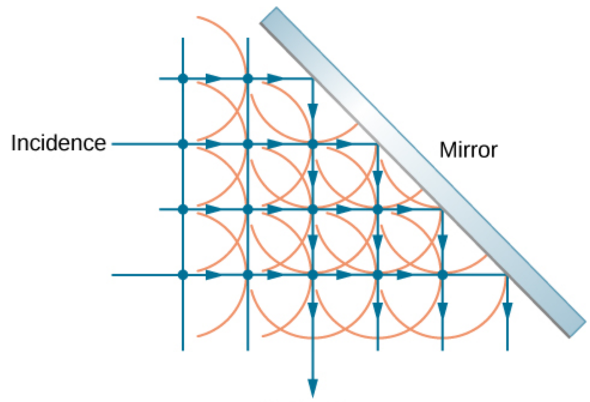


Figure 13.6.3: Huygens's principle applied to a plane wave front striking a mirror. The wavelets shown were emitted as each point on the wave front struck the mirror. The tangent to these wavelets shows that the new wave front has been reflected at an angle equal to the incident angle. The direction 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).

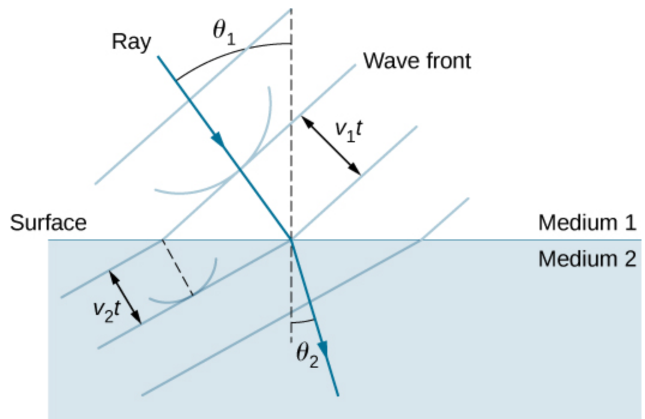


Figure 13.6.4: Huygens's principle applied to a plane wave front traveling from one medium to another, where its speed is less. The ray bends toward the perpendicular, since the wavelets have a lower speed in the second medium.

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 $n_1 < n_2$.

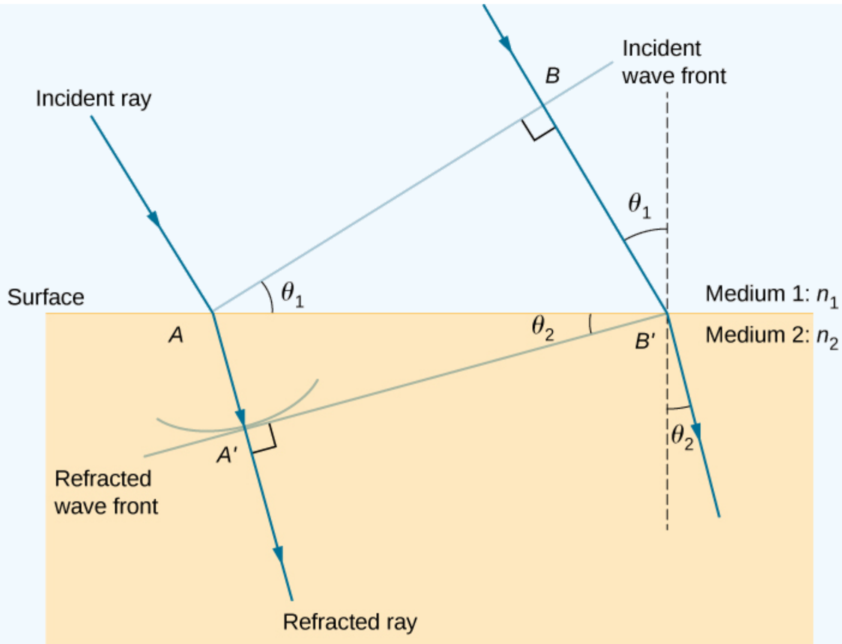


Figure 13.6.5: Geometry of the law of refraction from medium 1 to medium 2.

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 ∠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'=cΔt/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

$$\underbrace{n_1 \sin\theta_1 = n_2 \sin\theta_2}_{\text{Snell's law}}.$$

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,

$$\lambda = \frac{c}{f} = \frac{330\,m/s}{1000\,s^{-1}} = 0.33\,m,$$

about three times smaller than the width of the doorway).

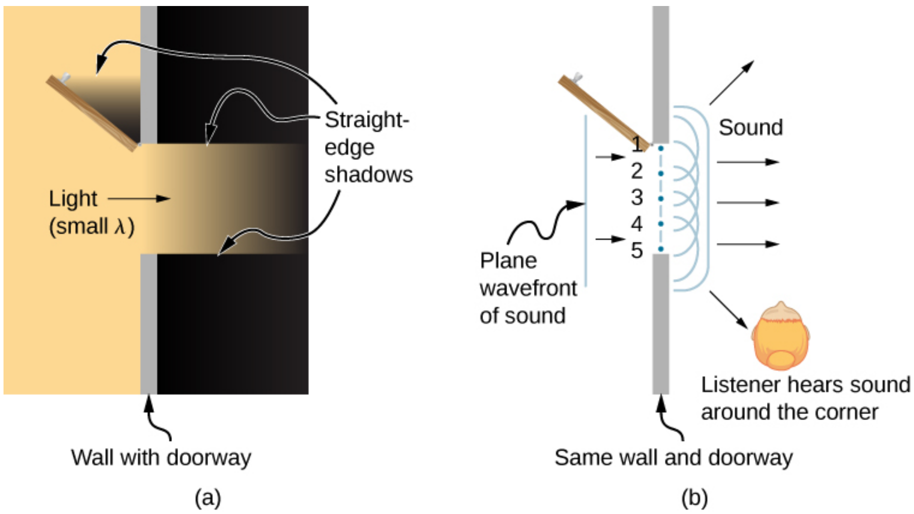


Figure 13.6.6: (a) Light passing through a doorway makes a sharp outline on the floor. Since light's wavelength is very small compared with the size of the door, it acts like a ray. (b) Sound waves bend into all parts of the room, a wave effect, because their wavelength is similar to the size of the 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.

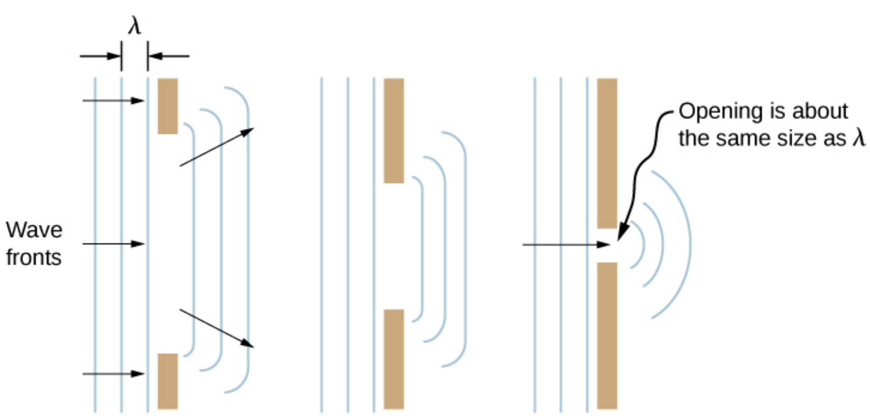


Figure 13.6.7: Huygens's principle applied to a plane wave front striking an opening. The edges of the wave front bend after passing through the opening, a process called diffraction. The amount of bending is more extreme for a small opening, consistent with the fact that wave characteristics are most noticeable for interactions with objects about the same size as the wavelength.

- **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.

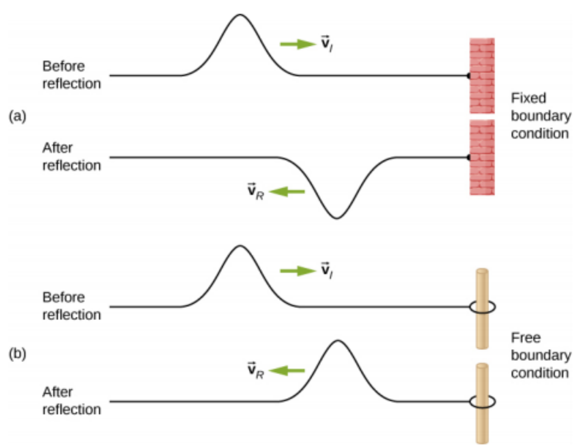


Figure $13.7.1$: (a) One end of a string is fixed so that it cannot move. A wave propagating on the string, encountering this fixed boundary condition, is reflected 180°(π rad) out of phase with respect to the incident wave. (b) One end of a string is tied to a solid ring of negligible mass on a frictionless 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°(0 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 way is exactly the same as the amplitude of the incident wave, but the reflected wave is reflected $180^o \pi$ 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.

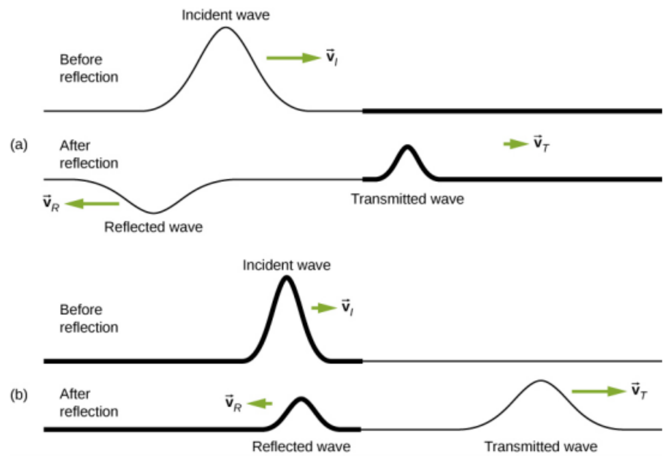


Figure $13.7.2$: Waves traveling along two types of strings: a thick string with a high linear density and a thin string with a low linear density. Both strings are under the same tension, so a wave moves faster on the low-density string than on the high-density string. (a) A wave moving from a low-speed to a high-speed medium results in a reflected wave that is 180°(π 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**.

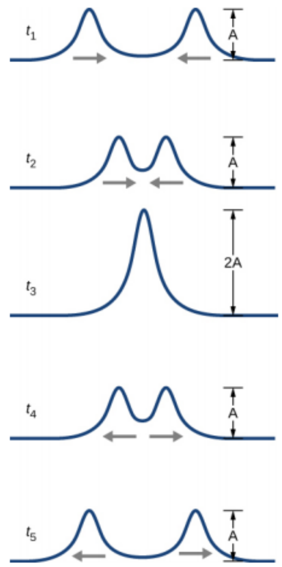


Figure 13.7.3: Two pulses moving toward one another experience interference. The term interference refers to what happens when two waves overlap.

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

$$\frac{\partial^2 y(x,t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y(x,t)}{\partial t^2}. \tag{13.7.1}$$

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.

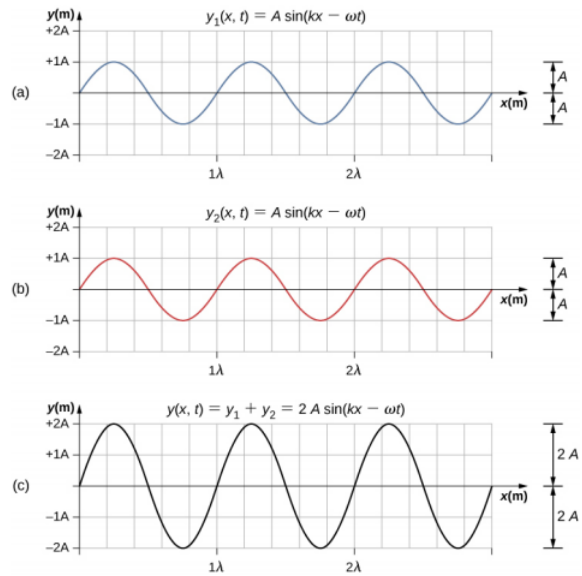


Figure $13.7.4$: Constructive interference of two identical waves produces a wave with twice the amplitude, but the same wavelength.

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.

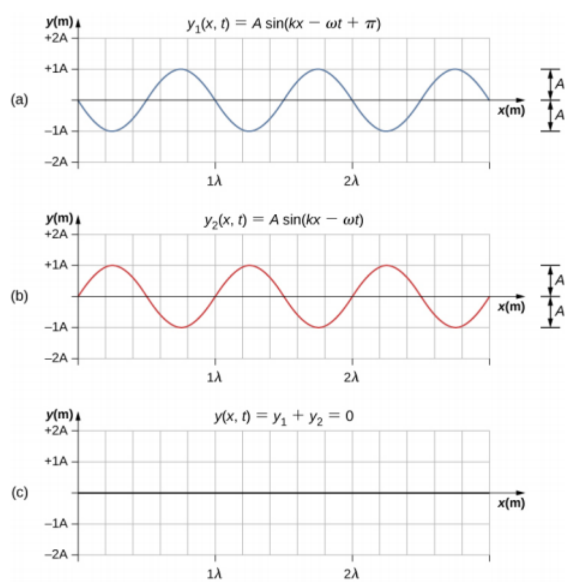


Figure 13.7.5: Destructive interference of two identical waves, one with a phase shift of 180°(π rad), produces zero amplitude, or complete cancellation.

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.

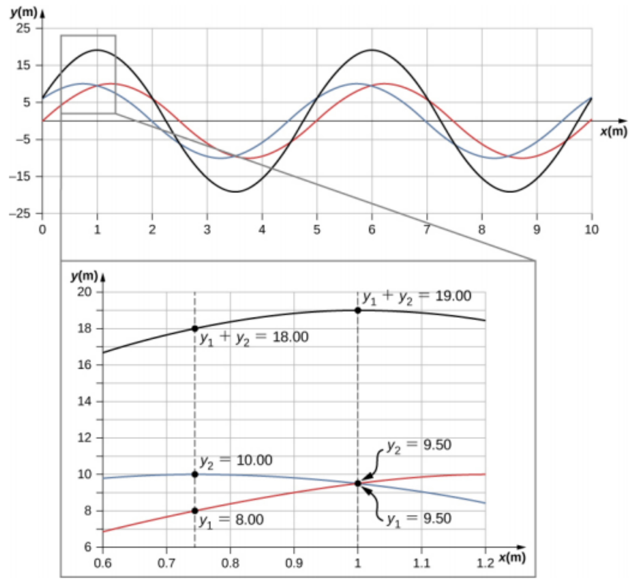


Figure 13.7.6: When two linear waves in the same medium interfere, the height of resulting wave is the sum of the heights of the individual waves, taken point by point. This plot shows two waves (red and blue) added together, along with the resulting wave (black). These graphs represent the height of 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.

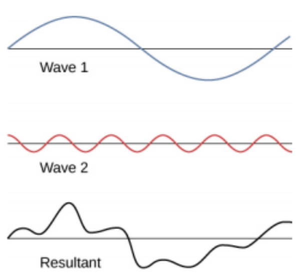


Figure 13.7.7: Superposition of nonidentical waves exhibits both constructive and destructive interference.

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:

$$\begin{aligned} y_1(x,t) &= A\sin(kx-\omega t+\phi), \\ y_2(x,t) &= A\sin(kx-\omega t). \end{aligned}$$

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:

$$y_R(x,t) = y_1(x,t)+y_2(x,t) = A\sin(kx-\omega t+\phi)+A\sin(kx-\omega t). \tag{13.7.2}$$

The resultant wave can be better understood by using the trigonometric identity:

$$\sin u + \sin v = 2\sin\left(\frac{u+v}{2}\right)\cos\left(\frac{u-v}{2}\right), \tag{13.7.3}$$

where u = kx - ωt + ϕ and v = kx - ωt. The resulting wave becomes

$$\begin{aligned} y_R(x,t) &= y_1(x,t)+y_2(x,t) = A\sin(kx-\omega t+\phi)+A\sin(kx-\omega t) \\ &= 2A\sin\left(\frac{(kx-\omega t+\phi)+(kx-\omega t)}{2}\right)\cos\left(\frac{(kx-\omega t+\phi)-(kx-\omega t)}{2}\right) \\ &= 2A\sin\left(kx-\omega t+\frac{\phi}{2}\right)\cos\left(\frac{\phi}{2}\right). \end{aligned}$$

This equation is usually written as

$$y_R(x,t) = 2A\cos\left(\frac{\phi}{2}\right)\sin\left(kx-\omega t+\frac{\phi}{2}\right). \tag{13.7.4}$$

The resultant wave has the same wave number and angular frequency, an amplitude of A_R = [2A cos$\left(\frac{\phi}{2}\right)$], 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.

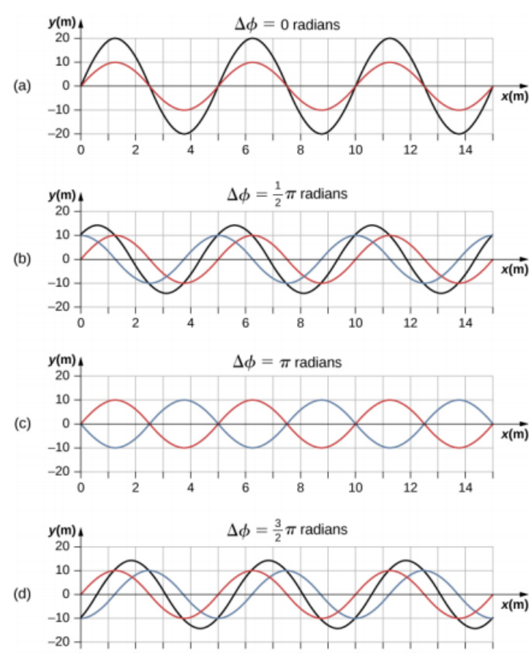


Figure 13.7.8: Superposition of two waves with identical amplitudes, wavelengths, and frequency, but that differ in a phase shift. The red wave is defined by the wave function y_1(x, t) = A sin(kx − ωt) and the blue wave is defined by the wave function y_2(x, t) = A sin(kx − ωt + ϕ). The black line shows 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).

13.8: 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.

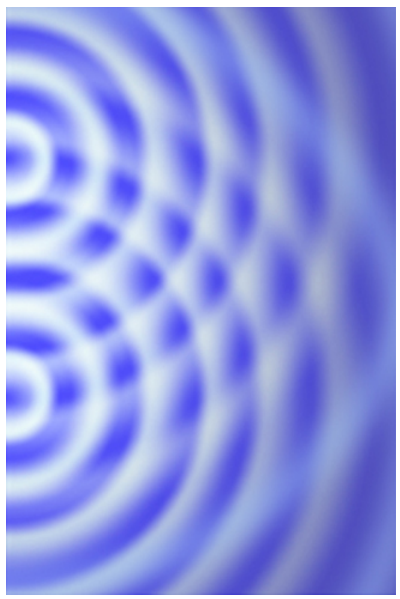


Figure 13.8.1: Photograph of an interference pattern produced by circular water waves in a ripple tank. Two thin plungers are vibrated up and down in phase at the surface of the water. Circular water waves are produced by and emanate from each plunger.

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 ($d \leq 1\,mm$), and the distance between the screen and the slits is D(≈1m), which is much greater than **d.**

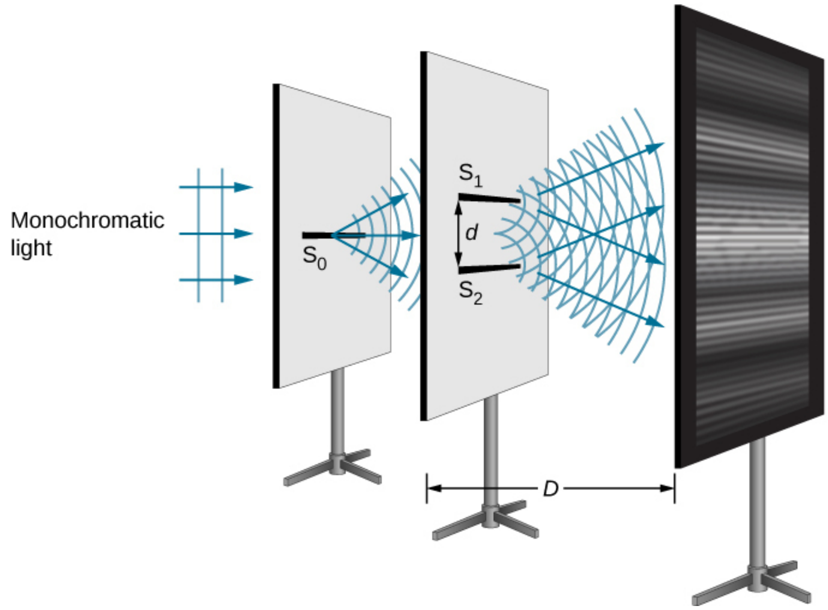


Figure 13.8.2: The double-slit interference experiment using monochromatic light and narrow slits. Fringes produced by interfering Huygens wavelets from slits S_1 and S_2 are observed on the screen.

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.

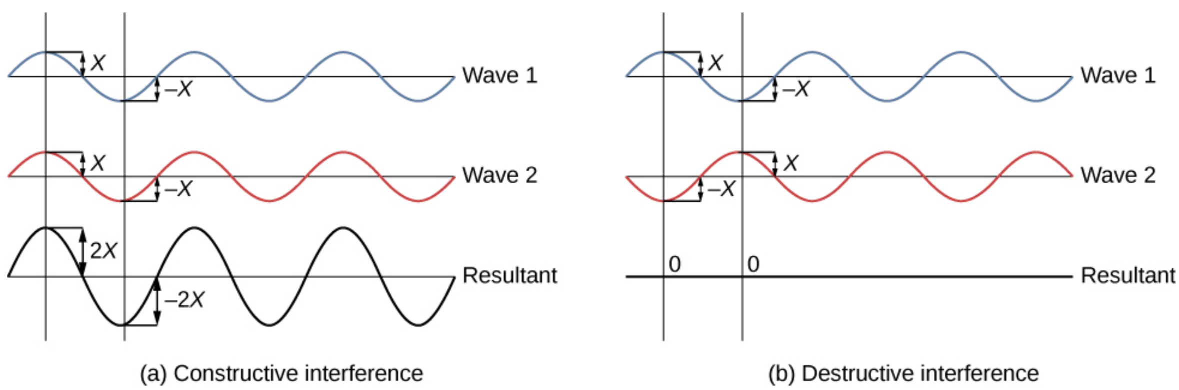


Figure 13.8.3: The amplitudes of waves add. (a) Pure constructive interference is obtained when identical waves are in phase. (b) Pure destructive interference occurs when identical waves are exactly out of phase, or shifted by half a wavelength.

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.

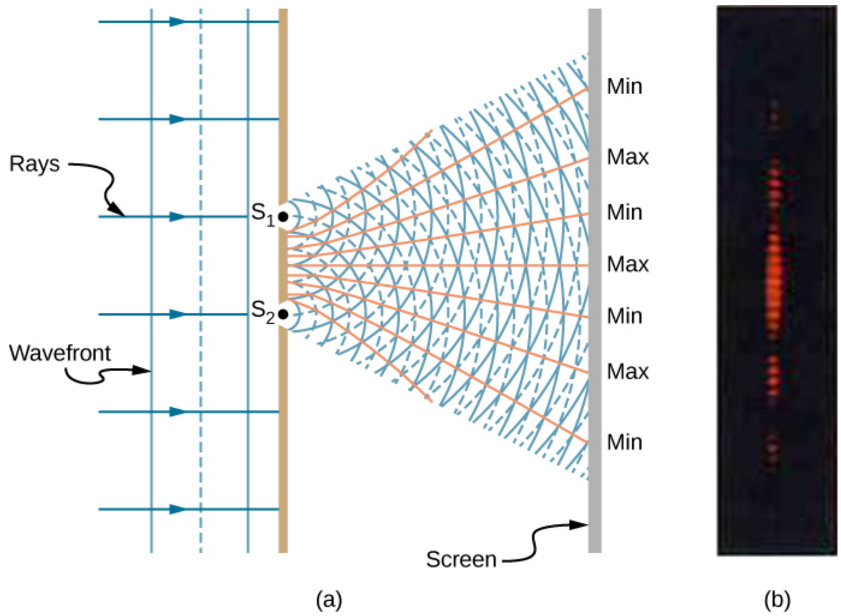


Figure 13.8.4: Double slits produce two coherent sources of waves that interfere. (a) Light spreads out (diffracts) from each slit, because the slits are narrow. These waves overlap and interfere constructively (bright lines) and destructively (dark regions). We can only see this if the light falls onto 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)λ, (3 / 2)λ, (5 / 2)λ, 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 $m = 0, \pm 1, \pm 2, \pm 3 \ldots$

$$\underbrace{\Delta l = \left(m + \frac{1}{2}\right)\lambda}_{\text{destructuve interference}}$$

for $m = 0, \pm 1, \pm 2, \pm 3 \ldots$

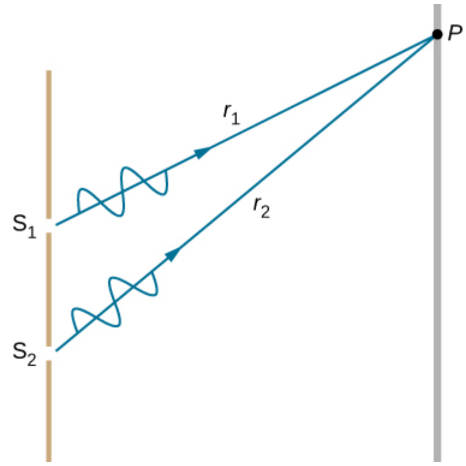


Figure 13.8.5: Waves follow different paths from the slits to a common point P on a screen. Destructive interference occurs where one path is a half wavelength longer than the other—the waves start in phase but arrive out of phase. Constructive interference occurs where one path is a whole wavelength longer than the other—the waves start out and arrive in phase.

13.9: 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	hen 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	$c = 2.99792458 \times 10^8 m/s \approx 3.00 \times 10^8 m/s$
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 $c = 2.99792458 \times 10^8 m/s \approx 3.00 \times 10^8 m/s$.
- 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.

13.10: 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 $2.012 \times 10^8 m/s$, and identify the most likely substance based on Table 1.1.

29. In what substance in Table 1.1 is the speed of light $2.290 \times 10^8 m/s$?

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 $3.84 \times 10^5 km$ away, would the light first arrive on Earth?

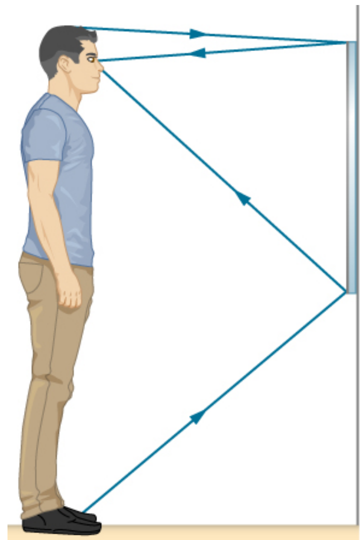
31. Components of some computers communicate with each other through optical fibers having an index of refraction $n = 1.55$. What time in nanoseconds is required for a signal to travel 0.200 m through such a fiber?

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 $1.50 \times 10^{-6}\,s$?

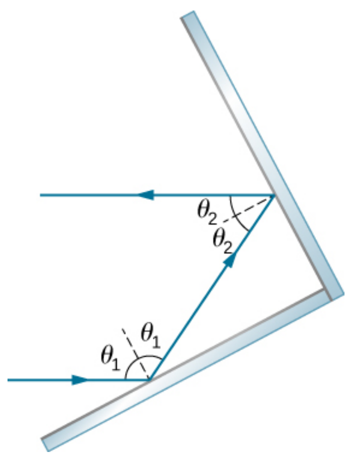
1.2 The Law of Reflection

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.

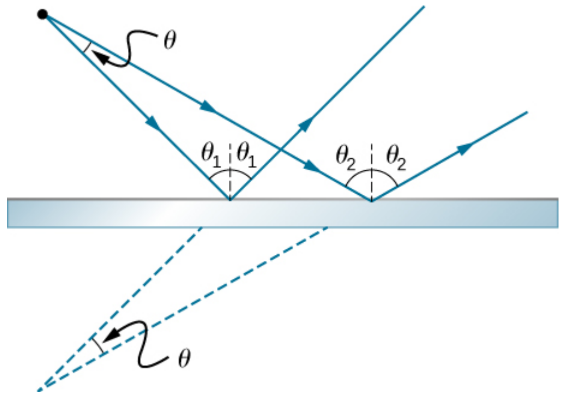
35. Show that when light reflects from two mirrors that meet each other at a right angle, the outgoing ray is parallel to the incoming ray, as illustrated below.



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 $3.84 \times 10^{8} m$ 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°20° with respect to the surface. What are the angles of reflection and refraction?

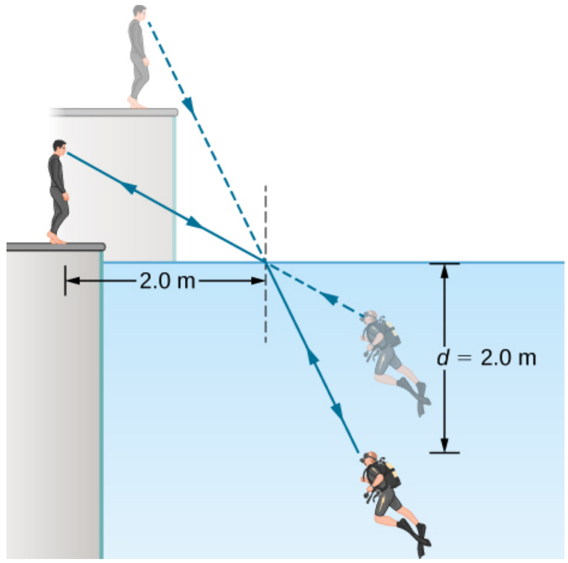
40. When a light ray crosses from water into glass, it emerges at an angle of 30° with respect to the normal of the interface. What is its angle of incidence?

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?

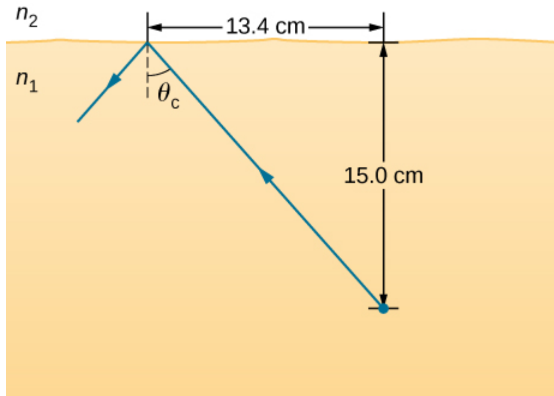
50. Suppose you are using total internal reflection to make an efficient corner reflector. If there is air outside and the incident angle is 45.0°, what must be the minimum index of refraction of the material from which the reflector is made?

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?

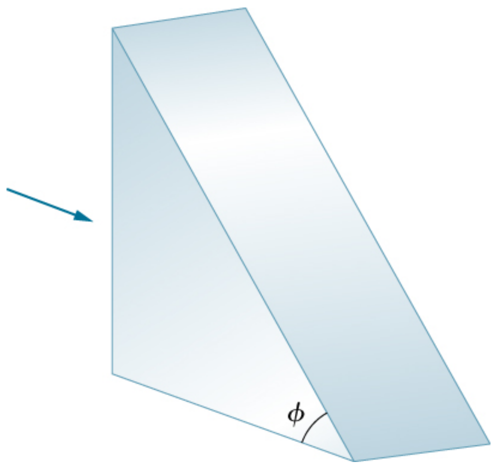


A light ray travels from an object placed in a medium n 1 at 15.0 centimeters below the horizontal interface with medium n 2. This ray gets totally internally reflected with theta c as critical angle. The horizontal distance between the object and the point of incidence is 13.4 centimeters.

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?

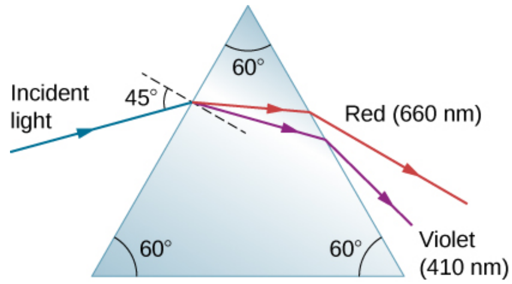
59. A ray of 610-nm light goes from air into fused quartz at an incident angle of 55.0°. At what incident angle must 470 nm light enter flint glass to have the same angle of refraction?

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 $150 W/m^2$, 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 $1.00 kW/m^2$ to reduce the intensity to $10.0 W/m^2$?

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.

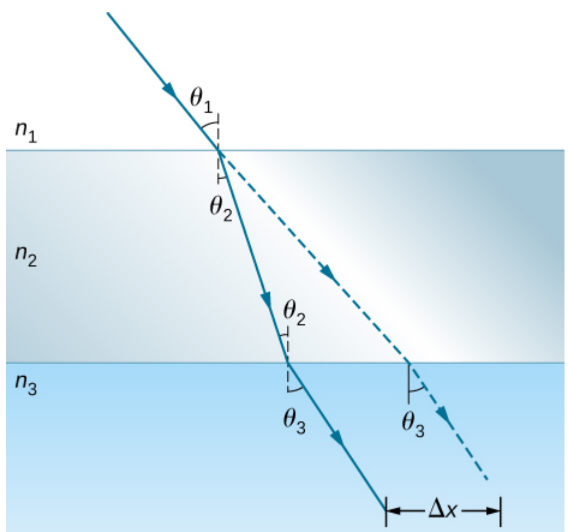
(a) Use this estimate along with the known diameter of Earth's orbit to obtain a rough value of the speed of light.

(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 $^{\circ}C/s$, assuming it is 80.0 absorbed? The aluminum beaker has a mass of 30.0 grams and contains 250 grams of water.

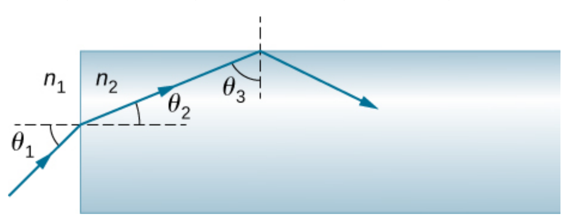
(b) Do the polarizing filters get hot? Explain.

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.

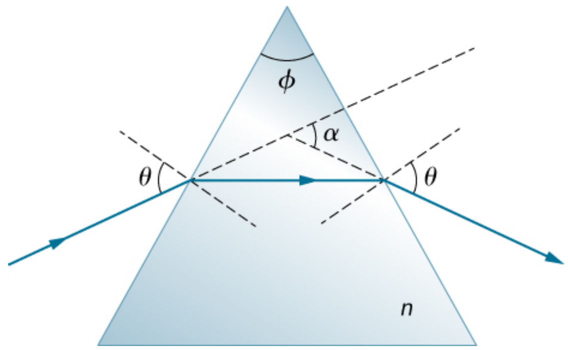
91. Consider sunlight entering Earth's atmosphere at sunrise and sunset—that is, at a 90.0°. incident angle. Taking the boundary between nearly empty space and the atmosphere to be sudden, calculate the angle of refraction for sunlight. This lengthens the time the Sun appears to be above the horizon, 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 phi and whose index of refraction is n. The angle of incidence of the ray relative to the normal to the left face is theta. 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 theta 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.

(a) What fraction of the incident light passes through P_1?

(b) What fraction of the incident light is passed by the combination?

(c) 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 $cos^2\theta + sin^2\theta = 1$.)

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

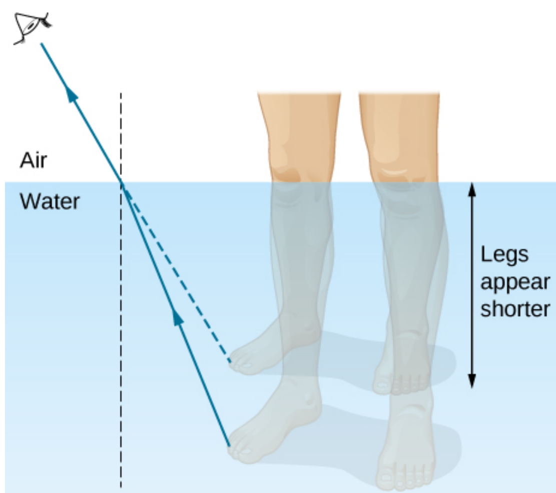
1. model as a ray when devices are large compared to wavelength, as a wave when devices are comparable or small compared to wavelength

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. $2.99705 \times 10^8 m/s$; $1.97 \times 10^8 m/s$

29. ice at $0^\circ C$

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. $0.250 I_0$

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. $1.69 \times 10^{-2}{}^{\circ}C/s$;

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

CHAPTER OVERVIEW

14: Introduction to Semiconductor Devices

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 essentially 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.

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.

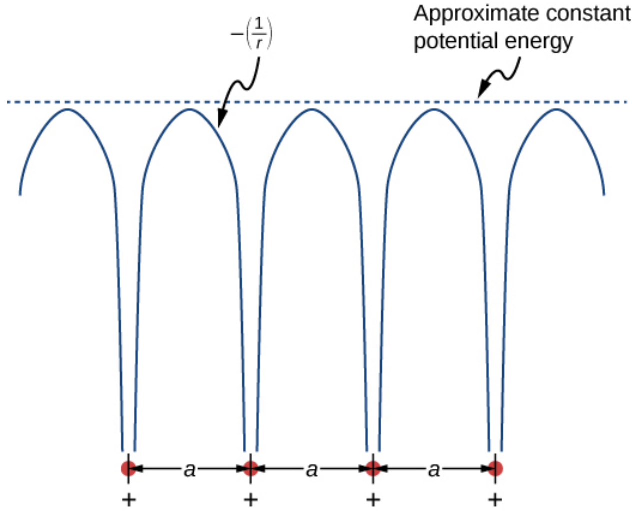


Figure 14.2.1: The periodic potential used to model electrons in a conductor. Each ion in the solid is the source of a Coulomb potential. Notice that the free electron model is productive because the average of this field is approximately constant.

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 1**s** 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 consider to be continuous or “in the continuum.”)

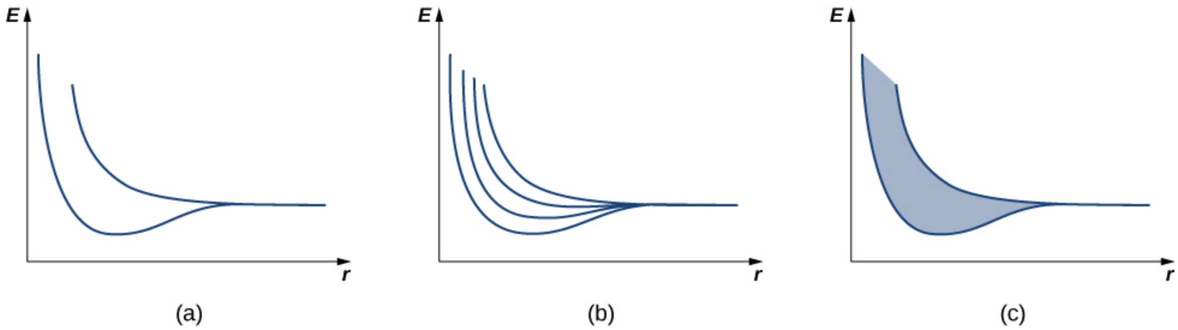


Figure 14.2.2: The dependence of energy-level splitting on the average distance between (a) two atoms, (b) four atoms, and (c) a large number of atoms. For a large number of electrons, a continuous band of energies is produced

Energy bands differ in the number of electrons they hold. In the 1**s** and 2**s** energy bands, each energy level holds up to two electrons (spin up and spin down), so this band has a maximum occupancy of 2**N** electrons. In the 2**p** energy band, each energy level holds up to six electrons, so this band has a maximum occupancy of 6**N** electrons (Figure $14.2.3$).

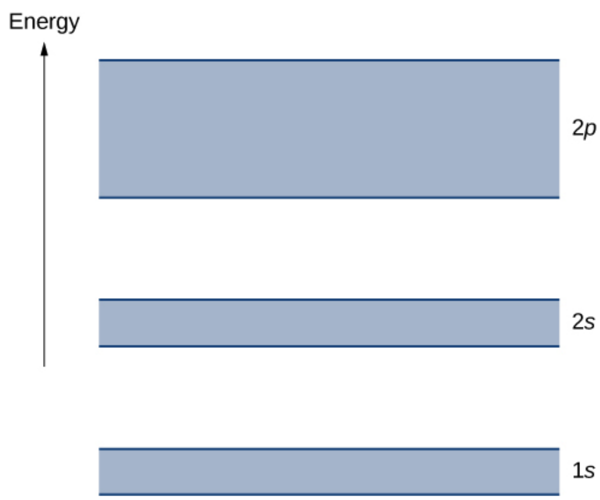


Figure 14.2.3: A simple representation of the energy structure of a solid. Electrons belong to energy bands separated by energy gaps.

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.

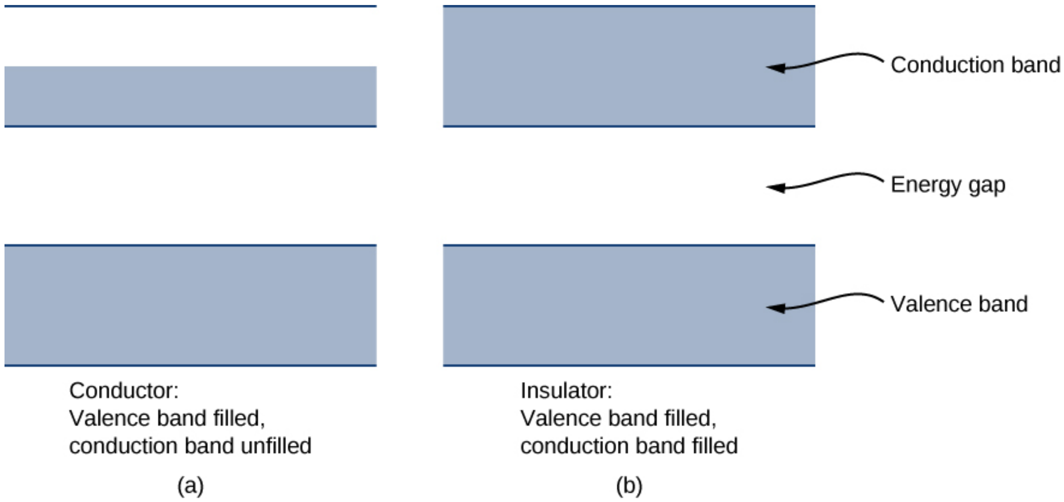


Figure 14.2.4: Comparison of a conductor and insulator. The highest energy band is partially filled in a conductor but completely filled in an insulator.

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 $T = 0\,K$, the semiconductor and insulator both have completely filled bands. The only difference is in the size of the energy gap (or **band gap**) $\mathbf{E_g}$ 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 KNote: 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

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, $n = p$.

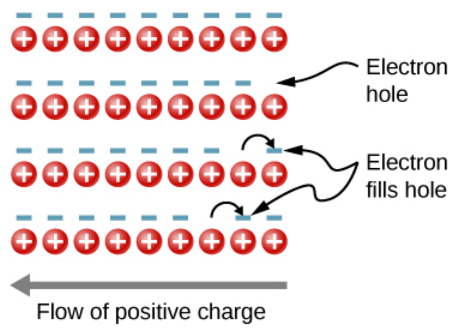


Figure 14.3.1: The motion of holes in a crystal lattice. As electrons shift to the right, an electron hole moves to the left.

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 $14.3.2a$).

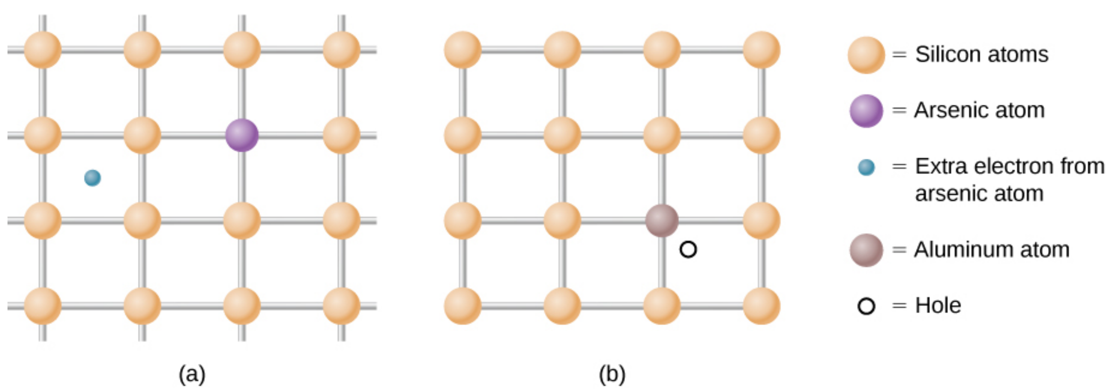


Figure 14.3.2: (a) A donor impurity and (b) an acceptor impurity. The introduction to impurities and acceptors into a semiconductor significantly changes the electronic properties of this material.

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.

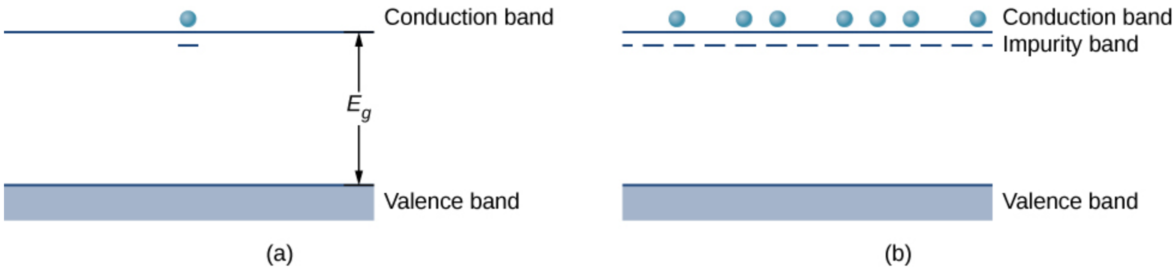


Figure 14.3.3: The extra electron from a donor impurity is excited into the conduction band; (b) formation of an impurity band in an n-type semiconductor.

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 $14.3.1b$

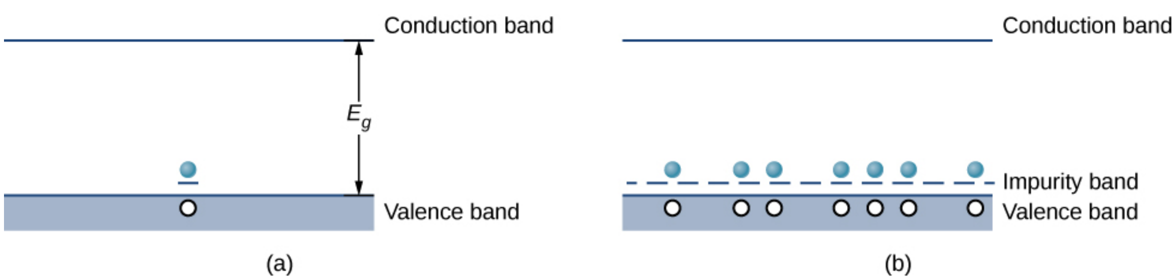


Figure 14.3.4: (a) An electron from the conduction band is excited into the empty state resulting from the acceptor impurity; (b) formation of an impurity band in a p-type semiconductor.

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 $14.3.5a$.

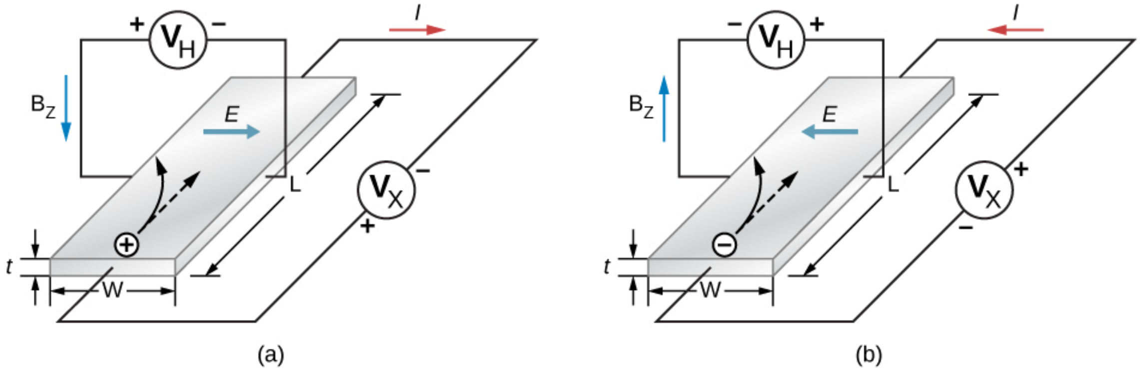


Figure 14.3.5: The Hall effect. (a) Positively charged electron holes are drawn to the left by a uniform magnetic field that points downward. An electric field is generated to the right. (b) Negative charged electrons are drawn to the left by a magnetic field that points up. An electric field is generated 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; **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 (**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

$$j = nqv, \tag{14.3.1}$$

where **q** is the magnitude of the charge, **n** is the number of charge carriers per unit volume, and **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 **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).

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 $14.4.2$.

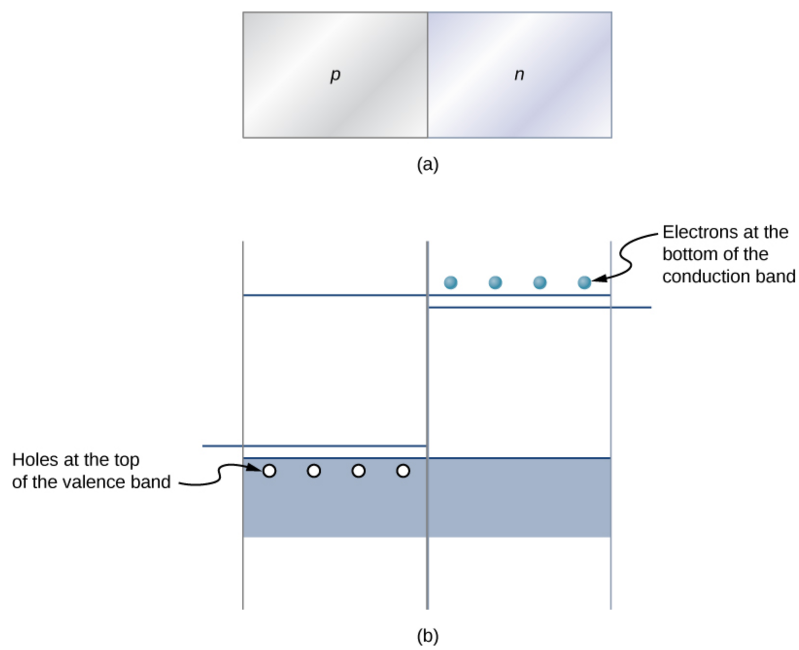


Figure 14.4.1: (a) Representation of a **p-n** junction. (b) A comparison of the energy bands of **p**-type and **n**-type silicon prior to equilibrium.

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.

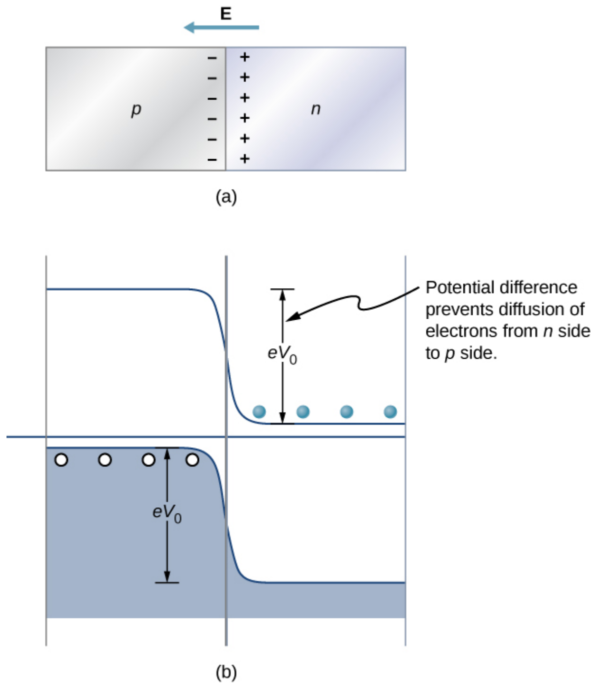


Figure 14.4.2: At equilibrium, (a) excess charge resides near the interface and the net current is zero, and (b) the potential energy difference for electrons (in light blue) prevents further diffusion of electrons into the **p**-side.

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_BT} . However, if a battery of voltage V_b is applied in the forward-bias configuration, this fraction improves to $Ne^{-e(V-V_b)/k_BT}$. The electric current due to the majority carriers from the **n**-side to the **p**-side is therefore

$$I = Ne^{-eV/k_BT}e^{eV_b/k_BT} = I_0e^{eV_b/k_BT},$$

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

$$I_{net} = I_0\left(e^{eV_b/k_BT} - 1\right).$$

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 $I_{net} \approx -I_0$. 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**.

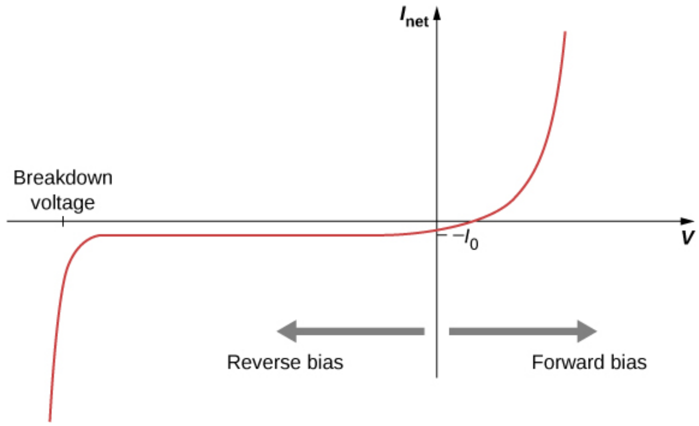


Figure 14.4.3: Current versus voltage across a **p-n** junction (diode). In the forward bias configuration, electric current flows easily. However, in the reverse bias configuration, electric current flow very little.

✔ 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 $4.5 \times 10^{-1} A$. The reverse saturation current is $2.2 \times 10^{-8} A$. (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

$$I_{net} = I_0 \left(e^{eV_b/k_BT} - 1\right).$$

The current with no bias is related to the reverse saturation current by

$$I_0 \approx -I_{sat} = 2.2 \times 10^{-8}.$$

Therefore

$$\frac{I_{net}}{I_0} = \frac{4.5 \times 10^{-1} A}{2.2 \times 10^{-8} A} = 2.0 \times 10^{8}.$$

this can be written as

$$\frac{I_{net}}{I_0} + 1 = e^{eV_b/k_BT}.$$

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

$$\frac{eV_b}{k_BT} = 19.$$

The temperature is therefore

$$T = \frac{eV_b}{k_B}\left(\frac{1}{19}\right) = \frac{e(0.12\,V)}{8.617 \times 10^{-5} eV/K}\left(\frac{1}{19}\right) = 73\,K.$$

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_BT, 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

$$I_c = \beta I_B.$$

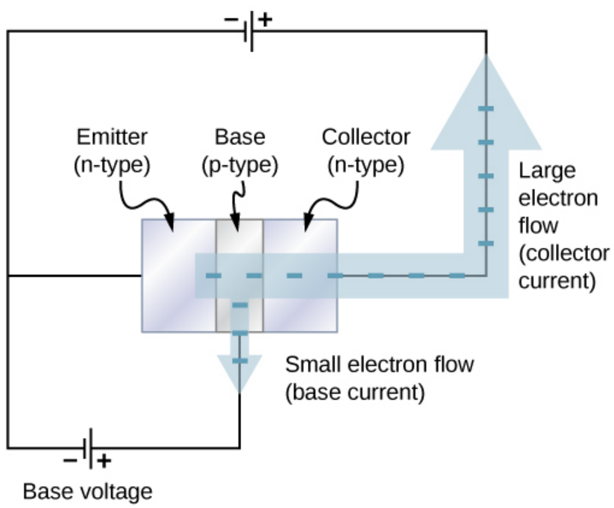


Figure 14.4.4: A junction transistor has three parts: emitter, base, and collector. Voltage applied to the base acts as a valve to control electric current from the emitter to the collector.

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**.

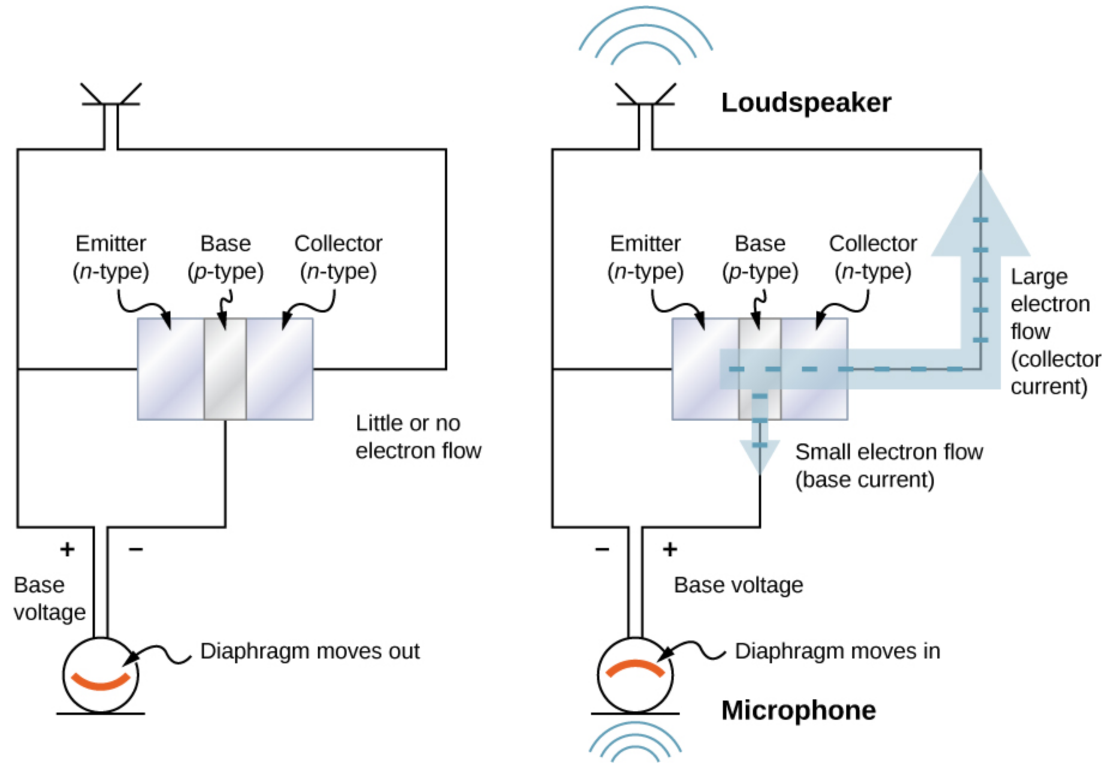


Figure 14.4.5: An audio amplifier based on a junction transistor. Voltage applied to the base by a microphone acts as a valve to control a larger electric current that passes through a loudspeaker.

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.

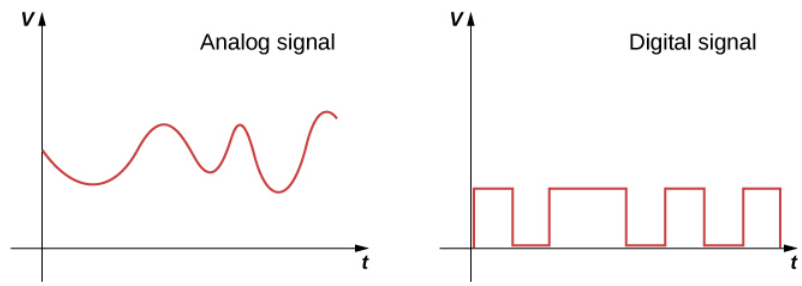
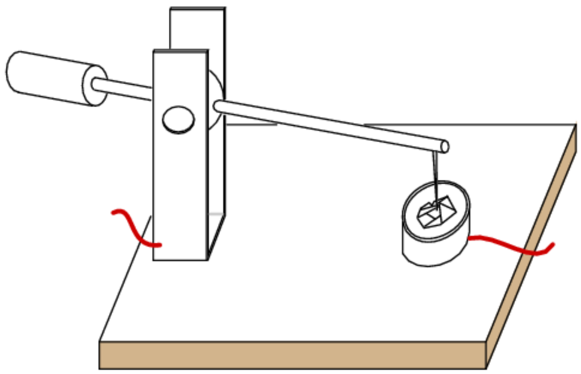


Figure 14.4.6: Real-world data are often analog, meaning data can vary continuously. Intensity values of sound or visual images are usually analog. These data are converted into digital signals for electronic processing in recording devices or computers. The digital signal is generated from the analog signal by requiring certain voltage cut-off value.

14.5: Junction Diodes

Selenium oxide rectifiers were used before modern power diode rectifiers became available. These and the Cu_2O rectifiers were polycrystalline devices. Photoelectric cells were once made from Selenium.

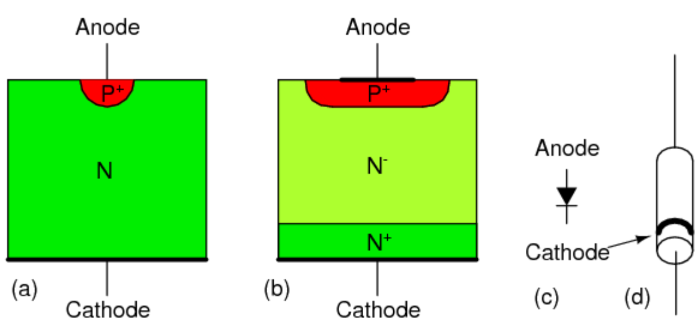
Before the modern semiconductor era, an early diode application was as a radio frequency *detector,* which recovered audio from a radio signal. The “semiconductor” was a polycrystalline piece of the mineral galena, lead sulfide, PbS. A pointed metallic wire known as a *cat whisker* was brought in contact 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.

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 posses **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 $14.6.1$.

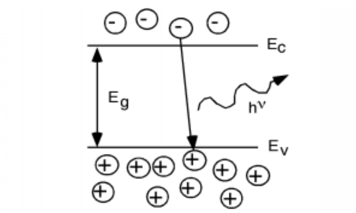


Figure $14.6.1$: Radiative recombination in a direct band-gap semiconductor

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 $4.14 \times 10^{-15}\ \text{eV} \cdot \text{seconds}$.

$$E = h\nu$$

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 $2\ \text{eV}$ band-gap should give off light at about $620\ \text{nm}$ (in the range of red light). A $3\ \text{eV}$ band-gap material would emit at $414\ \text{nm}$, 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.

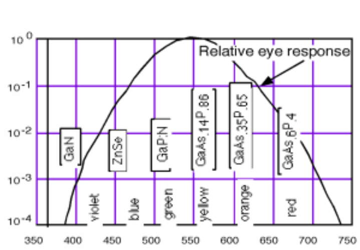


Figure $14.6.2$ Relative response of the human eye to various colors

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 \mathrm{GaAs}\) and make a mixed compound semiconductor $\mathrm{GaAs}_{1-x}\mathrm{P}_x$. 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 $14.6.4$.

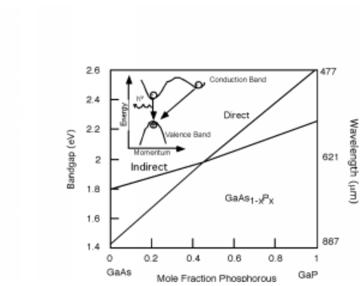


Figure $14.6.3$ Band gap for the GaAsP system

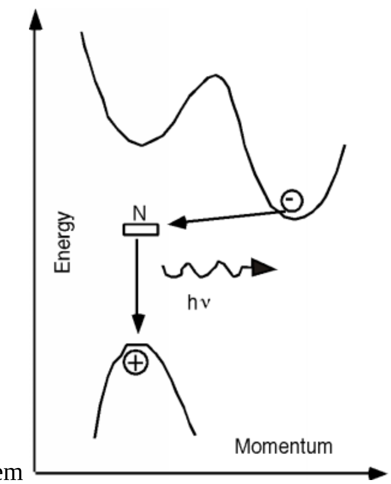


Figure $14.6.4$ Addition of a nitrogen recombination center to indirect GaAsP

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.

14.7: Solar Cells

Now let us look at the opposite process of light generation for a moment. Consider the following situation.

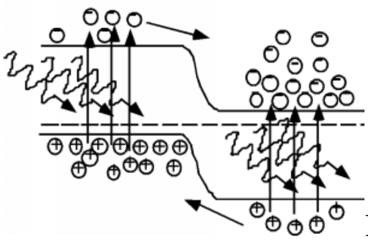


Figure 14.7.1: P-N diode under illumination

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.

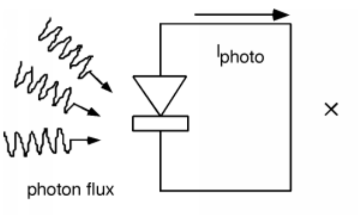


Figure 14.7.2 Schematic representation of a photovoltaic cell

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?

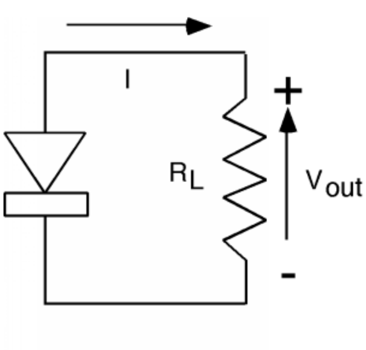


Figure 14.7.3 Photovoltaic cell with a load resistor

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.

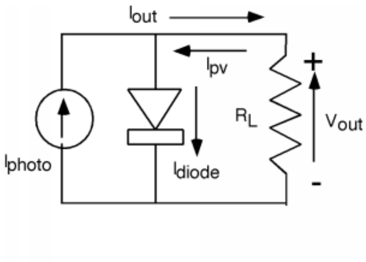


Figure 14.7.4 Model of PV cell

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!

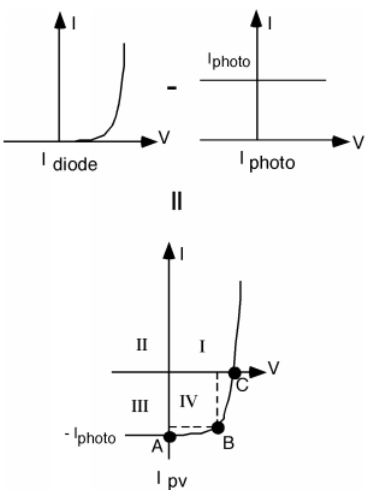


Figure 14.7.5 Combining the diode and the current source

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 our of out 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.

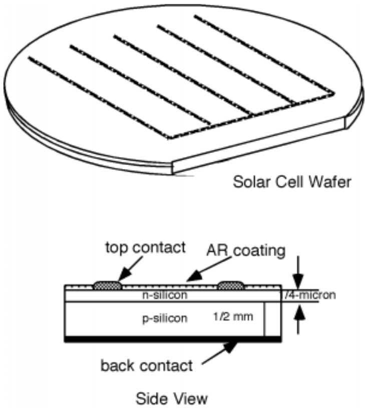


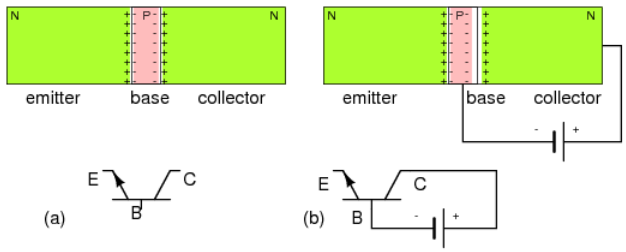
Figure $14.7.6$: A real solar cell

The solar power flux on the earth's surface is (conveniently) about $1\ \frac{\mathrm{kW}}{\mathrm{m}^2}$, or $100\ \frac{\mathrm{mW}}{\mathrm{cm}^2}$. So if we made a solar cell from a 4-inch-diameter wafer (typical) it would have an area of about $81\ \mathrm{cm}^2$ and so would be receiving a flux of about $8.1\ \mathrm{Watts}$. Typical cell efficiencies run from about $10\ \%$ to maybe $15\ \%$ unless special (and costly) tricks are used. This means that we will get about $1.2\ \mathrm{Watts}$ 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.

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.



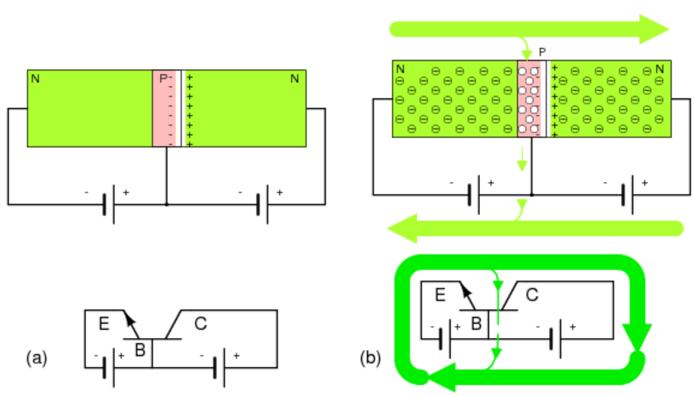
(a) NPN junction bipolar transistor. (b) Apply reverse bias to collector base junction.

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.

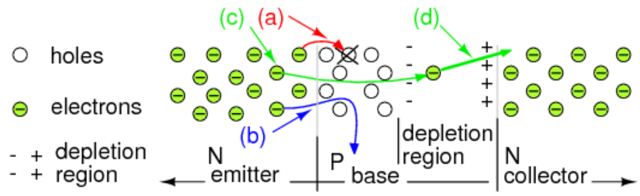
If the base region were thick, as in a pair of back-to-back diodes, all the current entering the base would flow out the base lead. In our NPN transistor example, electrons leaving the emitter for the base would combine with holes in the base, making room for more holes to be created at the (+) battery 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.

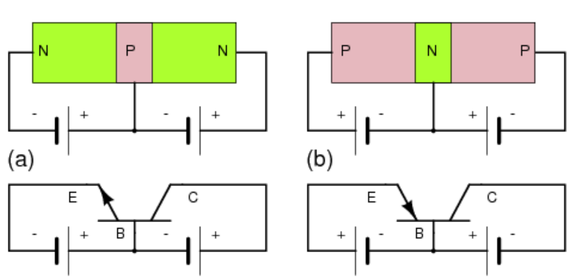
Majority carriers within the N-type emitter are electrons, becoming minority carriers when entering the P-type base. These electrons face four possible fates entering the thin P-type base. A few at Figure above(a) fall into holes in the base that contribute to base current flow to the (+) battery terminal. 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.

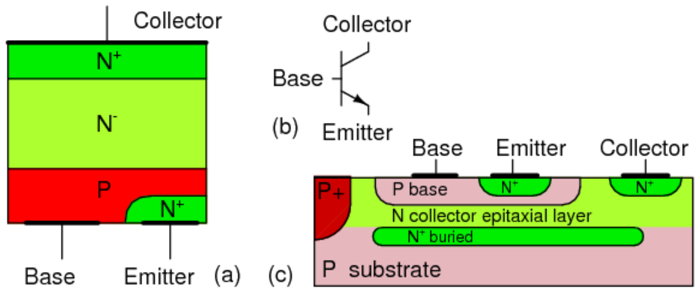
The thin base and the heavily doped emitter help keep the *emitter efficiency* high, 99% for example. This corresponds to 100% emitter current splitting between the base as 1% and the collector as 99%. The emitter efficiency is known as $\alpha = I_C/I_E$.

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 a good as the NPN variety within the same integrated circuit die. Thus, integrated circuits use the NPN variety as much as possible.

Review

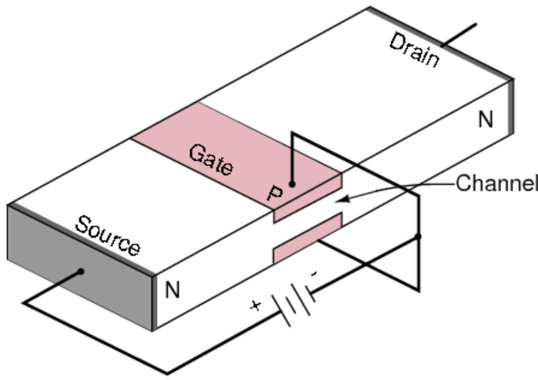
- Bipolar transistors conduct current using both electrons and holes in the same device.
- Operation of a bipolar transistor as a current amplifier requires that the collector-base junction be reverse biased and the emitter-base junction be forward biased.
- A transistor differs from a pair of back to 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.

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.

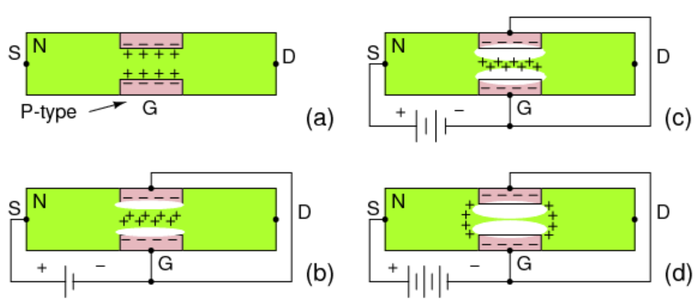
"Cleanliness is next to godliness" applies to the manufacture of field effect transistors. Though it is possible to make bipolar transistors outside of a *clean room*, it is a necessity for field effect transistors. Even in such an environment, manufacture is tricky because of contamination control issues. 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.

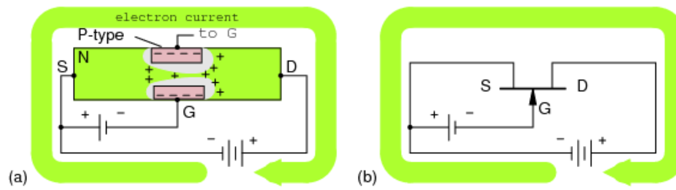
Figure below(a) shows the depletion region at the gate junction. This is due to diffusion of holes from the P-type gate region into the N-type channel, giving the charge separation about the junction, with a non-conductive depletion region at the junction. The depletion region extends more deeply into 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.

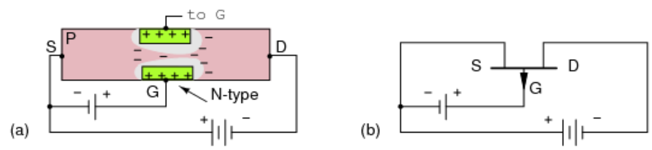
The source and drain are interchangeable, and the source to drain current may flow in either direction for low level drain battery voltage (< 0.6 V). That is, the drain battery may be replaced by a low voltage AC source. For a high drain power supply voltage, to 10's of volts for small signal devices, 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.

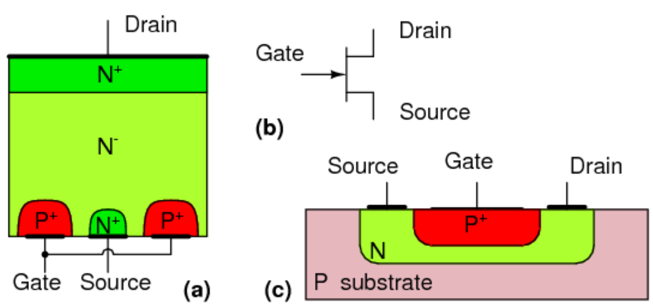
P-channel field effect transistors are also available. The channel is made of P-type material. The gate is a heavily dopped N-type region. All the voltage sources are reversed in the P-channel circuit (Figure below) as compared with the more popular N-channel device. Also note, the arrow points out of 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.

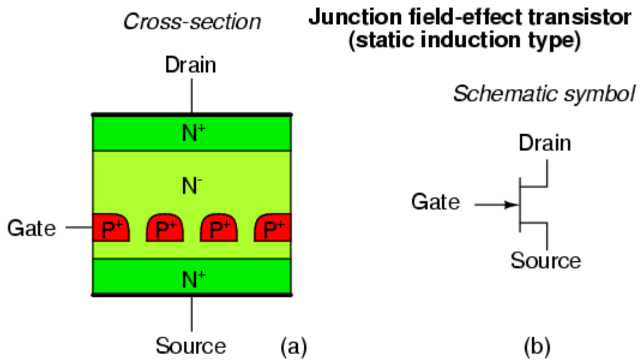
As the positive gate bias voltage is increased, the resistance of the P-channel increases, decreasing the current flow in the drain circuit.

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.



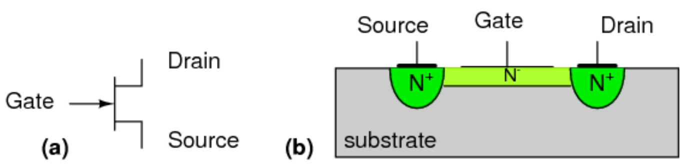
Junction field effect transistor: (a) Discrete device cross-section, (b) schematic symbol, (c) integrated circuit device cross-section.

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.

CHAPTER OVERVIEW

15: Part 2 - Detailed and/or Advanced Content

CHAPTER OVERVIEW

16: Direct Calculation of Electrical Quantities from Charge Distributions

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.

16.2: 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.

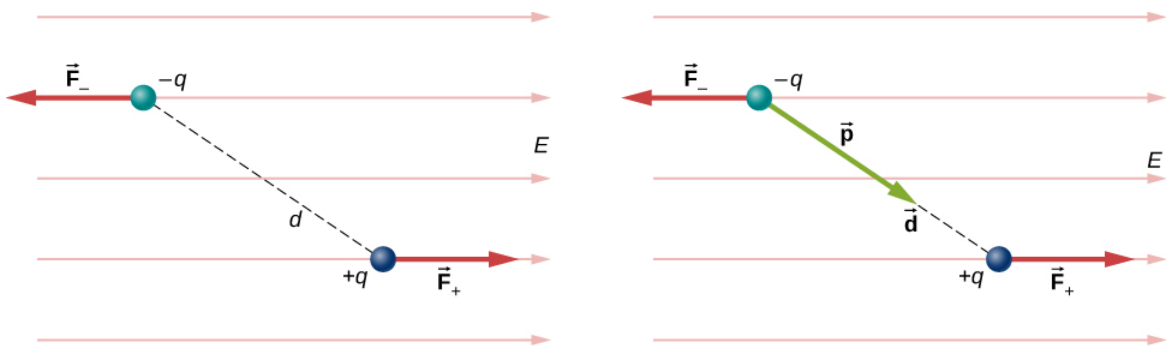


Figure 16.2.1: A dipole in an external electric field. (a) The net force on the dipole is zero, but the net torque is not. As a result, the dipole rotates, becoming aligned with the external field. (b) The dipole moment is a convenient way to characterize this effect. The $\vec{d}$ points in the same 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:

$$\vec{\tau} = \left(\frac{\vec{d}}{2} \times \vec{F}_+\right) + \left(-\frac{\vec{d}}{2} \times \vec{F}_-\right) \tag{16.2.1}$$

$$= \left[\left(\frac{\vec{d}}{2}\right) \times \left(+q\vec{E}\right) + \left(-\frac{\vec{d}}{2}\right) \times \left(-q\vec{E}\right)\right] \tag{16.2.2}$$

$$= q\vec{d} \times \vec{E}. \tag{16.2.3}$$

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:

$$\vec{p} \equiv q\vec{d}. \tag{16.2.4}$$

We can therefore write

$$\vec{\tau} = \vec{p} \times \vec{E}. \tag{16.2.5}$$

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 $16.2.2$).

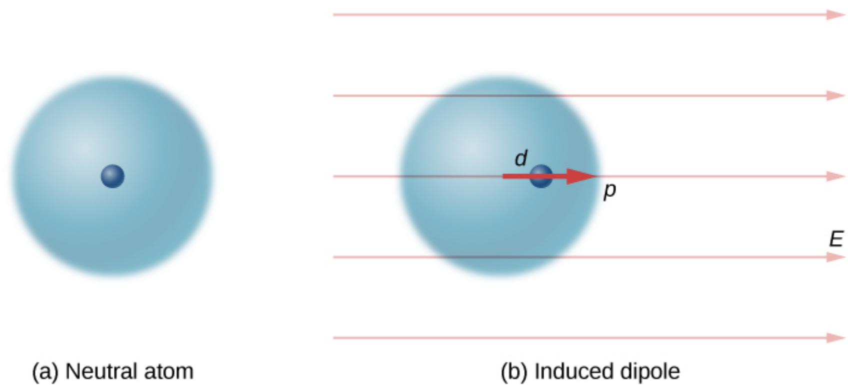


Figure $16.2.2$: A dipole is induced in a neutral atom by an external electric field. The induced dipole moment is aligned with the external field.

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

$$\vec{E}_{total} = \vec{E}_{external} + \vec{E}_{dipole} \tag{16.2.6}$$

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.

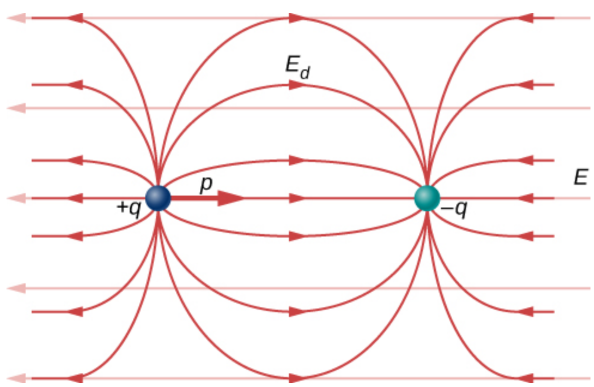


Figure $16.2.3$: The net electric field is the vector sum of the field of the dipole plus the external field.

Recall that we found the electric field of a dipole. If we rewrite it in terms of the dipole moment we get:

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{\vec{p}}{z^3}. \tag{16.2.7}$$

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 cm); (b) (0, 0, –5.0 cm); (c) (3.0 cm, 0, 2.0 cm).

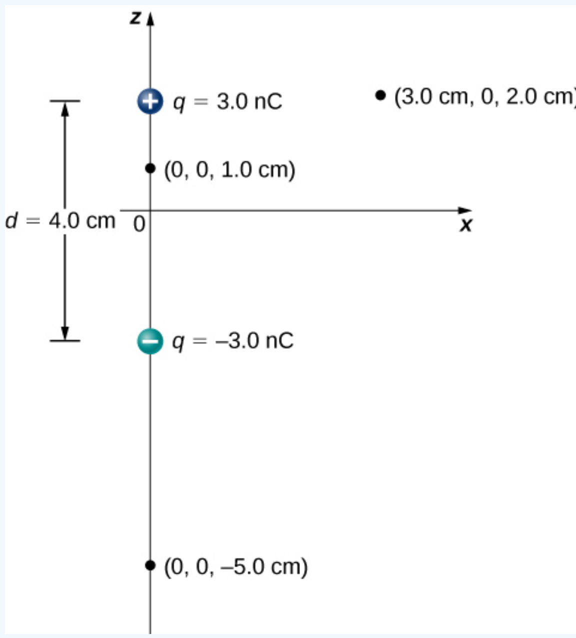


Figure 16.2.3: A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point P in space.

Strategy

Apply $V_p = k\sum_1^N \frac{q_i}{r_i}$ to each of these three points.

Solution

a. $V_p = k\sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9\,\text{N}\cdot\text{m}^2/\text{C}^2)\left(\frac{3.0\,\text{nC}}{0.010\,\text{m}} - \frac{3.0\,\text{nC}}{0.030\,\text{m}}\right) = 1.8 \times 10^3\,\text{V}$

b. $V_p = k\sum_1^N \frac{q_i}{r_i} = (9.0 \times 10^9\,\text{N}\cdot\text{m}^2/\text{C}^2)\left(\frac{3.0\,\text{nC}}{0.070\,\text{m}} - \frac{3.0\,\text{nC}}{0.030\,\text{m}}\right) = -5.1 \times 10^2\,\text{V}$

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 $z >> d$, 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, $r >> d$; 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

$$V_p = V_+ + V_- = k\left(\frac{q}{r_+} - \frac{q}{r_-}\right) \tag{16.2.8}$$

where

$$r_\pm = \sqrt{x^2 + \left(z \pm \frac{d}{2}\right)^2}. \tag{16.2.9}$$

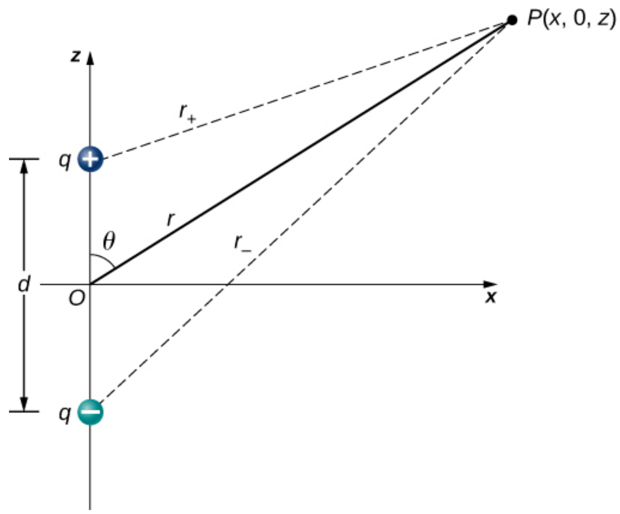


Figure 16.2.4: A general diagram of an electric dipole, and the notation for the distances from the individual charges to a point P in space.

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

$$r_\pm = \sqrt{r^2\sin^2\theta + \left(r\cos\theta \pm \frac{d}{2}\right)^2}. \tag{16.2.10}$$

We can simplify this expression by pulling r out of the root,

$$r_\pm = \sqrt{\sin^2\,\theta + \left(r\,\cos\,\theta \pm \frac{d}{2}\right)^2} \tag{16.2.11}$$

and then multiplying out the parentheses

$$r_\pm = r\sqrt{\sin^2\theta + \cos^2\theta \pm \cos\theta\frac{d}{r} + \left(\frac{d}{2r}\right)^2} = r\sqrt{1 \pm \cos\theta\frac{d}{r} + \left(\frac{d}{2r}\right)^2}. \tag{16.2.12}$$

The last term in the root is small enough to be negligible (remember $r >> d$, and hence $(d/r)^2$ is extremely small, effectively zero to the level we will probably be measuring), leaving us with

$$r_\pm = r\sqrt{1 \pm \cos\theta\frac{d}{r}}. \tag{16.2.13}$$

Using the **binomial approximation** (a standard result from the mathematics of series, when a is small)

$$\frac{1}{\sqrt{1 \pm a}} \approx 1 \pm \frac{a}{2} \tag{16.2.14}$$

and substituting this into our formula for V_p, we get

$$V_p = k\left[\frac{q}{r}\left(1+\frac{d\cos\theta}{2r}\right) - \frac{q}{r}\left(1-\frac{d\cos\theta}{2r}\right)\right] = k\frac{qd\cos\theta}{r^2}. \tag{16.2.15}$$

This may be written more conveniently if we define a new quantity, the **electric dipole moment**,

$$\vec{p} = q\vec{d}, \tag{16.2.16}$$

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

$$V_p = k\frac{\vec{p}\cdot\hat{r}}{r^2}. \tag{16.2.17}$$

A diagram of the application of this formula is shown in Figure $16.2.5$.

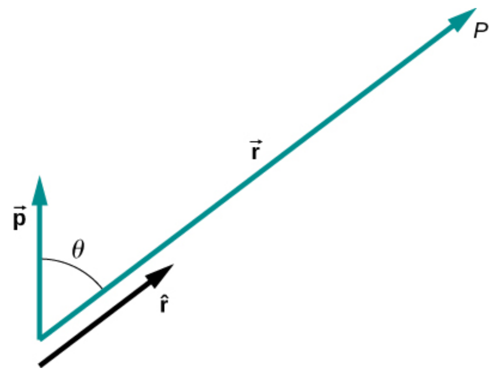


Figure $16.2.5$: The geometry for the application of the potential of a dipole.

There are also higher-order moments for quadrupoles, octupoles, and so on.

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 H_2O 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 $16.3.1$.

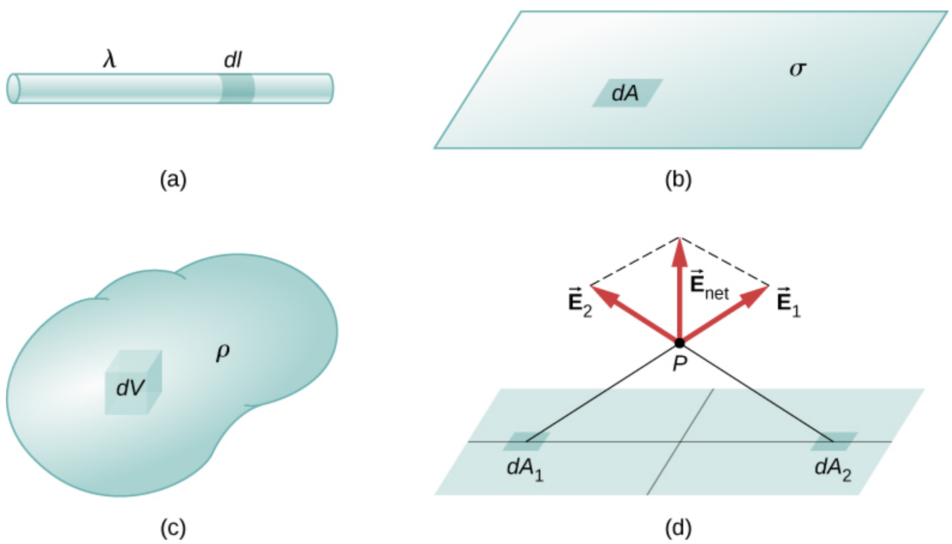


Figure 16.3.1: The configuration of charge differential elements for a (a) line charge, (b) sheet of charge, and (c) a volume of charge. Also note that (d) some of the components of the total electric field cancel out, with the remainder resulting in a net electric field.

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 $(\mathrm{C/m})$
- **surface charge density:** $\sigma \equiv$ charge per unit area (Figure $16.3.1b$); units are coulombs per square meter $(\mathrm{C/m^2})$
- **volume charge density:** $\rho \equiv$ charge per unit volume (Figure $16.3.1c$); units are coulombs per square meter $(\mathrm{C/m^3})$

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 $dq = \lambda\, dl$, $\sigma\, dA$, or $\rho\, dV$, respectively:

$$\underbrace{\vec{E}(P)=\frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\left(\frac{q_i}{r^2}\right)\hat{r}}_{\text{Point charges}} \tag{16.3.1}$$

$$\underbrace{\vec{E}(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{line}}\left(\frac{\lambda\,dl}{r^2}\right)\hat{r}}_{\text{Line charge}} \tag{16.3.2}$$

$$\underbrace{\vec{E}(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{surface}}\left(\frac{\sigma\,dA}{r^2}\right)\hat{r}}_{\text{Surface charge}} \tag{16.3.3}$$

$$\underbrace{\vec{E}(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{volume}}\left(\frac{\rho\,dV}{r^2}\right)\hat{r}}_{\text{Volume charge}} \tag{16.3.4}$$

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 **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, $\lambda\,dl$, $\sigma\,dA$, $\rho\,dV$) 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

$$E_x(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{line}}\left(\frac{\lambda\,dl}{r^2}\right)_x, \tag{16.3.5}$$

$$E_y(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{line}}\left(\frac{\lambda\,dl}{r^2}\right)_y, \tag{16.3.6}$$

$$E_z(P)=\frac{1}{4\pi\epsilon_0}\int_{\text{line}}\left(\frac{\lambda\,dl}{r^2}\right)_z \tag{16.3.7}$$

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.

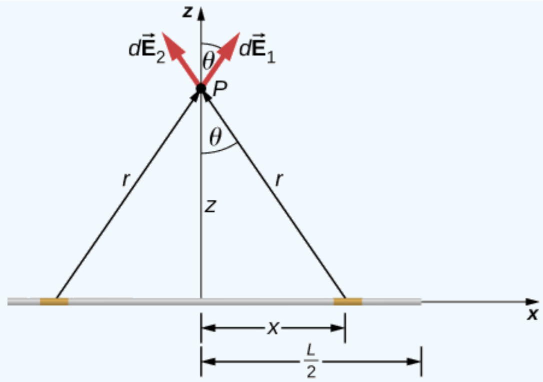


Figure 16.3.2: A uniformly charged segment of wire. The electric field at point P can be found by applying the superposition principle to symmetrically placed charge elements and integrating.

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

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{\lambda L}{z\sqrt{z^2+\frac{L^2}{4}}}\,\hat{k}. \tag{16.3.8}$$

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

$$\vec{E} \approx \frac{1}{4\pi\epsilon_0}\frac{\lambda L}{z^2}\hat{k}. \tag{16.3.9}$$

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 $L \to \infty$ 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:

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\frac{2\lambda}{z}\hat{k}. \tag{16.3.10}$$

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.

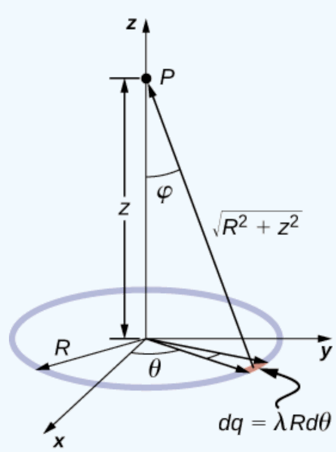


Figure 16.3.3: The system and variable for calculating the electric field due to a ring of charge.

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 $\cos\phi = \dfrac{z}{\sqrt{z^2+R^2}}$ and therefore the electric field is

$$\begin{aligned}\vec{E}(P) &= \frac{1}{4\pi\epsilon_0}\int_{\text{line}} \frac{\lambda dl}{r^2}\hat{r} = \frac{1}{4\pi\epsilon_0}\int_0^{2\pi} \frac{\lambda R d\theta}{z^2+R^2}\frac{z}{\sqrt{z^2+R^2}}\hat{z} \\ &= \frac{1}{4\pi\epsilon_0}\frac{\lambda R z}{(z^2+R^2)^{3/2}}\hat{z}\int_0^{2\pi} d\theta \\ &= \frac{1}{4\pi\epsilon_0}\frac{2\pi\lambda R z}{(z^2+R^2)^{3/2}}\hat{z} \\ &= \frac{1}{4\pi\epsilon_0}\frac{q_{tot} z}{(z^2+R^2)^{3/2}}\hat{z}.\end{aligned}$$

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

$$\vec{E} \approx \frac{1}{4\pi\epsilon_0}\frac{q_{tot}}{z^2}\hat{z},$$

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)

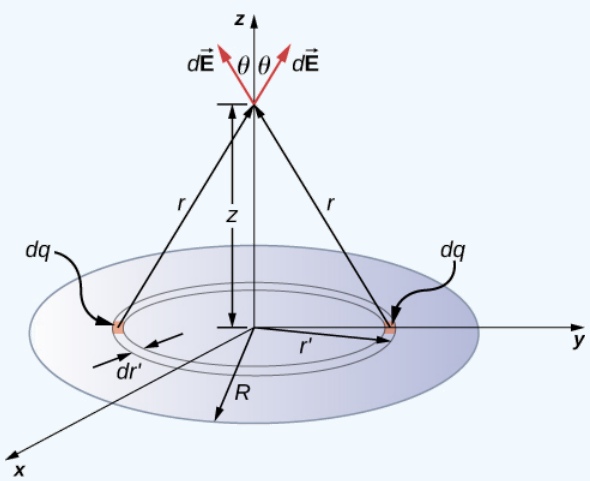


Figure 16.3.4: A uniformly charged disk. As in the line charge example, the field above the center of this disk can be calculated by taking advantage of the symmetry of the charge distribution.

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,

$$dA = 2\pi r'\,dr' \tag{16.3.11}$$

$$r^2 = r'^2 + z^2 \tag{16.3.12}$$

$$\cos\theta = \frac{z}{(r'^2+z^2)^{1/2}}. \tag{16.3.13}$$

(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,

$$\vec{E}(z) = \frac{1}{4\pi\epsilon_0}\left(2\pi\sigma - \frac{2\pi\sigma z}{\sqrt{R^2+z^2}}\right)\hat{k}. \tag{16.3.14}$$

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 $R \to \infty$, Equation 16.3.14reduces 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:

$$\vec{E} = \lim_{R\to\infty}\frac{1}{4\pi\epsilon_0}\left(2\pi\sigma - \frac{2\pi\sigma z}{\sqrt{R^2+z^2}}\right)\hat{k} \tag{16.3.15}$$

$$= \frac{\sigma}{2\epsilon_0}\hat{k}. \tag{16.3.16}$$

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 $16.3.16$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 $16.3.5$).

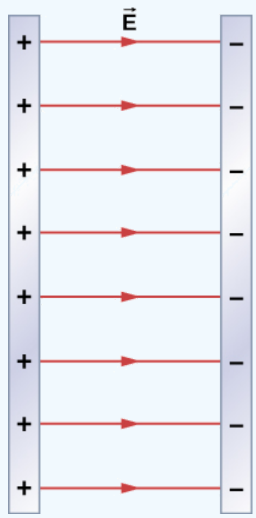


Figure $16.3.5$: Two charged infinite planes. Note the direction of the electric field.

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.

- **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,

$$V_p = k\sum \frac{q_i}{r_i}. \tag{16.4.1}$$

We may treat a continuous charge distribution as a collection of infinitesimally separated individual points. This yields the integral

$$V_p = \int \frac{dq}{r} \tag{16.4.2}$$

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

$$\underbrace{dq = \lambda\, dl}_{\text{one dimension}} \tag{16.4.3}$$

$$\underbrace{dq = \sigma\, dA}_{\text{two dimensions}} \tag{16.4.4}$$

$$\underbrace{dq = \rho\, dV}_{\text{three dimensions}} \tag{16.4.5}$$

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 $16.4.1$.

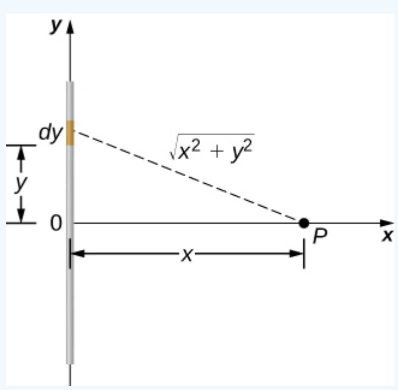


Figure 16.4.1: We want to calculate the electric potential due to a line of charge.

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.

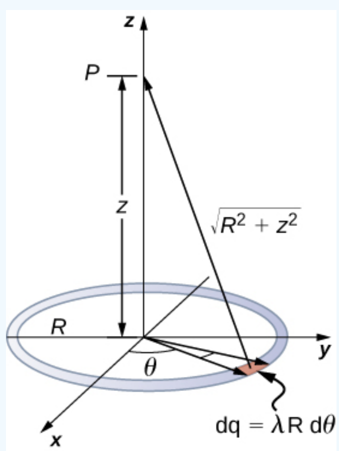


Figure 16.4.2: We want to calculate the electric potential due to a ring of charge.

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 $\lambda Rd\theta$. 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 $16.4.3$.

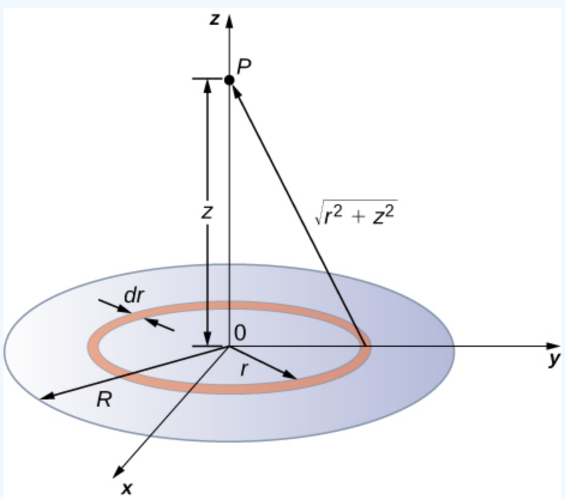


Figure 16.4.3: We want to calculate the electric potential due to a disk of charge.

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

$$dV_p = k\frac{dq}{\sqrt{z^2+r^2}} \tag{16.4.6}$$

where

$$dq = \sigma 2\pi r dr. \tag{16.4.7}$$

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 $r=R$:

$$\begin{aligned} V_p &= \int dV_p = k2\pi\sigma \int_0^R \frac{r\,dr}{\sqrt{z^2+r^2}}, \\ &= k2\pi\sigma(\sqrt{z^2+R^2} - \sqrt{z^2}). \end{aligned}$$

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 $L \to \infty$ in our previous result will work:

$$V_p = \lim_{L\to\infty} k\lambda \ln\left(\frac{L+\sqrt{L^2+4x^2}}{-L+\sqrt{L^2+4x^2}}\right). \tag{16.4.8}$$

However, this limit does not exist because the argument of the logarithm becomes $[2/0]$ as $L \to \infty$, 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 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

$$V_p = -\int_R^p \vec{E}\cdot d\vec{l} \tag{16.4.9}$$

where R is a finite distance from the line of charge, as shown in Figure 16.4.4.

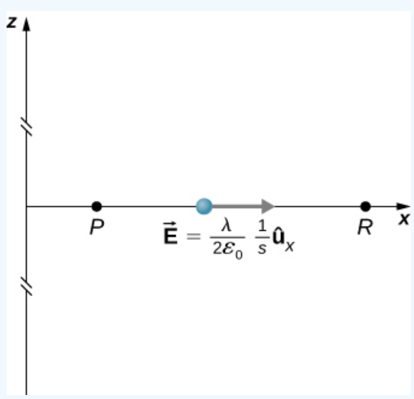


Figure 16.4.4: Points of interest for calculating the potential of an infinite line of charge.

With this setup, we use $\vec{E}_p = 2k\lambda\frac{1}{s}\hat{s}$ and $d\vec{l} = d\vec{s}$ to obtain

$$\begin{aligned} V_p - V_R &= -\int_R^p 2k\lambda\frac{1}{s}ds \\ &= -2k\lambda\ln\frac{s_p}{s_R}. \end{aligned}$$

Now, if we define the reference potential $V_R = 0$ at $s_R = 1\text{ m}$, this simplifies to

$$V_p = -2k\lambda \ln s_p. \tag{16.4.10}$$

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.

16.5: 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	$\overrightarrow{F_{12}}(r)=\frac{1}{4\pi\varepsilon_0}\frac{q_1q_2}{r_{12}^2}\hat{r_{12}}$
Superposition of electric forces	$\vec{F}(r)=\frac{1}{4\pi\varepsilon_0}Q\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i$
Electric force due to an electric field	$\vec{F}=Q\vec{E}$
Electric field at point P	$\vec{E}(P)\equiv\frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{q_i}{r_i^2}\hat{r}_i$
Field of an infinite wire	$\vec{E}(z)=\frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\hat{k}$
Field of an infinite plane	$\vec{E}=\frac{\sigma}{2\varepsilon_0}\hat{k}$
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).

16.6: 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.**

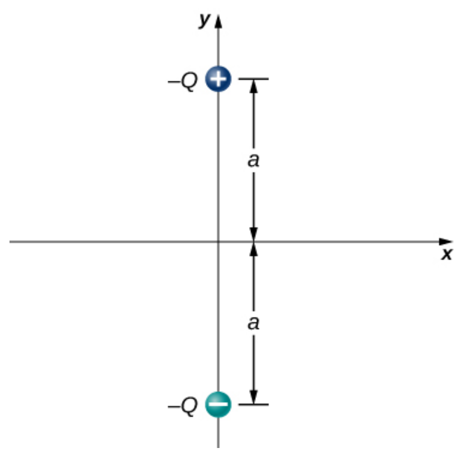
15. In what region of space is the potential due to a uniformly charged sphere the same as that of a point charge? In what region does it differ from that of a point charge?

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 $|x| \gg a, E \approx Qa/2\pi\varepsilon_0 x^3$. (b) Show that at all points on the **y**-axis for which $|y| \gg a, E \approx Qa/\pi\varepsilon_0 y^3$.



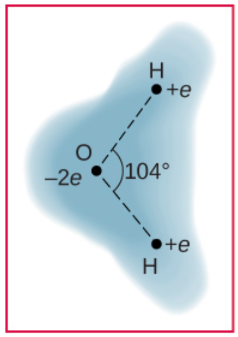
106. (a) What is the dipole moment of the configuration shown above? If **Q=4.0μC**,

(b) what is the torque on this dipole with an electric field of $4.0 \times 10^5 N/C\hat{i}$?

(c) What is the torque on this dipole with an electric field of $-4.0 \times 10^5 N/C\hat{i}$?

(d) What is the torque on this dipole with an electric field of $\pm 4.0 \times 10^5 N/C\hat{j}$?

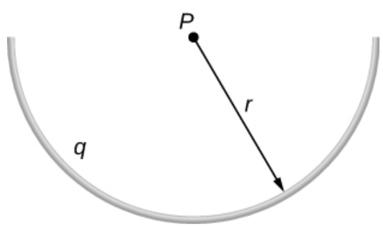
107. A water molecule consists of two hydrogen atoms bonded with one oxygen atom. The bond angle between the two hydrogen atoms is **104°** (see below). Calculate the net dipole moment of a hypothetical water molecule where the charge at the oxygen molecule is −2e and at each hydrogen atom is +e. The 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.



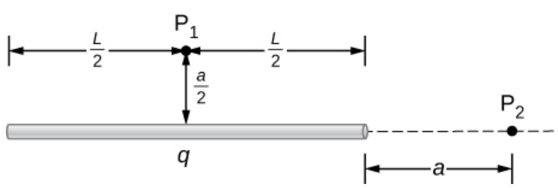
Calculating Electric Fields of Charge Distributions

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.)

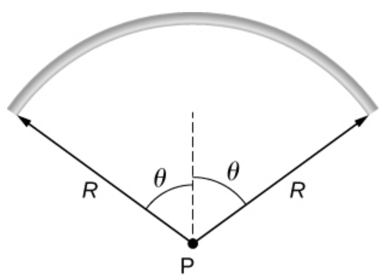
84. The charge per unit length on the thin semicircular wire shown below is **λ**. What is the electric field at the point **P**?



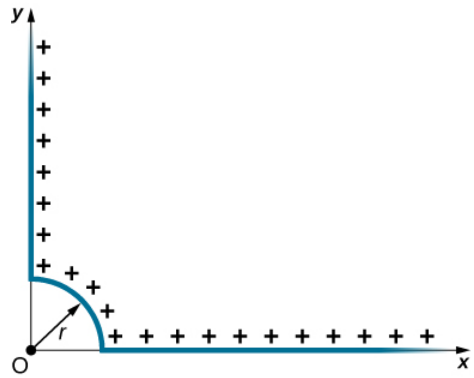
87. A total charge **q** is distributed uniformly along a thin, straight rod of length **L** (see below). What is the electric field at P_1? At P_2?



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**?

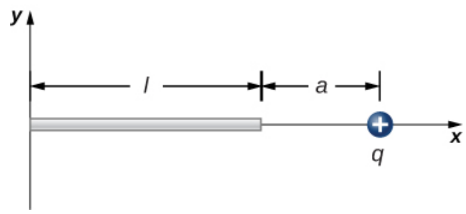


121. Charge is distributed uniformly along the entire **y**-axis with a density y_λ and along the positive **x**-axis from $x = a$ to $x = b$ with a density λ_x. What is the force between the two distributions?

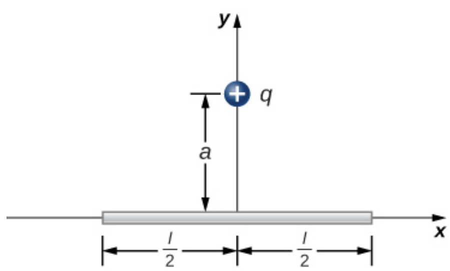
122. The circular arc shown below carries a charge per unit length $\lambda = \lambda_0 cos\theta$, where θ is measured from the **x**-axis. What is the electric field at the origin?

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.

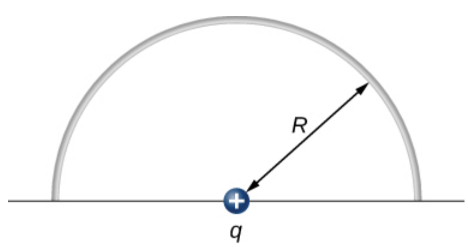
124. The charge per unit length on the thin rod shown below is λ. What is the electric force on the point charge **q**? Solve this problem by first considering the electric force $d\vec{F}$ on **q** due to a small segment dx of the rod, which contains charge λdx. Then, find the net force by integrating $d\vec{F}$ 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.)



- **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. $E_x = 0, E_y = \frac{1}{4\pi\varepsilon_0}[\frac{2q}{(x^2+a^2})\frac{a}{\sqrt{(x^2+a^2)}} \Rightarrow x \gg a \Rightarrow \frac{1}{2\pi\varepsilon_0}\frac{qa}{x^3}$

$E_y = \frac{q}{4\pi\varepsilon_0}[\frac{2ya+2ya}{(y-a)^2(y+a)^2}] \Rightarrow y \gg a \Rightarrow \frac{1}{\pi\varepsilon_0}\frac{qa}{y^3}$

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. $dE = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x+a)^2}, E = \frac{\lambda}{4\pi\varepsilon_0}[\frac{1}{l+a} - \frac{1}{a}]$

87. At $P_1 : \vec{E}(y) = \frac{1}{4\pi\varepsilon_0}\frac{\lambda L}{y\sqrt{y^2+\frac{L^2}{4}}}\hat{j} \Rightarrow \frac{1}{4\pi\varepsilon_0}\frac{q}{\frac{a}{2}\sqrt{(\frac{a}{2})^2+\frac{L^2}{4}}}\hat{j} = \frac{1}{\pi\varepsilon_0}\frac{q}{a\sqrt{a^2+L^2}}\hat{j}$

At P_2: Put the origin at the end of **L**.

$dE = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x+a)^2}, \vec{E} = -\frac{q}{4\pi\varepsilon_0 l}[\frac{1}{l+a} - \frac{1}{a}]\hat{i}$

97. circular arc $dE_x(-\hat{i}) = \frac{1}{4\pi\varepsilon_0}\frac{\lambda ds}{r^2}cos\theta(-\hat{i}$,

$\vec{E}_x = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{i}),$

$dEy(-\hat{i}\hat{}) = \frac{1}{4\pi\varepsilon_0}\frac{\lambda ds}{r^2}sin\theta(-\hat{j}),$

$\vec{E}_y = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{j});$

y-axis: $\vec{E}_x = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{i});$

x-axis: $\vec{E}_y = \frac{\lambda}{4\pi\varepsilon_0 r}(-\hat{j}),$

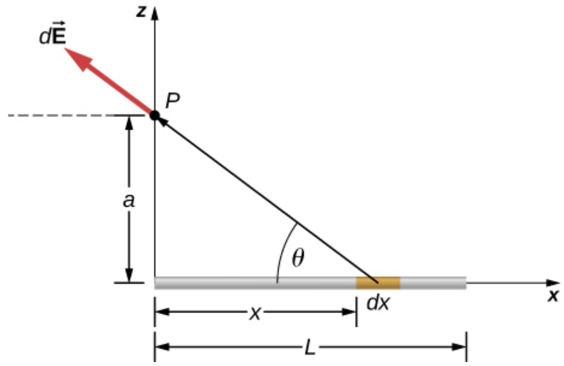
$\vec{E} = \frac{\lambda}{2\pi\varepsilon_0 r}(-\hat{i}) + \frac{\lambda}{2\pi\varepsilon_0 r}(-\hat{j})$

Additional Problems

121. Electric field of wire at **x**: $\vec{E}(x) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda_y}{x}\hat{i},$

$$dF = \frac{\lambda_y \lambda_x}{2\pi\varepsilon_0}(lnb - lna)$$

123.



$$dEx = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x^2+a^2)}\frac{x}{\sqrt{x^2+a^2}},$$

$$\vec{E}_x = \frac{\lambda}{4\pi\varepsilon_0}[\frac{1}{\sqrt{L^2+a^2}} - \frac{1}{a}]\hat{i},$$

$$dE_z = \frac{1}{4\pi\varepsilon_0}\frac{\lambda dx}{(x^2+a^2)}\frac{a}{\sqrt{x^2+a^2}},$$

$$\vec{E}_z = \frac{\lambda}{4\pi\varepsilon_0 a}\frac{L}{\sqrt{L^2+a^2}}\hat{k},$$

Substituting z for a, we have:

$$\vec{E}(z) = \frac{\lambda}{4\pi\varepsilon_0}[\frac{1}{\sqrt{L^2+z^2}} - \frac{1}{z}]\hat{i} + \frac{\lambda}{4\pi\varepsilon_0 z}\frac{L}{\sqrt{L^2+z^2}}\hat{k}$$

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.

$$dF_y = \frac{1}{4\pi\varepsilon_0}\frac{aq\lambda dx}{(x^2+a^2)^{3/2}},$$

$$Fy = \frac{1}{2\pi\varepsilon_0}\frac{q\lambda}{a}[\frac{l/2}{((l/2)^2+a^2)^{1/2}}]$$

CHAPTER OVERVIEW

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

---

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.

- **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.

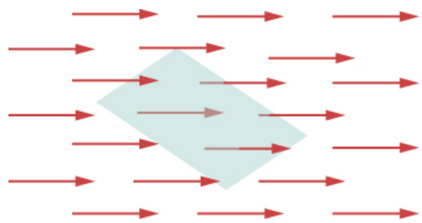


Figure 17.2.1: The flux of an electric field through the shaded area captures information about the “number” of electric field lines passing through the area. The numerical value of the electric flux depends on the magnitudes of the electric field and the area, as well as the relative orientation of 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 $\vec{E} = E\hat{y}$. 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**.

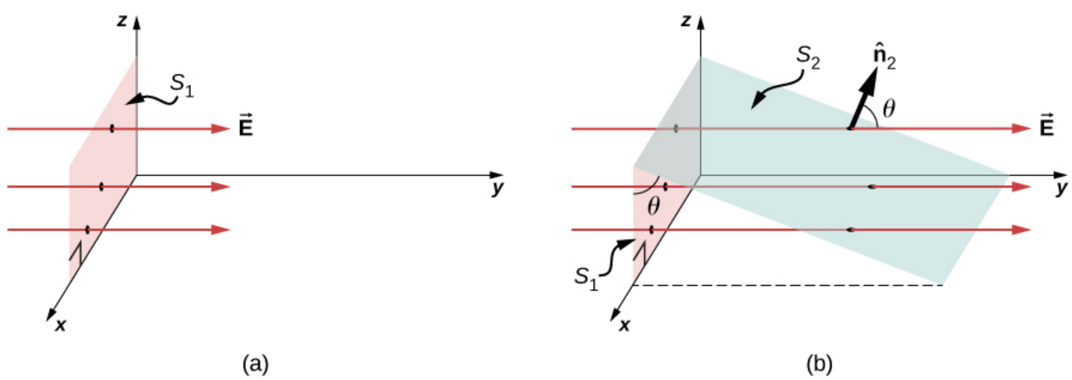


Figure 17.2.2: (a) A planar surface S_1 of area A_1 is perpendicular to the electric field $E\hat{j}$. N field lines cross surface S_1. (b) A surface S_2 of area A_2 whose projection onto the xz-plane is S_1. The same number of field lines cross each surface.

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

$$\Phi = \vec{E} \cdot \hat{n}_2 A_2. \tag{17.2.1}$$

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 $17.2.3$.

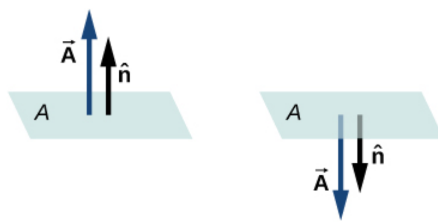


Figure 17.2.3: The direction of the area vector of an open surface needs to be chosen; it could be either of the two cases displayed here. The area vector of a part of a closed surface is defined to point from the inside of the closed space to the outside. This rule gives a unique direction.

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.

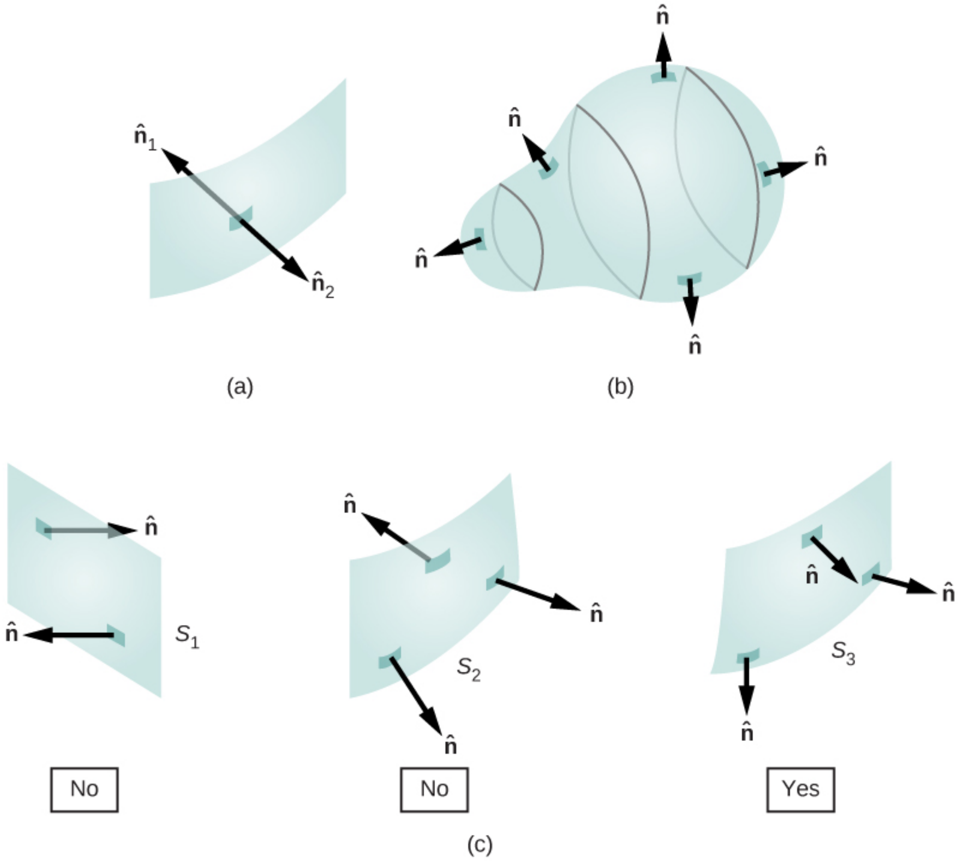


Figure 17.2.4: (a) Two potential normal vectors arise at every point on a surface. (b) The outward normal is used to calculate the flux through a closed surface. (c) Only S_3 has been given a consistent set of normal vectors that allows us to define the flux through the surface.

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:

$$\Phi = \vec{E} \cdot \vec{A} \, (uniform\, \hat{E},\ flat\, surface). \tag{17.2.2}$$

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?

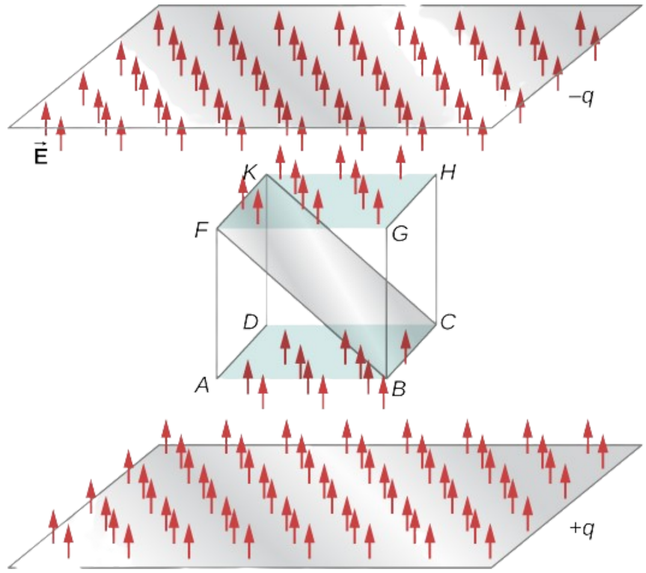


Figure 17.2.5: Electric flux through a cube, placed between two charged plates. Electric flux through the bottom face ($ABCD$) is negative, because $\vec{E}$ is in the opposite direction to the normal to the surface. The electric flux through the top face ($FGHK$) is positive, because the electric field 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.

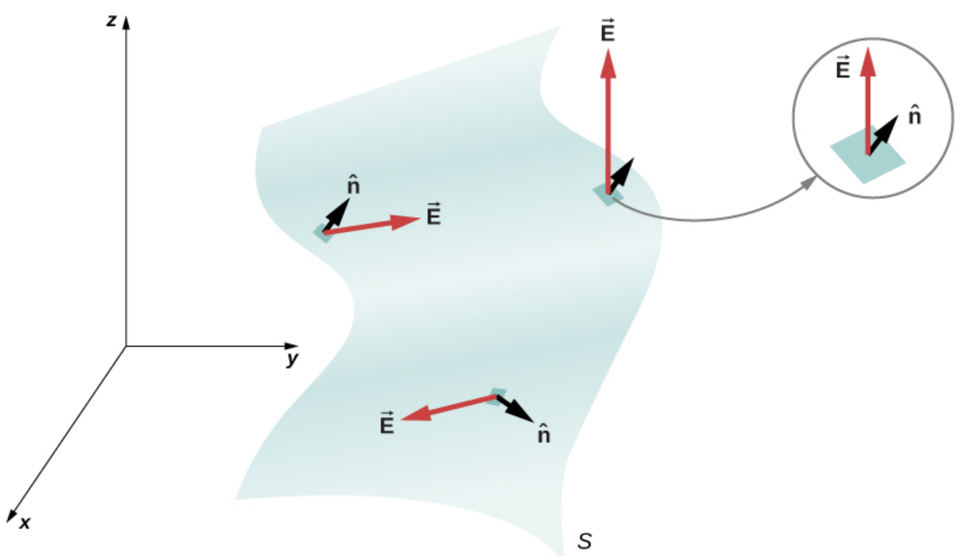


Figure 17.2.6: A surface is divided into patches to find the flux.

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 ith 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 $\vec{E}_i$.

$$\vec{E}_i = \text{average electric field over the } i\text{th patch.} \tag{17.2.3}$$

Therefore, we can write the electric flux Φ through the area of the ith patch as

$$\Phi_i = \vec{E}_i \cdot \delta\vec{A}_i \; (i\text{th patch}). \tag{17.2.4}$$

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.

$$\Phi = \sum_{i=1}^{N} \Phi_i = \sum_{i=1}^{N} \vec{E}_i \cdot \delta\vec{A}_i \; (N\, patch\, estimate). \tag{17.2.5}$$

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^o and negative when the angle is greater than 90^o. 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,

$$\Phi = \int_S \vec{E} \cdot \hat{n} dA = \int_S \vec{E} \cdot d\vec{A} \, (open\, surface). \tag{17.2.6}$$

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

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \oint_S \vec{E} \cdot d\vec{A} \, (closed\, surface) \tag{17.2.7}$$

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?

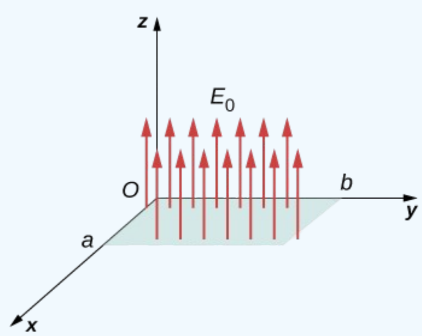


Figure 17.2.7: Calculating the flux of E_0 through a rectangular surface.

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?

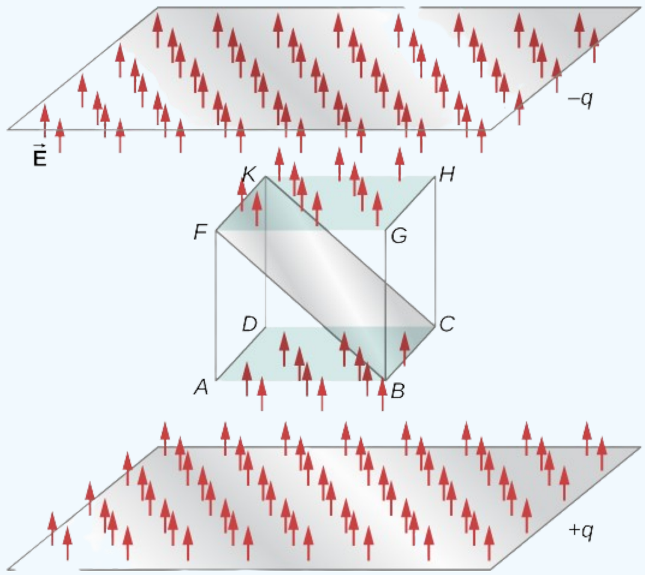


Figure 17.2.8: Calculating the flux of E_0 through a closed cubic surface.

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^o 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.

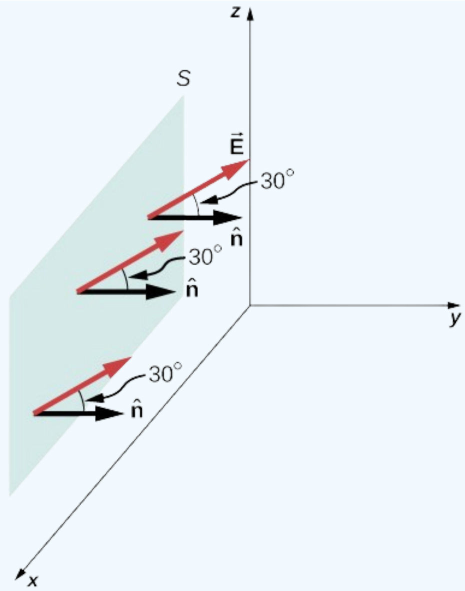


Figure 17.2.9: The electric field produces a net electric flux through the surface **S**.

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^o. 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

$$\Phi = \int_S \vec{E} \cdot \hat{n} dA = EA\, cos\, \theta \tag{17.2.8}$$

$$= (10\, \mathrm{N/C})(6.0\, m^2)(\cos\, 30^\circ) = 52\, \mathrm{N \cdot m^2/C}. \tag{17.2.9}$$

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?

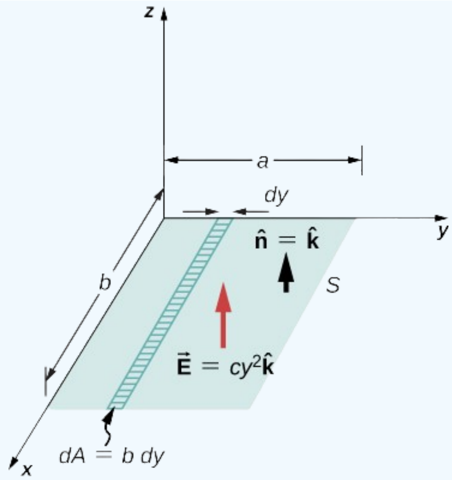


Figure 17.2.10: Since the electric field is not constant over the surface, an integration is necessary to determine the flux.

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

$$\begin{aligned}\Phi &=\int_S \vec{E}\cdot\hat{n}dA=\int_0^a (cy^2\hat{k})\cdot\hat{k}(b\,dy)\\ &=cb\int_0^a y^2dy=\frac{1}{3}a^3bc.\end{aligned}$$

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

$mab^2/2$

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

$$\vec{E}_p = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\hat{r}, \tag{17.3.1}$$

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 $17.3.1$.

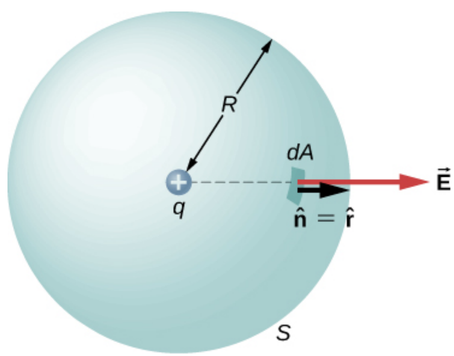


Figure 17.3.1: A closed spherical surface surrounding a point charge q.

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:

$$\Phi = \frac{1}{4\pi\epsilon_0}\frac{q}{R^2}\oint_S dA = \frac{1}{4\pi\epsilon_0}\frac{q}{R^2}(4\pi R^2) = \frac{q}{\epsilon_0}. \tag{17.3.2}$$

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

$$\Phi = \frac{q}{\epsilon_0}. \tag{17.3.3}$$

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 $17.3.2$).

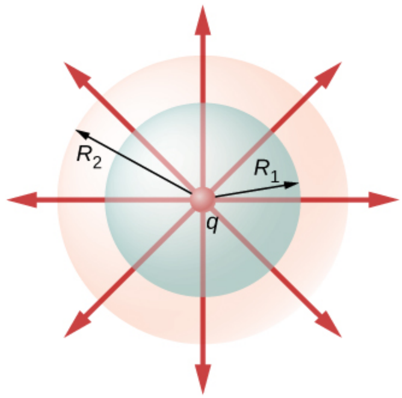


Figure $17.3.2$: Flux through spherical surfaces of radii R_1 and R_2 enclosing a charge q are equal, independent of the size of the surface, since all $\vec{E}$-field lines that pierce one surface from the inside to outside direction also pierce the other surface in the same direction.

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 $17.3.3c$).

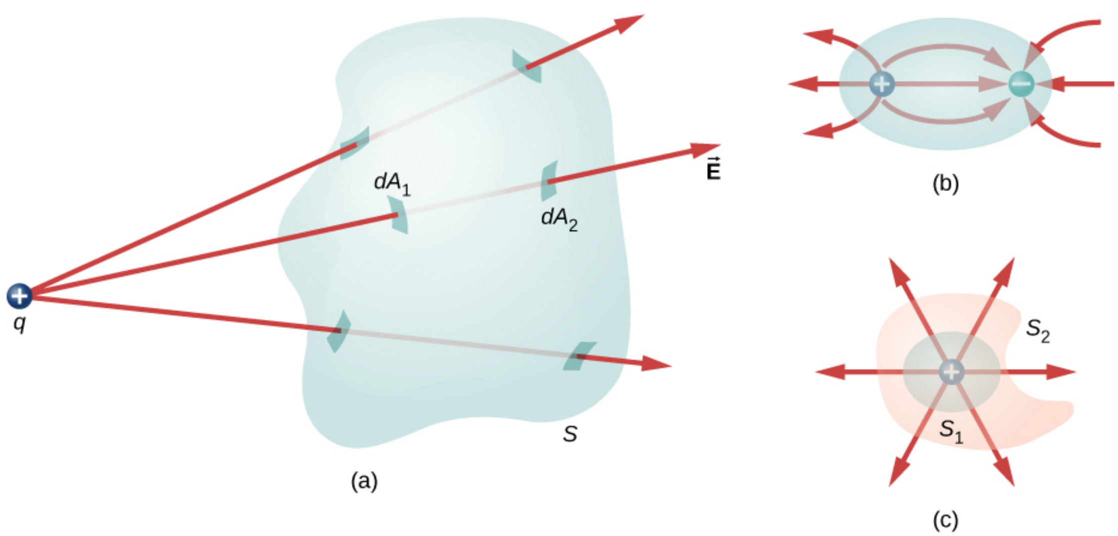


Figure $17.3.3$: Understanding the flux in terms of field lines. (a) The electric flux through a closed surface due to a charge outside that surface is zero. (b) Charges are enclosed, but because the net charge included is zero, the net flux through the closed surface is also zero. (c) The shape and 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):

$$\Phi_{\text{ClosedSurface}} = \frac{q_{enc}}{\epsilon_0}. \tag{17.3.4}$$

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.

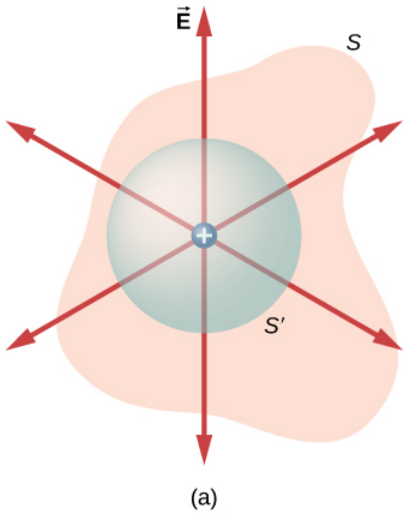


Figure 17.3.4: The electric flux through any closed surface surrounding a point charge q is given by Gauss's law. (a) Enclosed charge is positive. (b) Enclosed charge is negative.

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

$$\Phi = (q_1 + q_2 + q_5)/\epsilon_0. \tag{17.3.5}$$

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.

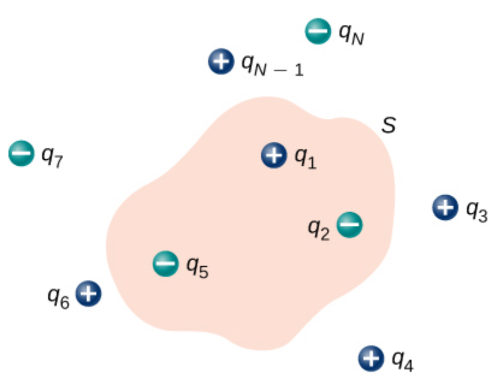


Figure 17.3.5: The flux through the Gaussian surface shown, due to the charge distribution, is $\Phi = (q_1 + q_2 + q_5)/\epsilon_0$.

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):

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\epsilon_0}. \tag{17.3.6}$$

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 $17.3.7$.

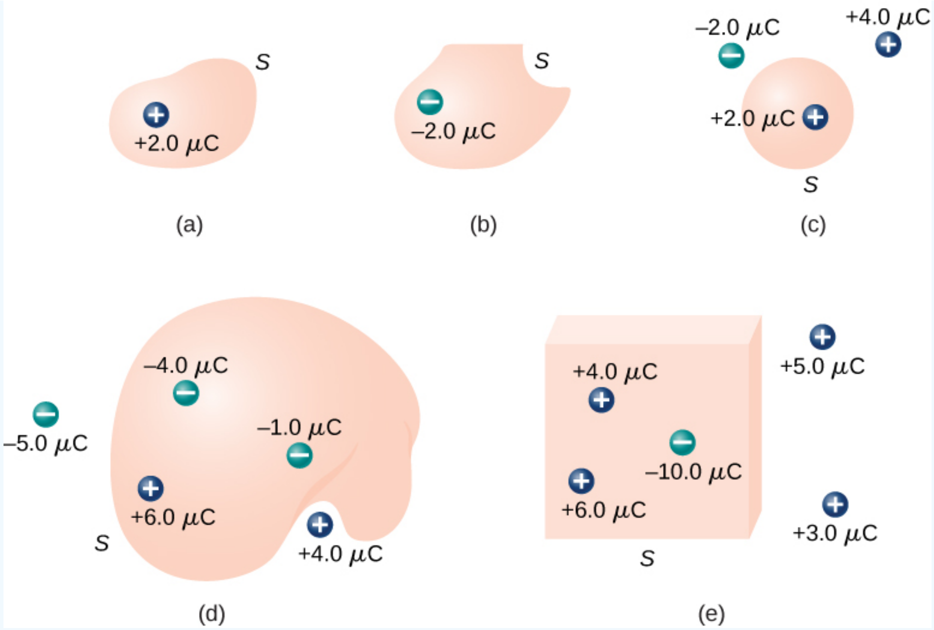


Figure 17.3.7: Various Gaussian surfaces and charges.

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

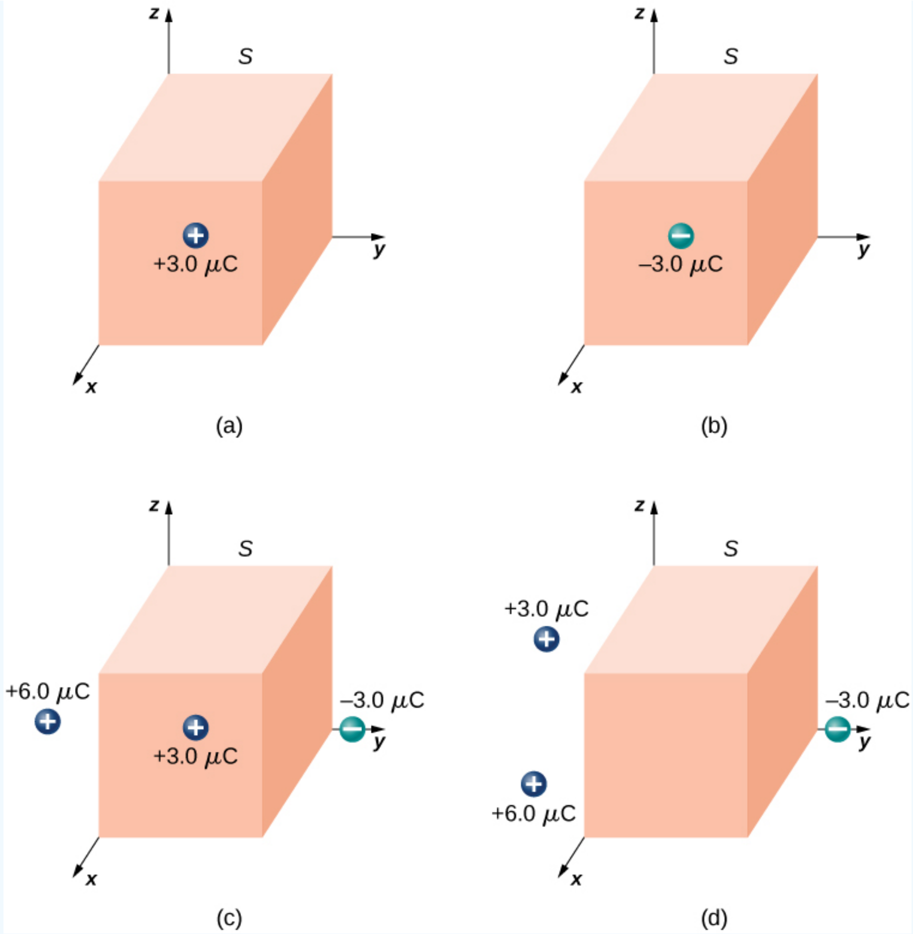


Figure 17.3.8: A cubical Gaussian surface with various charge distributions.

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.

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

$$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = E \oint_S dA = EA = \frac{q_{enc}}{\epsilon_0}, \tag{17.4.1}$$

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

$$E \approx \frac{q_{enc}}{\epsilon_0 \, area}. \tag{17.4.2}$$

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.

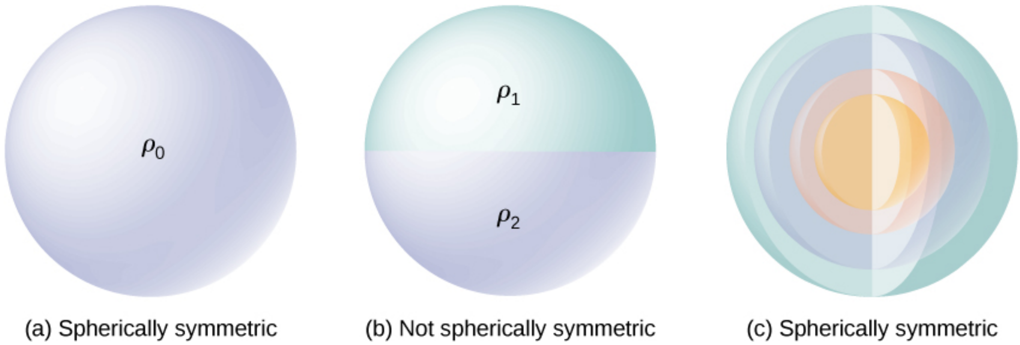


Figure $17.4.1$: Illustrations of spherically symmetrical and nonsymmetrical systems. Different shadings indicate different charge densities. Charges on spherically shaped objects do not necessarily mean the charges are distributed with spherical symmetry. The spherical symmetry occurs only when the 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:

$$Spherical\, symmetry: \vec{E}_p = E_p(r)\hat{r}, \tag{17.4.3}$$

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 $17.4.2$).

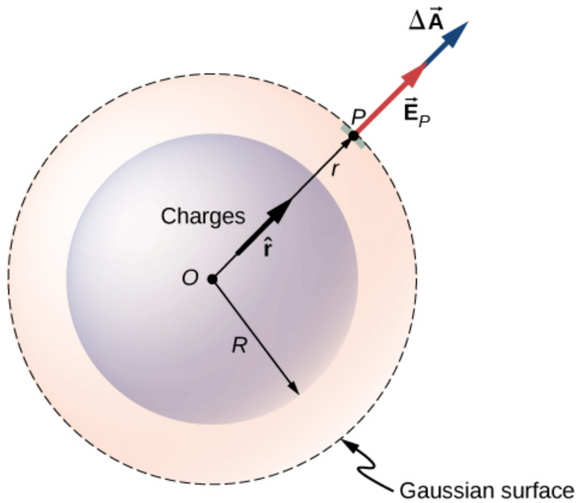


Figure 17.4.2: The electric field at any point of the spherical Gaussian surface for a spherically symmetrical charge distribution is parallel to the area element vector at that point, giving flux as the product of the magnitude of electric field and the value of the area. Note that the radius R of the 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:**

$$\Phi = \oint_S \vec{E}_p \cdot \hat{n} dA = E_p \oint_S dA = E_p \, 4\pi r^2. \tag{17.4.4}$$

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:

$$4\pi r^2 E = \frac{q_{enc}}{\epsilon_0}. \tag{17.4.5}$$

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:

$$Magnitude: \; E(r) = \frac{1}{4\pi\epsilon_0}\frac{q_{enc}}{r^2} \tag{17.4.6}$$

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 $17.4.3$.

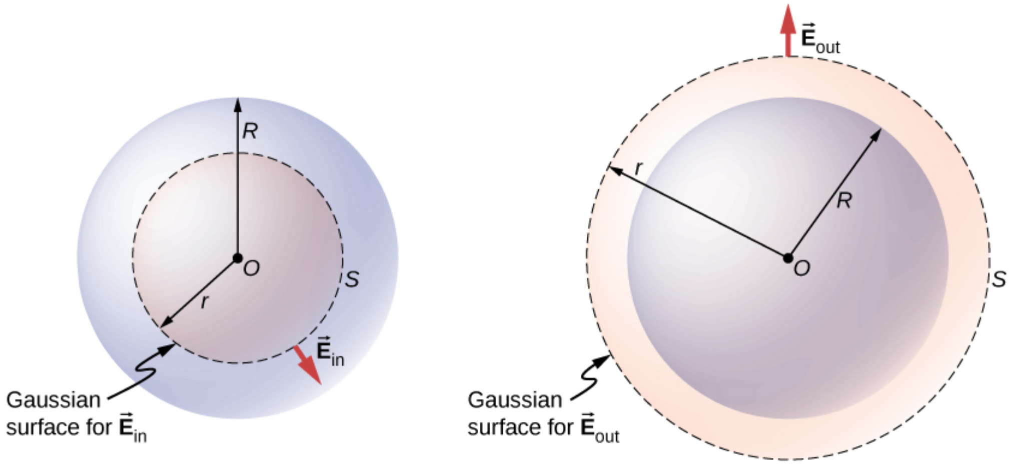


Figure 17.4.3: A spherically symmetrical charge distribution and the Gaussian surface used for finding the field (a) inside and (b) outside the distribution.

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

$$q_{enc} = q_{tot} \, (total \, charge) \, if \, r \geq R \tag{17.4.7}$$

$$q_{enc} = q_{within \, r<R} \, (only \, charge \, within \, r < R) \, if \, r < R \tag{17.4.8}$$

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

$$P \, outside \, sphere \, E_{out} = \frac{1}{4\pi\epsilon_0} \frac{q_{tot}}{r^2} \tag{17.4.9}$$

$$P \, inside \, sphere \, E_{in} = \frac{1}{4\pi\epsilon_0} \frac{q_{within \, r<R}}{r^2}. \tag{17.4.10}$$

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

$$q_{enc} = \int \rho_0 dV = \int_0^r \rho_0 4\pi r'^2 dr' = \rho \left(\frac{4}{3} \pi r^3 \right). \tag{17.4.11}$$

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}.

$$E_{out} = \frac{1}{4\pi\epsilon_0} \frac{q_{tot}}{r^2}, \; q_{tot} = \frac{4}{3} \pi R^3 \, \rho_0, \tag{17.4.12}$$

$$E_{in} = \frac{q_{enc}}{4\pi\epsilon_0 r^2} = \frac{\rho_0 r}{3\epsilon_0},\ since\ q_{enc} = \frac{4}{3}\pi r^3 \rho_0. \tag{17.4.13}$$

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.

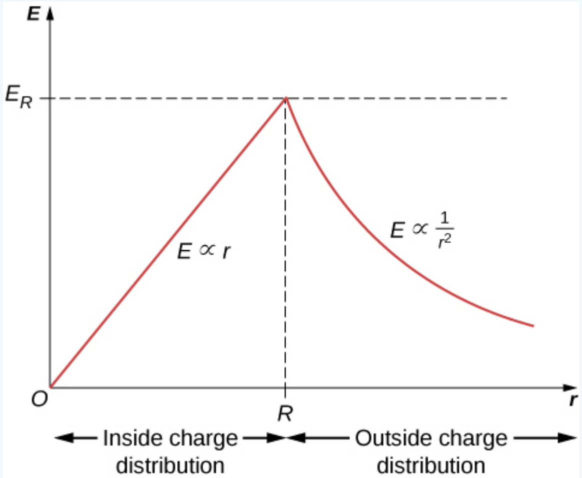


Figure 17.4.4: Electric field of a uniformly charged, non-conducting sphere increases inside the sphere to a maximum at the surface and then decreases as $1/r^2$. Here, $E_R = \frac{\rho_0 R}{3\epsilon_0}$. The electric field is due to a spherical charge distribution of uniform charge density and total 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 $r = 2R$.

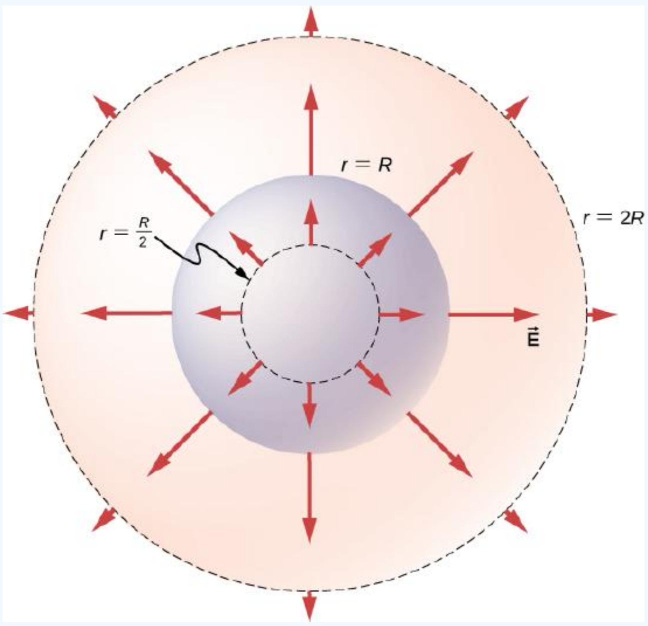


Figure 17.4.5: Electric field vectors inside and outside a uniformly charged sphere.

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

$$\rho(r) = ar^n (r \leq R;\ n \geq 0),$$

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:

$$dq = ar'^n 4\pi r'^2 dr'.$$

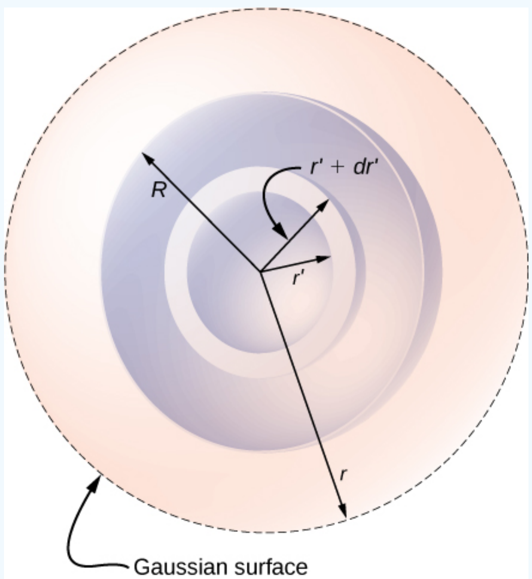


Figure 17.4.6: Spherical symmetry with non-uniform charge distribution. In this type of problem, we need four radii: R is the radius of the charge distribution, r is the radius of the Gaussian surface, r' is the inner radius of the spherical shell, and $r' + dr'$ is the outer radius of the spherical 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 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 r_{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 FiFigure $17.4.7d$, 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.

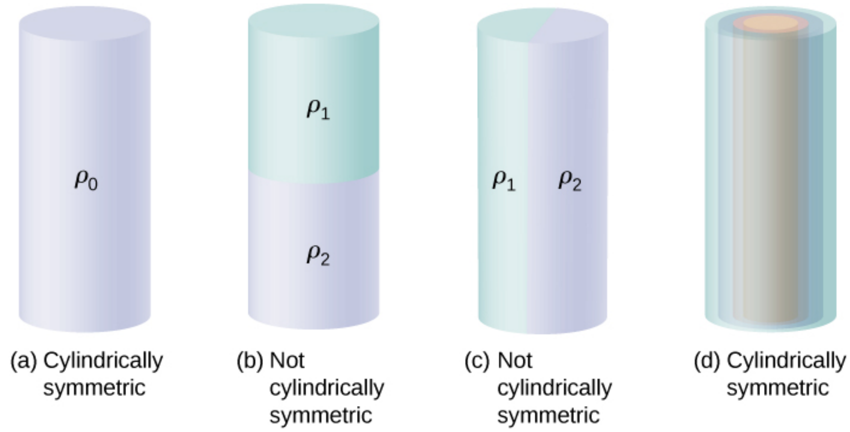


Figure 17.4.7: To determine whether a given charge distribution has cylindrical symmetry, look at the cross-section of an "infinitely long" cylinder. If the charge density does not depend on the polar angle of the cross-section or along the axis, then you have cylindrical symmetry. (a) Charge density 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 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).

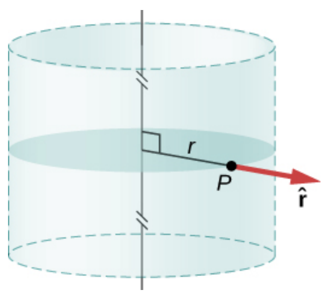


Figure 17.4.8: The electric field in a cylindrically symmetrical situation depends only on the distance from the axis. The direction of the electric field is pointed away from the axis for positive charges and toward the axis for negative charges.

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).

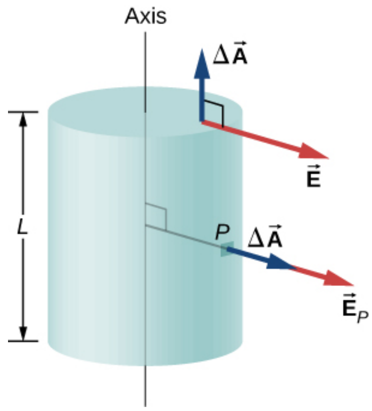


Figure 17.4.9: The Gaussian surface in the case of cylindrical symmetry. The electric field at a patch is either parallel or perpendicular to the normal to the patch of the Gaussian surface.

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:

$$q_{enc} = \lambda_{enc} L. \tag{17.4.14}$$

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:

$$Magnitude: \; E(r) = \frac{\lambda_{enc}}{2\pi\epsilon_0}\frac{1}{r}. \tag{17.4.15}$$

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:

$\lambda_{enc} = (\text{totalcharge}) \, if \, r \geq R$

$\lambda_{enc} = (\text{onlychargewithin} \, t < R) \, if \, r < R$

✔ 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

$$\lambda_{enc} = \frac{\sigma_0 2\pi RL}{L} = 2\pi R\sigma_0. \tag{17.4.16}$$

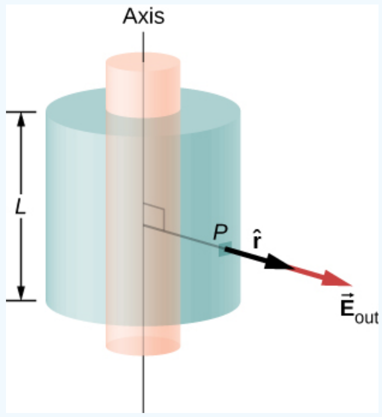


Figure 17.4.10: A Gaussian surface surrounding a cylindrical shell.

Hence, the electric field at a point P outside the shell at a distance **r** away from the axis is

$$\vec{E} = \frac{2\pi R\sigma_0}{2\pi\epsilon_0}\frac{1}{r}\hat{r} = \frac{R\sigma_0}{\epsilon_0}\frac{1}{r}\hat{r}\ (r > R) \tag{17.4.17}$$

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:

$$\lambda_{enc} = 0. \tag{17.4.18}$$

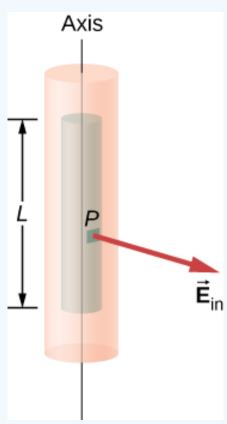


Figure 17.4.11: A Gaussian surface within a cylindrical shell.

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.

$$E_{in}2\pi rL = 0(r < R),$$

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

$\vec{E} = \frac{\lambda_0}{2\pi\epsilon_0}\frac{1}{d}\hat{r}$; 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.

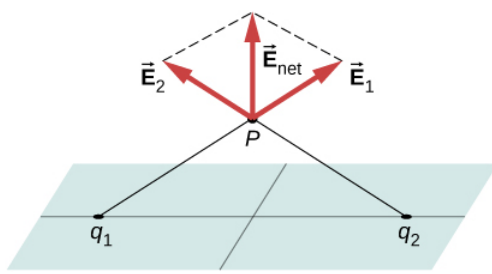


Figure 17.4.12: The components of the electric field parallel to a plane of charges cancel out the two charges located symmetrically from the field point **P**. Therefore, the field at any point is pointed vertically from the plane of charges. For any point **P** and charge q_1, we can always find a 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.

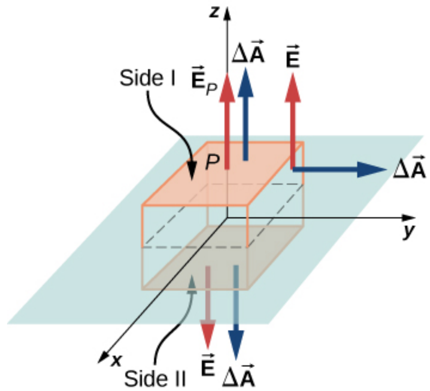


Figure 17.4.13: A thin charged sheet and the Gaussian box for finding the electric field at the field point P. The normal to each face of the box is from inside the box to outside. On two faces of the box, the electric fields are parallel to the area vectors, and on the other four faces, the electric 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

$$\Phi = \int_S \vec{E}_p \cdot \hat{n} dA = E_p A + E_p A + 0 + 0 + 0 + 0 = 2E_p A \tag{17.4.19}$$

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.

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.

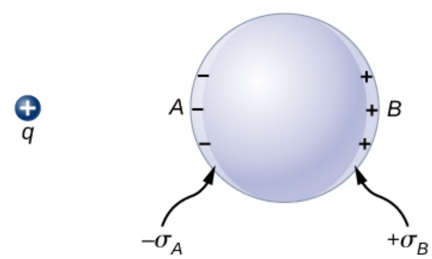


Figure $17.5.1$: Polarization of a metallic sphere by an external point charge $+q$. The near side of the metal has an opposite surface charge compared to the far side of the metal. The sphere is said to be polarized. When you remove the external charge, the polarization of the metal also disappears.

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.

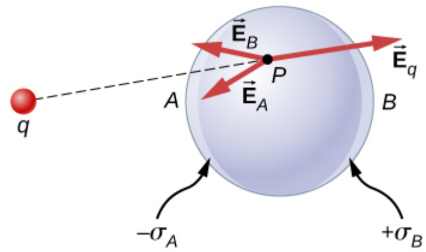


Figure 17.5.2: In the presence of an external charge q, the charges in a metal redistribute. The electric field at any point has three contributions, from $+q$ and the induced charges $-\sigma_A$ and $+\sigma_B$. Note that the surface charge distribution will not be uniform in this case.

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, $q_{enc} = 0$ and hence

$$\vec{E}_{net} = \vec{0}\,(\mathrm{at points inside a conductor}). \tag{17.5.1}$$

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.

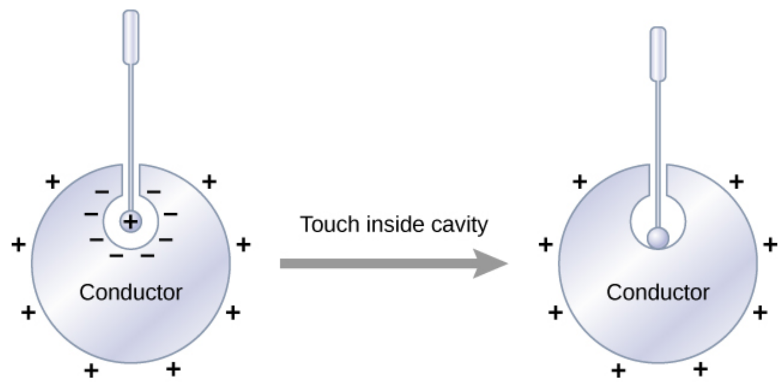


Figure 17.5.3: Electric charges on a conductor migrate to the outside surface no matter where you put them initially.

To see why this happens, note that the Gaussian surface in iFigure $17.5.4$(the dashed line) follows the contour of the actual surface of the conductor and is located an infinitesimal distance **within** it. Since $E = 0$ everywhere inside a conductor,

$$\oint \vec{E} \cdot \hat{n} dA = 0. \tag{17.5.2}$$

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.

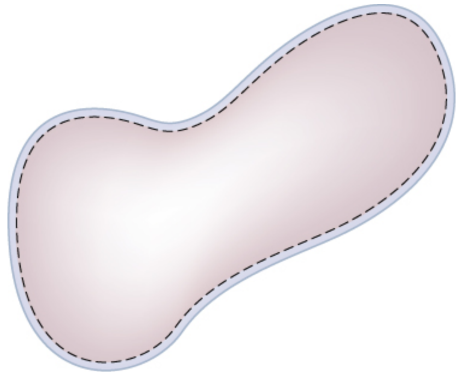


Figure 17.5.4: The dashed line represents a Gaussian surface that is just beneath the actual surface of the conductor.

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^{-16} [2] , a number so small that the validity of Coulomb's law seems indisputable.

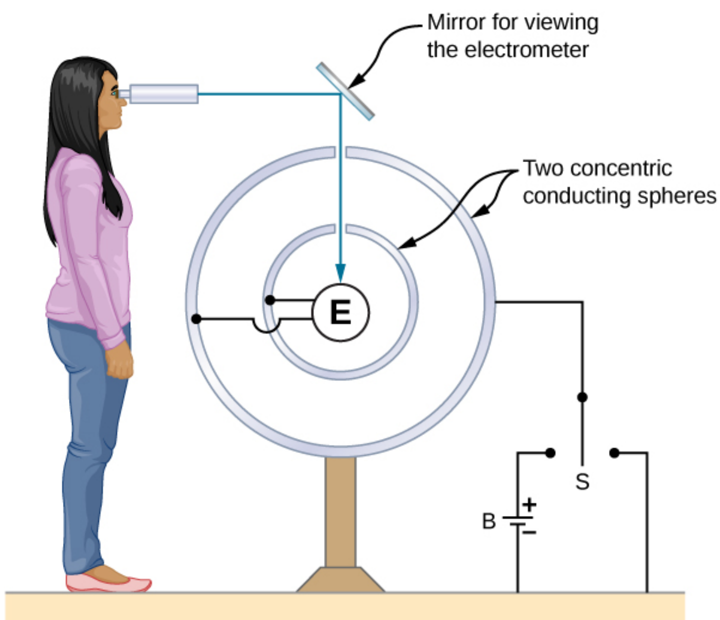


Figure 17.5.5: A representation of the apparatus used by Plimpton and Lawton. Any transfer of charge between the spheres is detected by the electrometer E.

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 δ and the magnitude of the electric field $\vec{E}$ are related by

$$E = \frac{\sigma}{\epsilon_0}. \tag{17.5.3}$$

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,

$$E\Delta A = \frac{\sigma \Delta A}{\epsilon_0}. \tag{17.5.4}$$

Thus

$$E = \frac{\sigma}{\epsilon_0}. \tag{17.5.5}$$

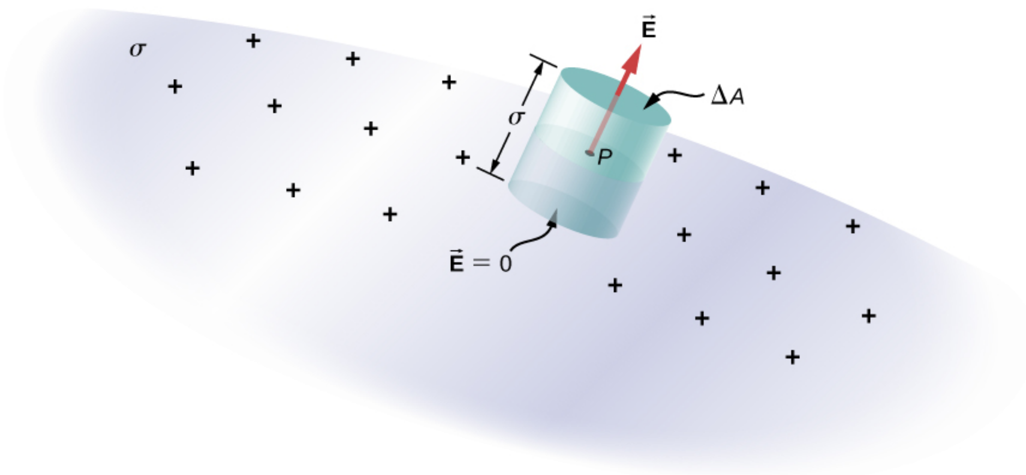


Figure 17.5.6: An infinitesimally small cylindrical Gaussian surface surrounds point **P**, which is on the surface of the conductor. The field $\vec{E}$ is perpendicular to the surface of the conductor outside the conductor and vanishes within it.

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.

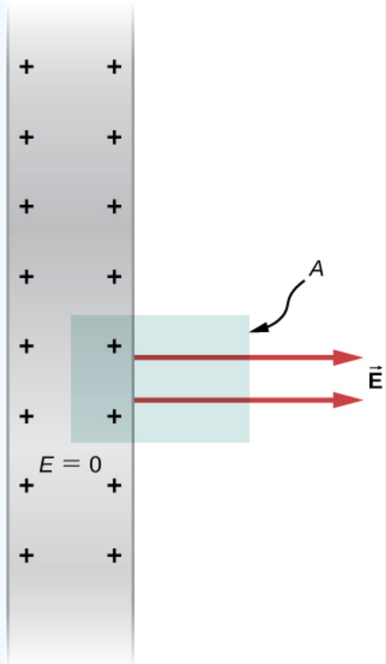


Figure 17.5.7: A side view of an infinite conducting plate and Gaussian cylinder with cross-sectional area **A**.

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 $l = 6.50\, mm$. What is the electric field between the plates?

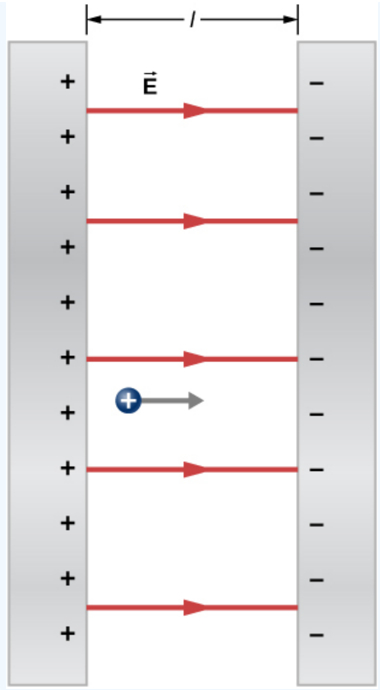


Figure 17.5.8: The electric field between oppositely charged parallel plates. A test charge is released at the positive plate.

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

$$E = \frac{\sigma}{\epsilon_0} = \frac{6.81\times10^{-7}C/m^2}{8.85\times10^{-12}C^2/Nm^2} = 7.69\times10^4 N/C \tag{17.5.6}$$

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?

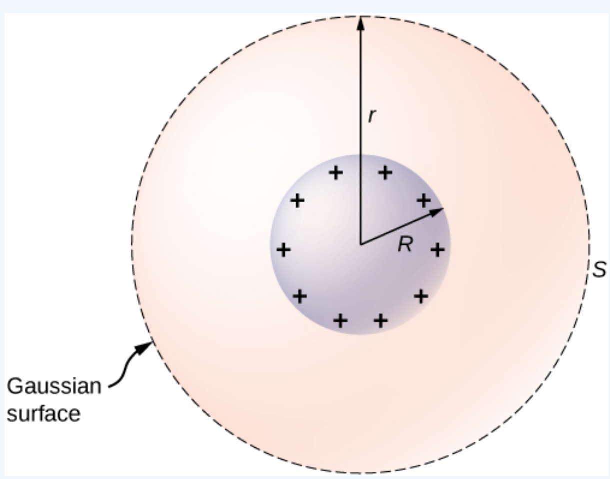


Figure 17.5.9: An isolated conducting sphere.

Strategy The sphere is isolated, so its surface change 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 **r** is constant and $\hat{n} = \hat{r}$ on the sphere,

$$\oint_S \vec{E} \cdot \hat{n} dA = E(r) \oint_S dA = E(r) 4\pi r^2. \tag{17.5.7}$$

For $r < R$, S is within the conductor, so $q_{enc} = 0$, and Gauss's law gives

$$E(r) = 0, \tag{17.5.8}$$

as expected inside a conductor. If $r > R$, S encloses the conductor so $q_{enc} = q$. From Gauss's law,

$$E(r) 4\pi r^2 = \frac{q}{\epsilon_0}. \tag{17.5.9}$$

The electric field of the sphere may therefore be written as

$$\vec{E} = \vec{0} \, (r < R), \tag{17.5.10}$$

$$\vec{E} = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} \hat{r} \, (r \geq R). \tag{17.5.11}$$

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.

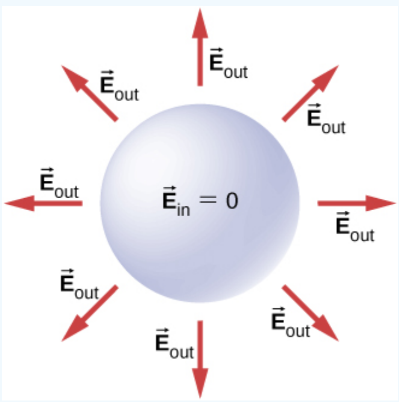


Figure 17.5.10: Electric field of a positively charged metal sphere. The electric field inside is zero, and the electric field outside is same as the electric field of a point charge at the center, although the charge on the metal sphere is at the surface.

? 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 direction 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).

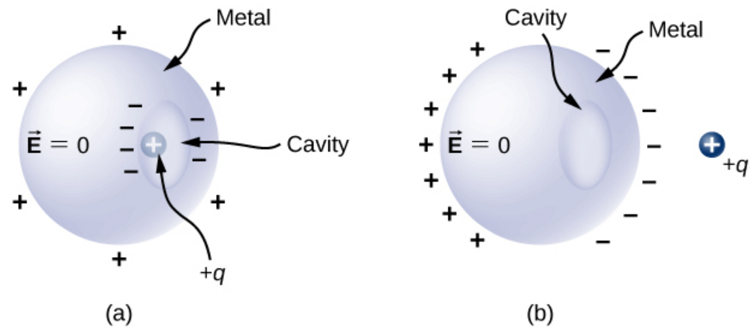


Figure 17.5.11: (a) A charge inside a cavity in a metal. The distribution of charges at the outer surface does not depend on how the charges are distributed at the inner surface, since the **E**-field inside the body of the metal is zero. That magnitude of the charge on the outer surface does depend 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.

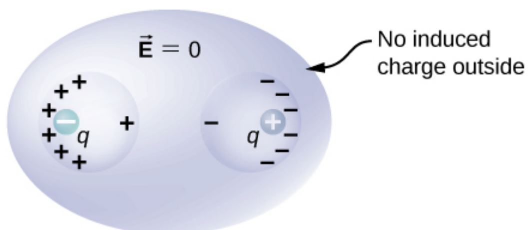


Figure 17.5.12: The charges induced by two equal and opposite charges in two separate cavities of a conductor. If the net charge on the cavity is nonzero, the external surface becomes charged to the amount of the net charge.

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	$\Phi = \vec{E} \cdot \vec{A} \to EA\cos\theta$
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	$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\varepsilon_0}$
Gauss's Law for systems with symmetry	$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = E \oint_S dA = EA = \frac{q_{enc}}{\varepsilon_0}$
The magnitude of the electric field just outside the surface of a conductor	$E = \frac{\sigma}{\varepsilon_0}$

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,

  $$\Phi = \oint_S \vec{E} \cdot \hat{n} dA = \frac{q_{enc}}{\varepsilon_0} \;,$$

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 $4\pi r^2 E = \frac{q_{enc}}{\varepsilon_0}$.
- For cylindrical symmetry, we use a cylindrical Gaussian surface, and find that Gauss's law simplifies to $2\pi rLE = \frac{q_{enc}}{\varepsilon_0}$.
- 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 $2AE = \frac{q_{enc}}{\varepsilon_0}$.

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 $E = \frac{\sigma}{\varepsilon_0}$.

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.

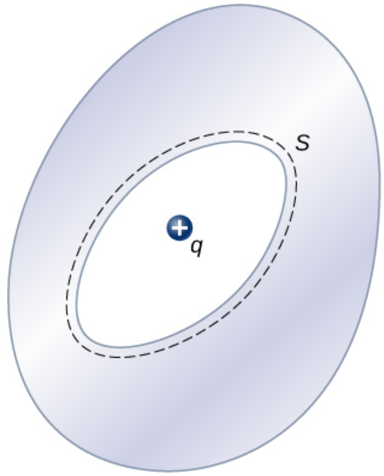
15. Discuss the restrictions on the Gaussian surface used to discuss planar symmetry. For example, is its length important? Does the cross-section have to be square? Must the end faces be on opposite sides of the sheet?

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 **3cm×2cm** 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° +30°**.

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 $150 cm^2$ face each other with a separation of 3 mm between them. The plates are charged with equal amount of opposite charges, **±20μC**. 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°+5°**.

25. A square surface of area $2 cm^2$ 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°+θ**.

26. A vector field is pointed along the **z**-axis, \(\displaystyle \vec{v} =\frac{α}{x^2+y^2}\hat{z}.

(a) Find the flux of the vector field through a rectangle in the xy-plane between a<x<ba<x<b and c<y<dc<y<d.

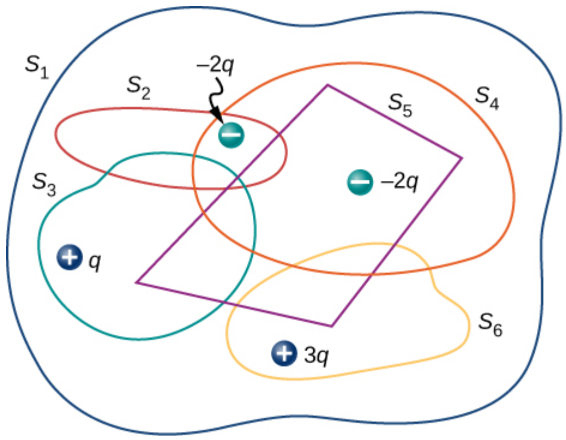
(b) Do the same through a rectangle in the yz-plane between a<z<ba<z<b and c<y<dc<y<d. (Leave your answer as an integral.)

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?

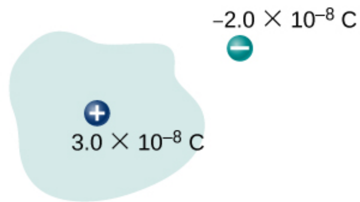
28. Repeat the previous problem, given that the circular area is (a) in the **yz**-plane and (b) **45°** above the **xy**-plane.

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?

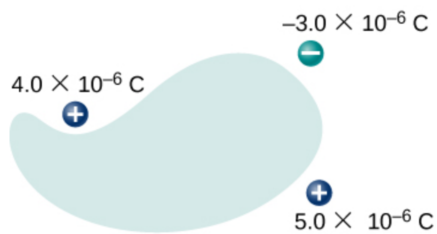
30. Determine the electric flux through each closed surface whose cross-section inside the surface is shown below.



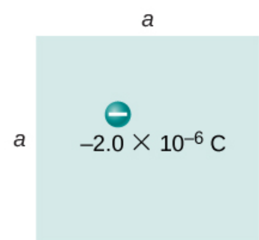
31. Find the electric flux through the closed surface whose cross-sections are shown below.



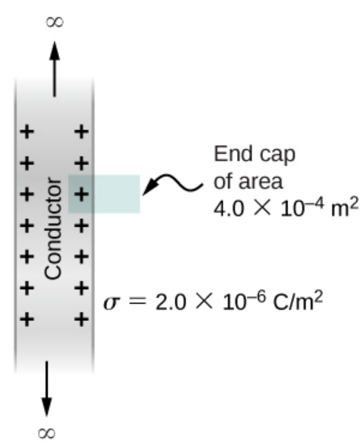
(a)



(b)



(c)



(d)

32. A point charge **q** is located at the center of a cube whose sides are of length **a**. If there are no other charges in this system, what is the electric flux through one face of the cube?

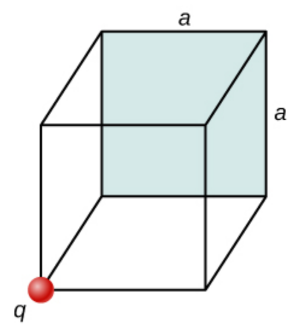
33. A point charge of $10\mu C$ is at an unspecified location inside a cube of side 2 cm. Find the net electric flux though the surfaces of the cube.

34. A net flux of $1.0 \times 10^4 N \cdot^{m} 2/C$ 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

$$\vec{E} = \vec{0} \ (r \leq r_1);$$

$$\vec{E} = \frac{Q}{4\pi\varepsilon_0 r^2}\left(\frac{r^3 - r_1^3}{r_2^3 - r_1^3}\right)\hat{r} \ \ (r_1 \leq r \leq r_2);$$

$$\vec{E}=\frac{Q}{4\pi\varepsilon_0 r^2}\hat{r}\ \ (r\geq r_2)\,.$$

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

$$E=\frac{\rho r}{2\varepsilon_0}\ \ (r\leq R)\,;$$

$$E=\frac{\rho R^2}{2\varepsilon_0 r}\ \ (r\geq R)$$

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

$$E=\frac{\alpha r^2}{3\varepsilon_0}\ \ (r\leq R)\,;$$

$$E=\frac{\alpha R^3}{3\varepsilon_0 r}\ \ (r\geq R)\,.$$

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 $4.0 \times 10^2 N/C$. 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 αα 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 $R = 7.4 \times 10^{-15} m$ with a charge of **92e** distributed uniformly throughout its volume. (a) What is the electric force exerted on an electron when it is $3.0 \times 10^{-15} m$ 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?

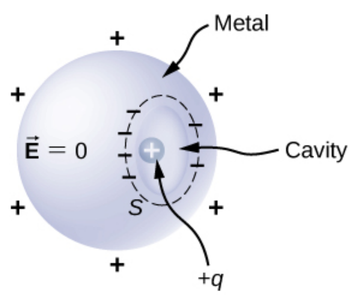
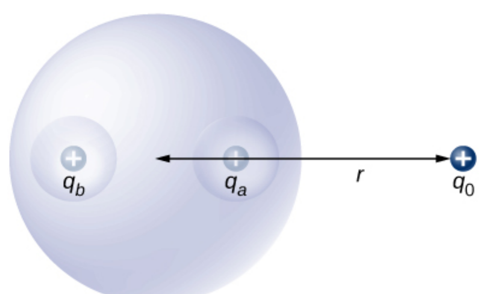


Figure 6.46: A charge inside a cavity of a metal. Charges at the outer surface do not depend on how the charges are distributed at the inner surface since E field inside the body of the metal is zero.

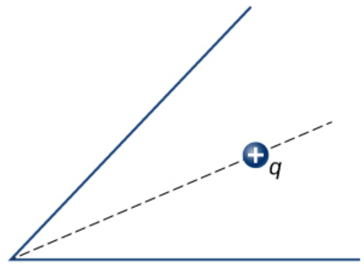
61. An uncharged spherical conductor **S** of radius **R** has two spherical cavities **A** and **B** of radii **a** and **b**, respectively as shown below. Two point charges $+q_a$ and $+q_b$ are placed at the center of the two cavities by using non-conducting supports. In addition, a point charge $+q_0$ 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 **5μC** of charge. A copper spherical shell of inner radius 6 cm and outer radius 8 cm surrounds it. A total charge of **−8μC** 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.

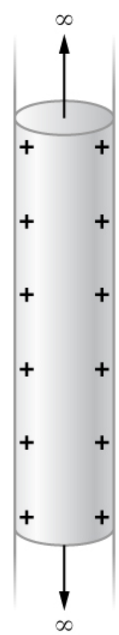
65. A long cylinder of aluminum of radius **R** meters is charged so that it has a uniform charge per unit length on its surface of λ. (a) Find the electric field inside and outside the cylinder. (b) Plot electric field as a function of distance from the center of the rod.

66. At the surface of any conductor in electrostatic equilibrium, $E=\sigma/\varepsilon_0$. 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 $5.0\times10^{-9}C$. 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?

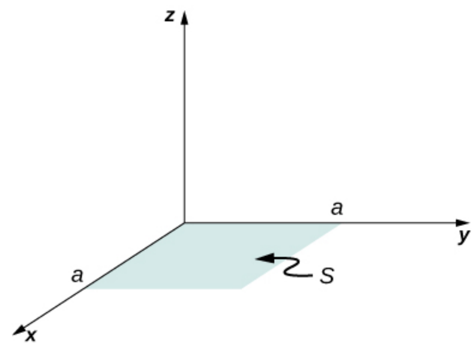
(b) What is the charge density on the outer surface of the shell?

(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 $\hat{n} = \hat{k}$.

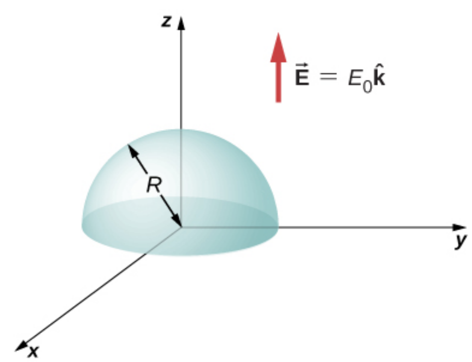


73. Repeat the preceding problem, with $\vec{E} = 2x\hat{i} + 3x^2\hat{k}$.

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 $\vec{E} = 3z^2\hat{i}$.

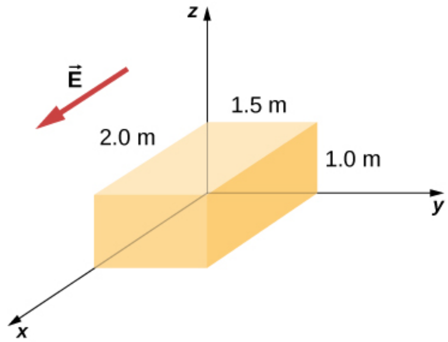
75. (a) Calculate the electric flux through the open hemispherical surface due to the electric field $\vec{E} = E_0\hat{k}$ (see below).

(b) If the hemisphere is rotated by **90°** around the **x**-axis, what is the flux through it?

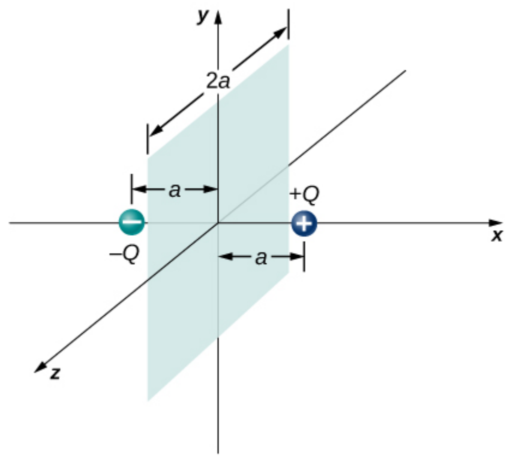


76. Suppose that the electric field of an isolated point charge were proportional to $1/r^{2+\sigma}$ rather than $1/r^2$. Determine the flux that passes through the surface of a sphere of radius **R** centered at the charge. Would Gauss's law remain valid?

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 $-30 \mu C$.

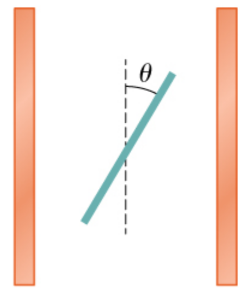
(a) Find the charge density at all surfaces, i.e., on those facing each other and those facing away.

(b) Find the electric field between the plates near the center assuming planar symmetry.

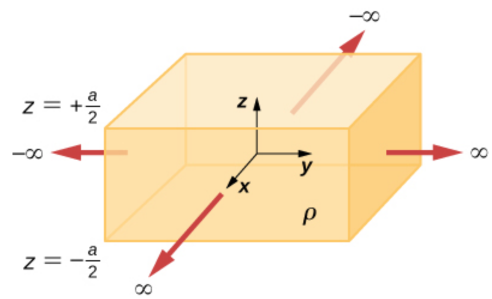
82. Two large copper plates facing each other have charge densities $\pm 4.0 C/m^2$ on the surface facing the other plate, and zero in between the plates. Find the electric flux through a **3cm×4cm** rectangular area between the plates, as shown below, for the following orientations of the area.

(a) If the area is parallel to the plates, and

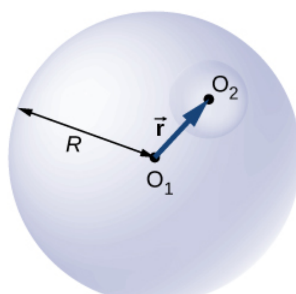
(b) if the area is tilted $\theta = 30^\circ$ from the parallel direction. Note, this angle can also be $\theta = 180^\circ + 30^\circ$.



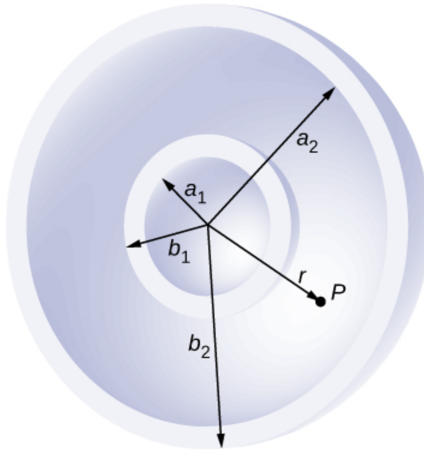
83. The infinite slab between the planes defined by $z = -a/2$ and $z = a/2$ contains a uniform volume charge density ρ(see below). What is the electric field produced by this charge distribution, both inside and outside the distribution?



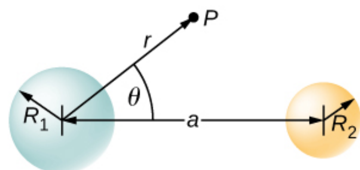
84. A total charge **Q** is distributed uniformly throughout a spherical volume that is centered at O_1 and has a radius **R**. Without disturbing the charge remaining, charge is removed from the spherical volume that is centered at O_2 (see below). Show that the electric field everywhere in the empty region is given by $\vec{E} = \frac{Q\vec{r}}{4\pi\varepsilon_0 R^3}$, 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 charged 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 **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) $r < a_1$.

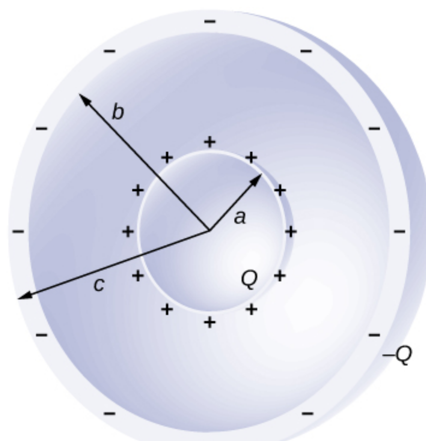


86. Two non-conducting spheres of radii R_1 and R_2 are uniformly charged with charge densities ρ_1 and ρ_2, respectively. They are separated at center-to-center distance a (see below). Find the electric field at point **P** located at a distance r from the center of sphere 1 and is in the direction 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.)

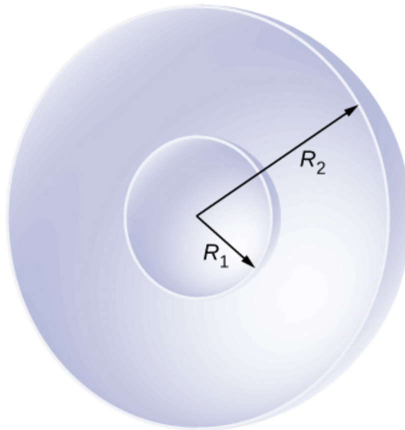


87. A disk of radius **R** is cut in a non-conducting large plate that is uniformly charged with charge density **σ** (coulomb per square meter). See below. Find the electric field at a height h above the center of the disk. ($h >> R, h << l$ or w). (**Hint:** Fill the hole with $\pm\sigma$.)

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?



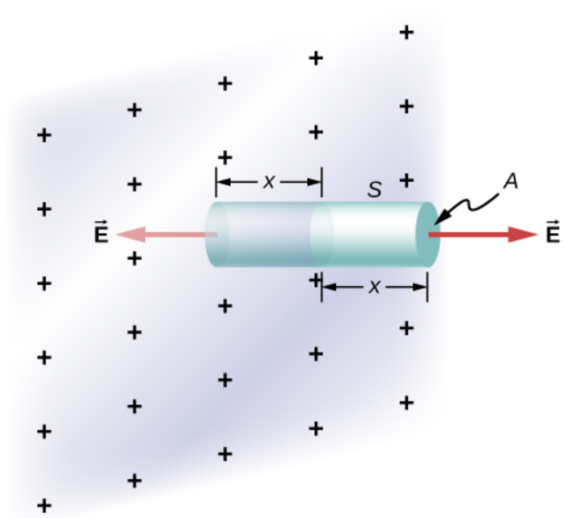
90. A point charge of q=5.0×10−8Cq=5.0×10−8C is placed at the center of an uncharged spherical conducting shell of inner radius 6.0 cm and outer radius 9.0 cm. Find the electric field at (a) r=4.0cmr=4.0cm, (b) r=8.0cmr=8.0cm, and (c) r=12.0cmr=12.0cm. (d) What are the charges induced on the inner and 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.



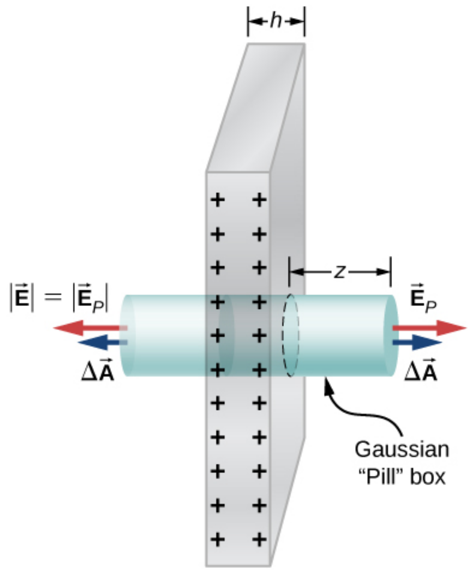
94. A spherical rubber balloon carries a total charge **Q** distributed uniformly over its surface. At $t = 0$, the radius of the balloon is **R**. The balloon is then slowly inflated until its radius reaches **2R** at the time t_0. Determine the electric field due to this charge as a function of time

(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 **q**. Let **A** be area of one side of the plate and **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.



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. $\vec{E} = \frac{\lambda_0}{2\pi\varepsilon_0}\frac{1}{d}\hat{r}$; 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 direction 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 $\frac{1}{4\pi\varepsilon_0}$, 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. $\Phi = \vec{E} \cdot \vec{A} \rightarrow EAcos\theta = 2.2 \times 10^4 N \cdot m^2/C$ electric field in direction of unit normal; $\Phi = \vec{E} \cdot \vec{A} \rightarrow EAcos\theta = -2.2 \times 10^4 N \cdot m^2/C$ 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. $\Phi = EAcos0^\circ = 1.0 \times 10^3 N/C (2.0 \times 10^{-4} m)^2 cos0^\circ = 0.20 N \cdot m^2/C$

27. $\Phi = 3.8\times10^4 N\cdot m^2/C$

29. $\vec{E}(z) = \frac{1}{4\pi\varepsilon_0}\frac{2\lambda}{z}\hat{k}, \int \vec{E}\cdot\hat{n}dA = \frac{\lambda}{\varepsilon_0}l$

31. a. $\Phi = 3.39\times10^3 N\cdot m^2/C$;

b. $\Phi = 0$;

c. $\Phi = -2.25\times10^5 N\cdot m^2/C$;

d. $\Phi = 90.4 N\cdot m^2/C$

33. $\Phi = 1.13\times10^6 N\cdot m^2/C$

35. Make a cube with **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 $\Phi = \frac{1}{24}\frac{q}{\varepsilon_0}$.

37. $q = 3.54\times10^{-7} C$

39. zero, also because flux in equals flux out

41. $r > R, E = \frac{Q}{4\pi\varepsilon_0 r^2}; r < R, E = \frac{qr}{4\pi\varepsilon_0 R^3}$

43. $EA = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = 4.50\times10^7 N/C$

45. a. 0;

b. 0;

c. $\vec{E} = 6.74\times10^6 N/C(-\hat{r})$

47. a. 0;

b. $E = 2.70\times10^6 N/C$

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. $\vec{E} = \frac{R\sigma_0}{\varepsilon_0}\frac{1}{r}\hat{r} \Rightarrow \sigma_0 = 5.31\times10^{-11} C/m^2, \lambda = 3.33\times10^{-12} C/m$;

b. $\Phi = \frac{q_{enc}}{\varepsilon_0} = \frac{3.33\times10^{-12} C/m(0.05m)}{\varepsilon+0} = 0.019 N\cdot m^2/C$

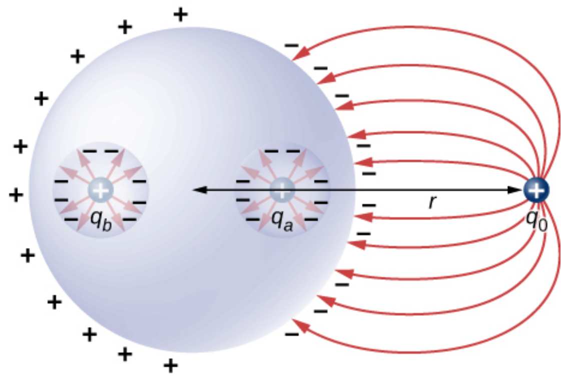
53. $E2\pi rl = \frac{\rho\pi r^2 l}{\varepsilon_0} \Rightarrow E = \frac{\rho r}{2\varepsilon_0}(r\leq R); E2\pi rl = \frac{\rho\pi R^2 l}{\varepsilon_0} \Rightarrow E = \frac{\rho R^2}{2\varepsilon_0 r}(r\geq R)$

55. $\Phi = \frac{q_{enc}}{\varepsilon_0} \Rightarrow q_{enc} = -1.0\times10^{-9} C$

57. $q_{enc} = \frac{4}{5}\pi\alpha r^5, E4\pi r^2 = \frac{4\pi\alpha r^5}{5\varepsilon_0} \Rightarrow E = \frac{\alpha r^3}{5\varepsilon_0}(r\leq R), q_{enc} = \frac{4}{5}\pi\alpha R^5, E4\pi r^2 = \frac{4\pi\alpha R^5}{5\varepsilon_0} \Rightarrow E = \frac{\alpha R^5}{5\varepsilon_0 r^2}(r\geq R)$

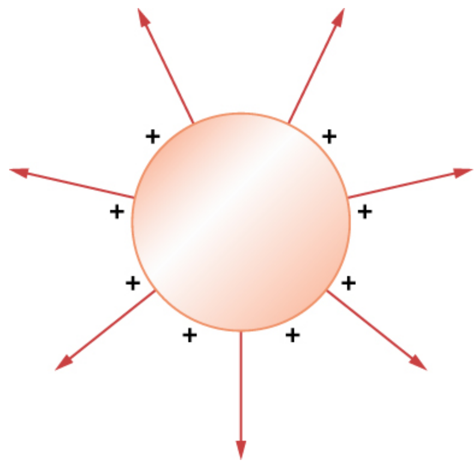
59. integrate by parts: $q_{enc} = 4\pi\rho_0[-e^{-\alpha r}(\frac{(r)^2}{\alpha}+\frac{2r}{\alpha^2}+\frac{2}{\alpha^3})+\frac{2}{\alpha^3}] \Rightarrow E = \frac{\rho_0}{r^2\varepsilon_0}[-e^{-\alpha r}(\frac{(r)^2}{\alpha}+\frac{2r}{\alpha^2}+\frac{2}{\alpha^3})+\frac{2}{\alpha^3}]$

61.



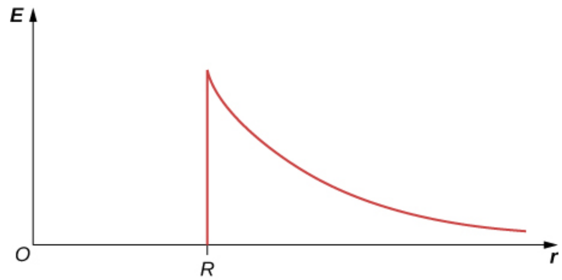
63. a. Outside: $E2\pi rl = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{3.0C/m}{2\pi\varepsilon_0 r}$; Inside $E_{in} = 0$;

b.



65. a. $E2\pi rl = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{\lambda}{2\pi\varepsilon_0 r} r \geq R$ **E** inside equals 0;

b.



67. $E = 5.65 \times 10^4 N/C$

69. $\lambda = \frac{\lambda l}{\varepsilon_0} \Rightarrow E = \frac{a\sigma}{\varepsilon_0 r} r \geq a, E = 0$ inside since q **enclosed**=0

71. a. $E = 0$;

b. $E2\pi rL = \frac{Q}{\varepsilon_0} \Rightarrow E = \frac{Q}{2\pi\varepsilon_0 rL}$; 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. $\Phi = \frac{q_{enc}}{\varepsilon_0}$; 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 $x = 2.0m$; $-E(0)[1.5m^2] + E(2.0m)[1.5m^2] = \frac{q_{enc}}{\varepsilon_0} = -100Nm2/C$

where the minus sign indicates that at $x = 0$, the electric field is along positive **x** and the unit normal is along negative **x**. At $x = 2$, the unit normal and the electric field vector are in the same direction: $q_{enc} = \varepsilon_0 \Phi = -8.85 \times 10^{-10} C$

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. $E = 3.39 \times 10^8 N/C E = 3.39 \times 10^8 N/C$

83. Construct a Gaussian cylinder along the **z**-axis with cross-sectional area **A.**

$|z| \geq \frac{a}{2} q_{enc} = \rho A a, \Phi = \frac{\rho A a}{\varepsilon_0} \Rightarrow E = \frac{\rho a}{2\varepsilon_0}$,

$|z| \leq \frac{a}{2} q_{enc} = \rho A 2z, E(2A) = \frac{\rho A 2z}{\varepsilon_0} \Rightarrow E = \frac{\rho z}{\varepsilon_0}$

85. a. $r > b_2 \;\; E 4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3 - a_1^3) + \rho_2(b_2^3 - a_2^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3) + \rho_2(b_2^3 - a_2^3)}{3\varepsilon_0 r^2}$;

b. $a_2 < r < b_2 \;\; E 4\pi r^2 = \frac{\frac{4}{3}\pi[\rho_1(b_1^3 - a_1^3) + \rho_2(r^3 - a_2^3)]}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3) + \rho_2(r^3 - a_2^3)}{3\varepsilon_0 r^2}$;

c. $b_1 < r < a_2 \;\; E 4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(b_1^3 - a_1^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(b_1^3 - a_1^3)}{3\varepsilon_0 r^2}$;

d. $a_1 < r < b_1 \;\; E 4\pi r^2 = \frac{\frac{4}{3}\pi\rho_1(r^3_{-} a_1^3)}{\varepsilon_0} \Rightarrow E = \frac{\rho_1(r^3 - a_1^3)}{3\varepsilon_0 r^2}$;

e. 0

87. Electric field due to plate without hole: $E = \frac{\sigma}{2\varepsilon_0}$.

Electric field of just hole filled with \(\displaystyle −σE=\frac{−σ}{2ε_0}(1−\frac{z}{\sqrt{R^2+z^2}}).

Thus, $E_{net} = \frac{\sigma}{2\varepsilon_0} \frac{h}{\sqrt{R^2 + h^2}}$

89. a. $E = 0$; b. $E = \frac{q_1}{4\pi\varepsilon_0 r^2}$; c. $E = \frac{q_1 + q_2}{4\pi\varepsilon_0 r^2}$; 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 **A** and **2x**, respectively, and the cylinder is positioned so that it is bisected by the plane sheet. Since **E** is perpendicular to each end and parallel to the side of the cylinder, we have **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, $2EA = \frac{\sigma A}{\varepsilon_0}$, and the electric field of an infinite sheet of charge is

$E = \frac{\sigma}{2\varepsilon_0}$, 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 $Q : \sigma = \frac{Q}{2A}$,

$$\oint_S \vec{E} \cdot \hat{n} dA = \frac{2\sigma \Delta A}{\varepsilon_0} \Rightarrow E_P = \frac{\sigma}{\varepsilon_0} = \frac{Q}{\varepsilon_0 2A}$$

CHAPTER OVERVIEW

18: Calculation of Magnetic Quantities from Currents

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.

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 **I**. What is the magnetic field at a point **P**, located a distance **R** from the wire?

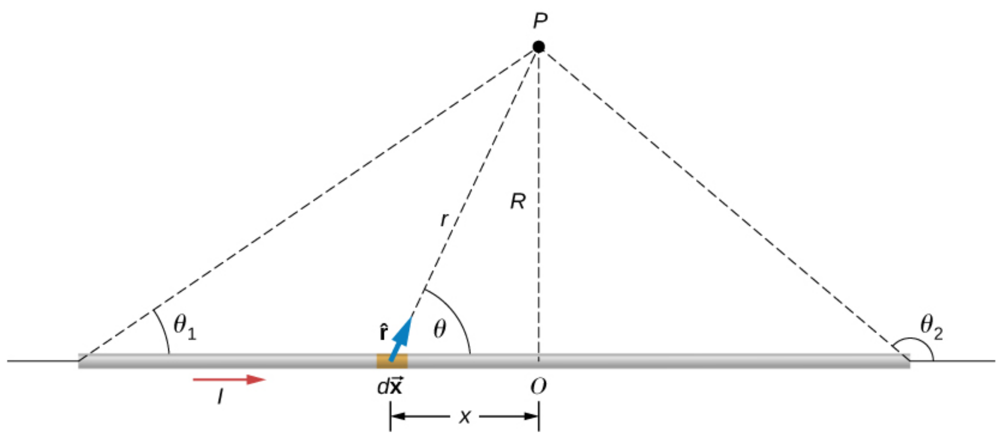


Figure 18.2.1: A section of a thin, straight current-carrying wire. The independent variable θ has the limits θ_1 and θ_2.

Let's begin by considering the magnetic field due to the current element $I\,d\vec{x}$ located at the position **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 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**

$$B = \frac{\mu_0}{4\pi}\int_{wire} \frac{I\,\sin\theta\,dx}{r^2}. \tag{18.2.1}$$

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 **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

$$\boxed{B = \frac{\mu_0 I}{2\pi R}.}$$

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 $\vec{B}$.

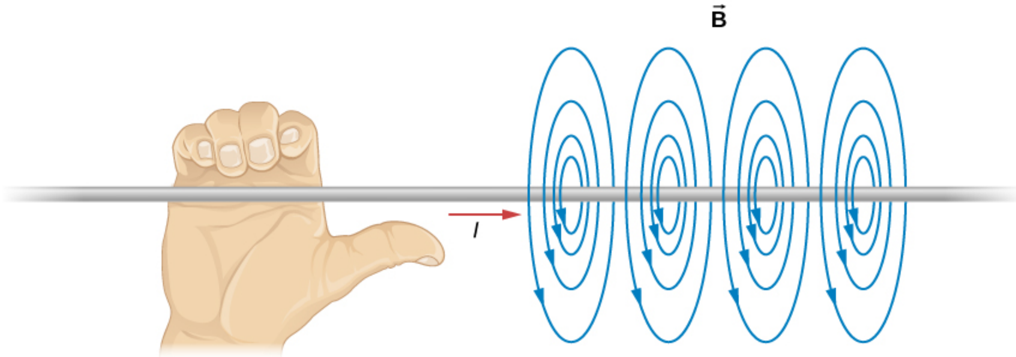


Figure 18.2.2:. Some magnetic field lines of an infinite wire. The direction of B can be found with a form of the right-hand rule.

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 $18.2.3b$

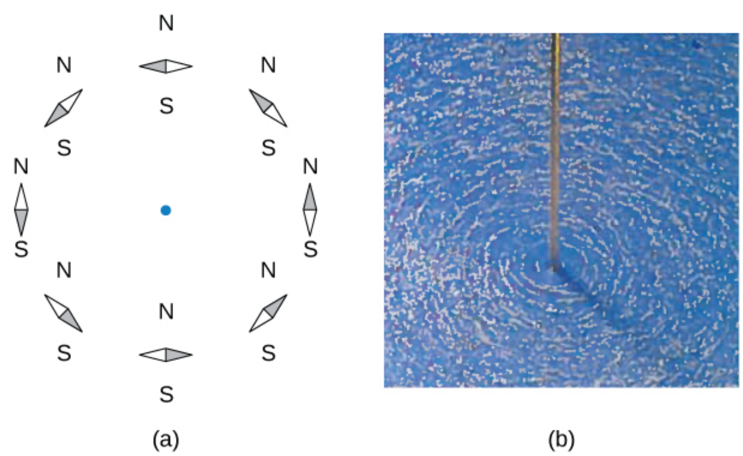


Figure 18.2.3: The shape of the magnetic field lines of a long wire can be seen using (a) small compass needles and (b) iron filings.

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.

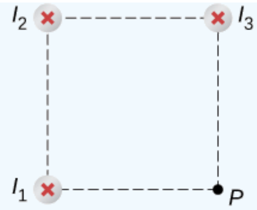


Figure 18.2.4: Three wires have current flowing into the page. The magnetic field is determined at the fourth corner of the square.

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

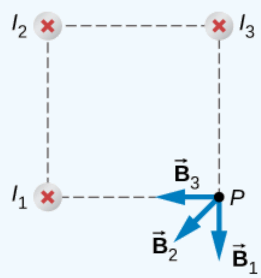
Wires 1 and 3 both have the same magnitude of magnetic field contribution at point **P**:

$$B_1 = B_3 = \frac{\mu_0 I}{2\pi R} = \frac{(4\pi \times 10^{-7} T \cdot m/A)(2\ A)}{2\pi(0.01\ m)} = 4 \times 10^{-5} T.$$

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:

$$B_{net\ x} = -4 \times 10^{-5} T - 2.83 \times 10^{-5} T\ \cos(45^o) = -6 \times 10^{-5} T.$$

The **y**-component is similarly the contributions from wire 1 and the **y**-component of wire 2:

$$B_{net\ y} = -4 \times 10^{-5} T - 2.83 \times 10^{-5} T\ \sin(45^o) = -6 \times 10^{-5} T.$$

Therefore, the net magnetic field is the resultant of these two components:

$$B_{net} = \sqrt{B^2_{net\ x} + B_{net\ y}} \tag{18.2.2}$$

$$= \sqrt{(-6 \times 10^{-5} T)^2 + (-6 \times 10^{-5} T)^2} \tag{18.2.3}$$

$$= 8.48 \times 10^{-5} T. \tag{18.2.4}$$

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

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 **R**, carries a current **I**, and lies in the **xz**-plane. What is the magnetic field due to the current at an arbitrary point **P** along the axis of the loop?

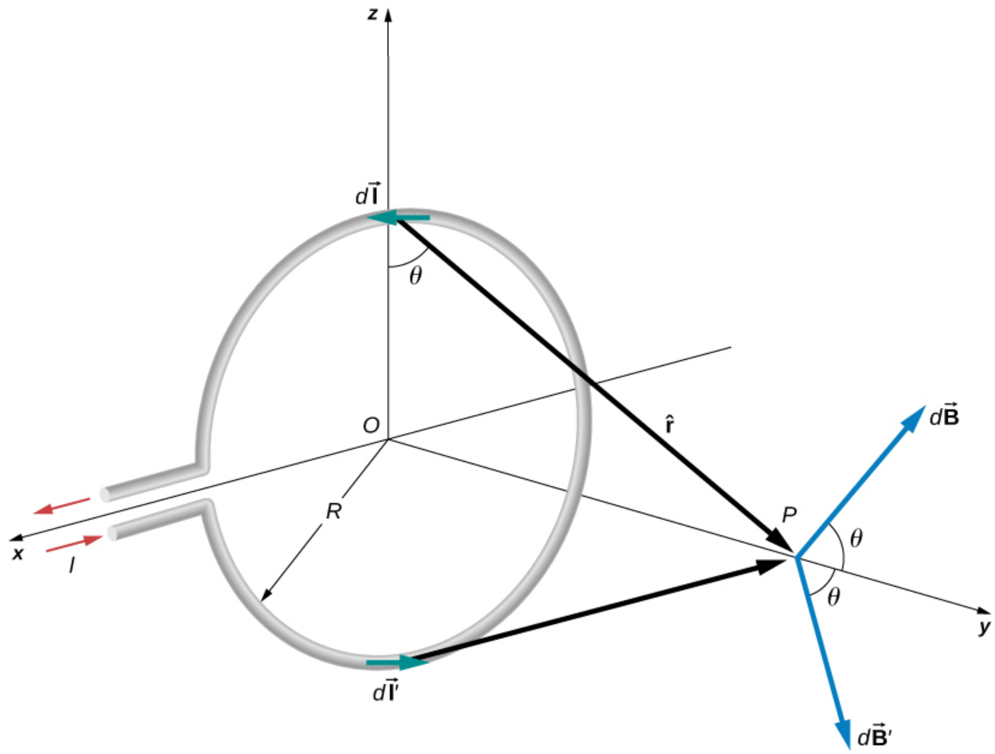


Figure 18.3.1: Determining the magnetic field at point **P** along the axis of a current-carrying loop of wire.

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 **P** be a distance **y** from the center of the loop. From the right-hand rule, the magnetic field $d\vec{B}$ at **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 $\hat{r}$ is in the **yz**-plane, the two vectors are perpendicular, so we have

$$dB = \frac{\mu_0}{4\pi}\frac{I\,dl\,sin\,\pi/2}{r^2} = \frac{\mu_0}{4\pi}\frac{I\,dl}{y^2+R^2} \tag{18.3.1}$$

where we have used $r^2 = y^2 + R^2$.

Now consider the magnetic field $d\vec{B}'$ due to the current element $I\,d\vec{l}'$, which is directly opposite $I\,d\vec{l}$ on the loop. The magnitude of $d\vec{B}'$ is also given by Equation 18.3.1, but it is directed at an angle below the **y**-axis. The components of $d\vec{B}$ and $d\vec{B}'$ 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 **P**:

$$\vec{B} = \hat{j} \int_{loop} dB\, cos\,\theta = \hat{j} \frac{\mu_0 I}{4\pi} \int_{loop} \frac{cos\,\theta\, dl}{y^2 + R^2}. \tag{18.3.2}$$

For all elements $\vec{dl}$ on the wire, **y**, **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 **P** is

$$\vec{B} = \hat{j} \frac{\mu_0 I R}{4\pi (y^2 + R^2)^{3/2}} \int_{loop} dl = \frac{\mu_0 I R^2}{2(y^2 + R^2)^{3/2}} \hat{j} \tag{18.3.3}$$

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 $\vec{\mu} = I\, A\hat{n}$. For this example, $A = \pi R^2$ and $\hat{n} = \hat{j}$, so the magnetic field at **P** can also be written as

$$\vec{B} = \frac{\mu_0 \mu \hat{j}}{2\pi (y^2 + R^2)^{3/2}}. \tag{18.3.4}$$

By setting $y = 0$ in Equation 18.3.4, we obtain the magnetic field at the center of the loop:

✓ Note

$$\vec{B} = \frac{\mu_0 I}{2R} \hat{j}. \tag{18.3.5}$$

This equation becomes $B = \mu_0 n I / (2R)$ for a flat coil of **n** loops per length. It can also be expressed as

$$\vec{B} = \frac{\mu_0 \vec{\mu}}{2\pi R^3}. \tag{18.3.6}$$

If we consider $y >> R$ in Equation 18.3.4, the expression reduces to an expression known as the magnetic field from a dipole:

$$\vec{B} = \frac{\mu_0 \vec{\mu}}{2\pi y^3}. \tag{18.3.7}$$

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.

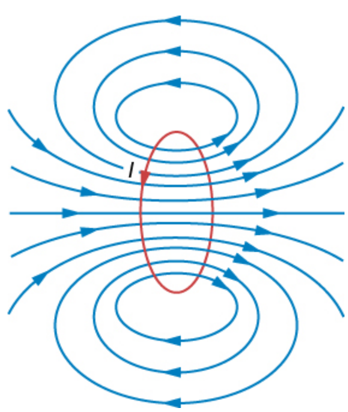


Figure 18.3.2: Sketch of the magnetic field lines of a circular current loop.

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 $R = 50\,cm$ while the other loop has a radius of $2R = 100\,cm$. 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**?

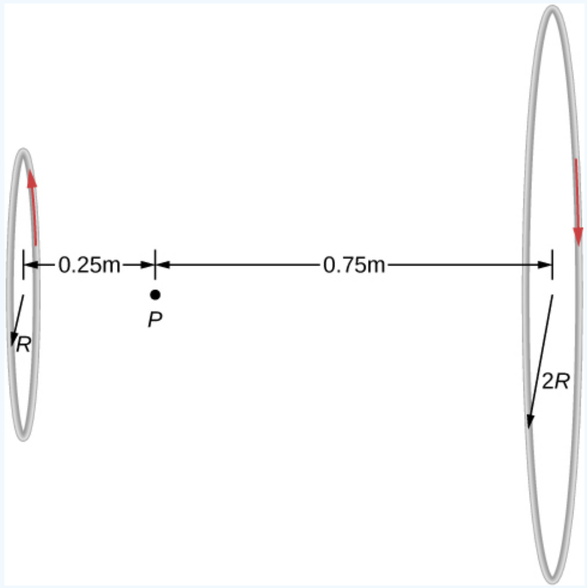


Figure 18.3.3: Two loops of different radii have the same current but flowing in opposite directions. The magnetic field at point **P** is measured to be zero.

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

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 directed counterclockwise and 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 **B** and the projection of **dl** onto the circle passing through $d\vec{l}$. If the radius of this particular circle is **r**, the projection is $r d\theta$, and

$$\vec{B} \cdot d\vec{l} = Br\, d\theta.$$

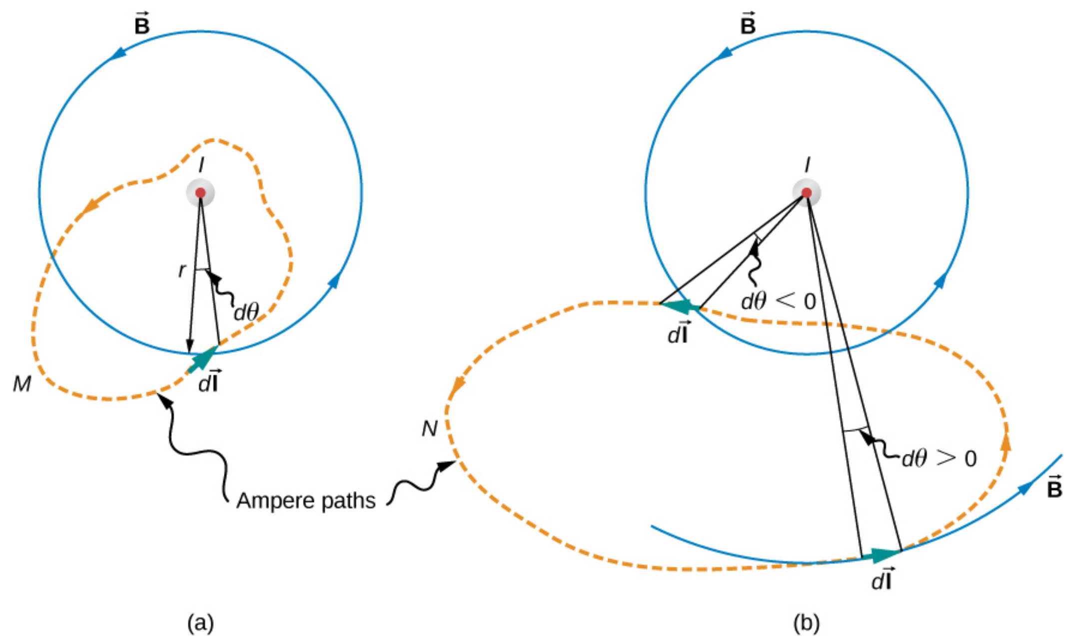


Figure 18.4.1: The current I of a long, straight wire is directed out of the page. The integral $\oint d\theta$ equals 2π and 0, respectively, for paths M and N.

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. Chose 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.

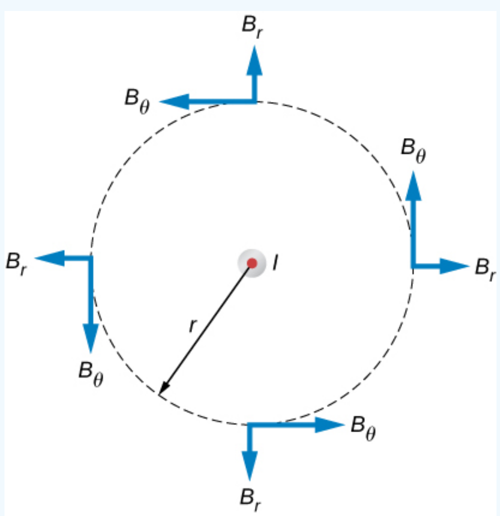


Figure 18.4.2: The possible components of the magnetic field **B** due to a current **I**, which is directed out of the page. The radial component is zero because the angle between the magnetic field and the path is at a right angle.

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 $\vec{B}_r \cdot d\vec{l} = 0$. 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.

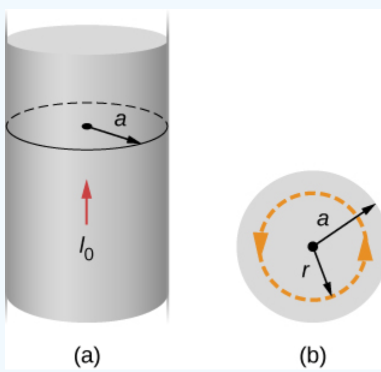


Figure 18.4.3: (a) A model of a current-carrying wire of radius **a** and current I_0. (b) A cross-section of the same wire showing the radius **a** and the Ampère’s loop of radius **r**.

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 **B** with **r** is shown in Figure 18.4.4.

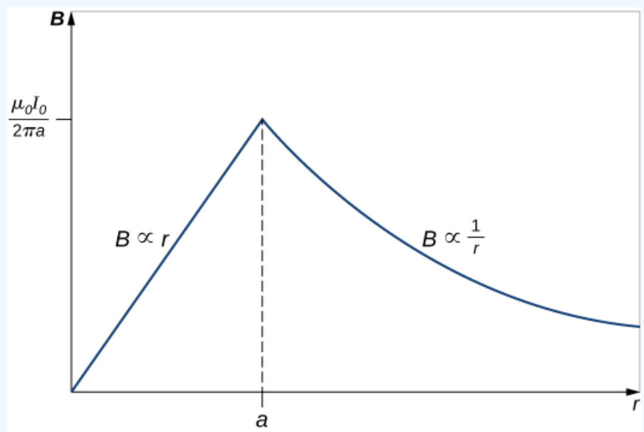


Figure 18.4.4: Variation of the magnetic field produced by a current I_0 in a long, straight wire of radius **a**.

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.

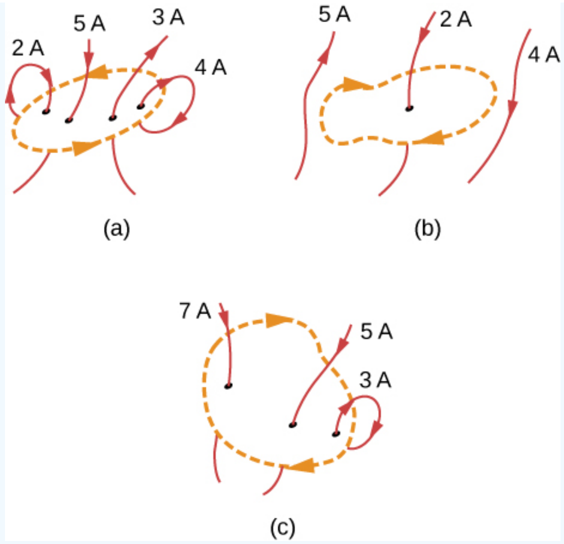


Figure 18.4.5: Current configurations and paths for Example 18.4.3.

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 $\oint \vec{B} \cdot d\vec{l} = \mu_0 (2\ A) = 2.51 \times 10^{-6} T \cdot m$.

(c) The right-hand rule shows us the current going downward through the loop is in the positive direction. There are $7A + 5A = 12A$ of current going downward and –3 A going upward. Therefore, the total current is 9 A and $\oint \vec{B} \cdot d\vec{l} = \mu_0 (9\ A) = 5.65 \times 10^{-6} T \cdot m$.

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 Ampèrian loop are very difficult because there is no symmetry, so this method would not be useful.

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

$$dI = \frac{NI}{L}dy. \tag{18.5.1}$$

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:

$$d\vec{B} = \frac{\mu_0 R^2 dI}{2(y^2+R^2)^{3/2}}\hat{j} = \left(\frac{\mu_0 I R^2 N}{2L}\hat{j}\right)\frac{dy}{(y^2+R^2)^{3/2}} \tag{18.5.2}$$

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:

$$sin\,\theta = \frac{y}{\sqrt{y^2+R^2}}. \tag{18.5.3}$$

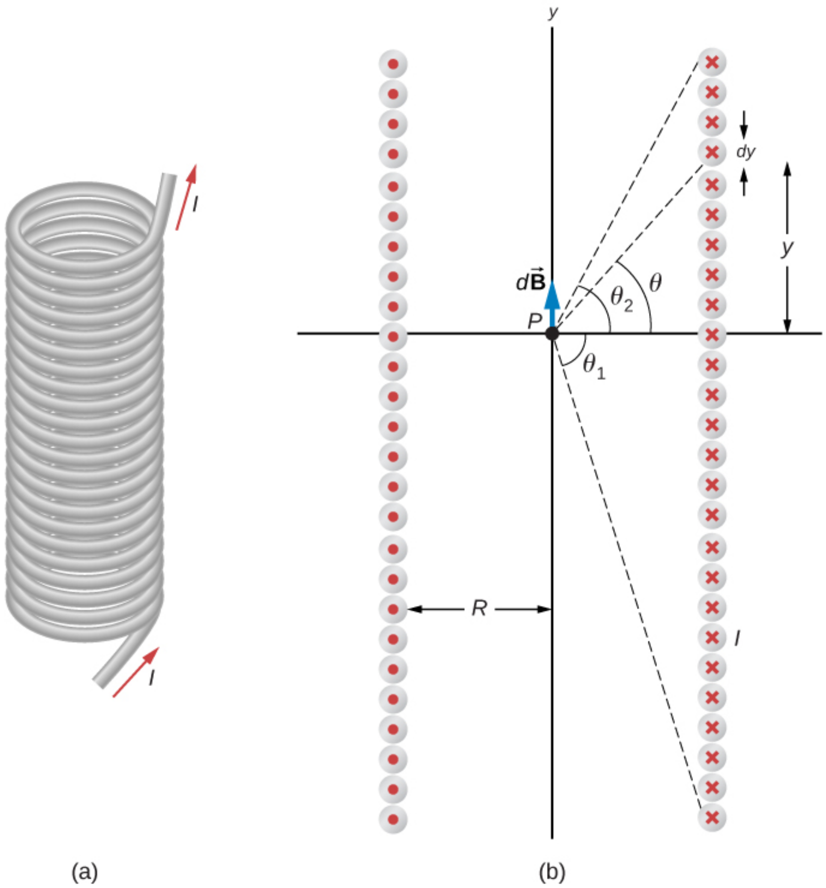


Figure 18.5.1: (a) A solenoid is a long wire wound in the shape of a helix. (b) The magnetic field at the point **P** on the axis of the solenoid is the net field due to all of the current loops.

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

$$\vec{B} = \frac{\mu_0 IN}{2L}\hat{j}\int_{\theta_1}^{\theta_2} cos\,\theta\, d\theta = \frac{\mu_0 IN}{2L}(sin\,\theta_2 - sin\,\theta_1)\hat{j}, \tag{18.5.4}$$

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 $(L >> R)$. 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

$$\vec{B} = \frac{\mu_0 IN}{2L}\hat{j}[sin(\pi/2) - sin(-\pi/2)] = \frac{\mu_0 IN}{L}\hat{j}$$

or

$$\vec{B} = \mu_0 nI\hat{j}, \tag{18.5.5}$$

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

$$\oint \vec{B} \cdot d\vec{l} = \int_1 \vec{B} \cdot d\vec{l} = Bl. \tag{18.5.6}$$

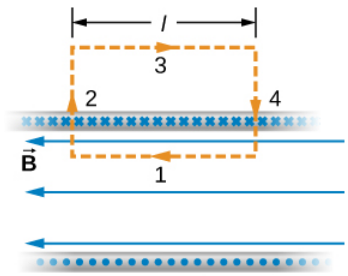


Figure 18.5.2: The path of integration used in Ampère's law to evaluate the magnetic field of an infinite solenoid.

The solenoid has **n** turns per unit length, so the current that passes through the surface enclosed by the path is **nlI**. Therefore, from Ampère's law,

$$Bl = \mu_0 nlI$$

and

Note

$$B = \mu_0 nI \tag{18.5.7}$$

within the solenoid. This agrees with what we found earlier for **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

$$n = \frac{300\, turns}{0.140\, m} = 2.14 \times 10^3\, turns/m.$$

The magnetic field produced inside the solenoid is

$$B = \mu_0 nI = (4\pi \times 10^{-7} T \cdot m/A)(2.14 \times 10^3 turns/m)(0.410\, A)$$

$$B = 1.10 \times 10^{-3} T.$$

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^o? (b) 89^o? 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?

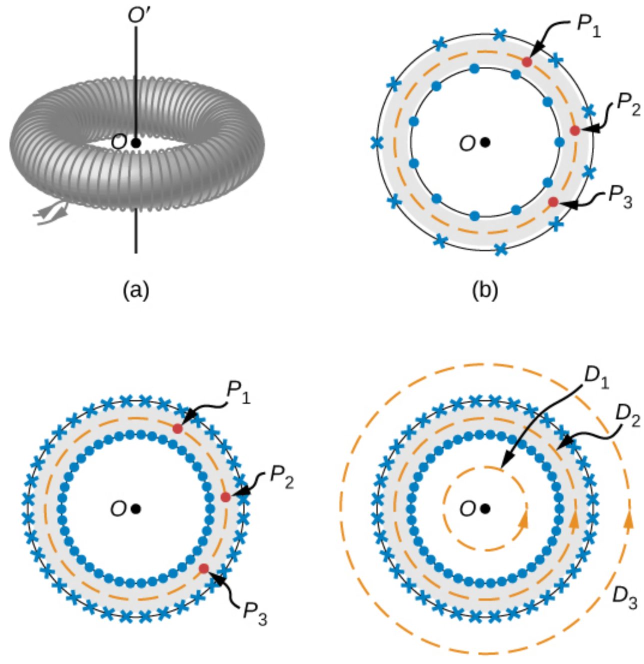


Figure 18.5.4: (a) A toroid is a coil wound into a donut-shaped object. (b) A loosely wound toroid does not have cylindrical symmetry. (c) In a tightly wound toroid, cylindrical symmetry is a very good approximation. (d) Several paths of integration for Ampère's law.

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 $18.5.4$,

$$\oint \vec{B}\cdot d\vec{l} = B(2\pi r). \tag{18.5.8}$$

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

$$B = 0\,(outside\, the\, toroid). \tag{18.5.9}$$

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

$$B=\frac{\mu_0 NI}{2\pi r}\,(within\,the\,toroid). \tag{18.5.10}$$

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 **OO'**. However, if the central radius **R** (the radius midway between the inner and outer radii of the toroid) is much larger than the cross-sectional diameter of the coils **r**, the variation is fairly small, and the magnitude of the magnetic field may be calculated by Equation 18.5.10 where $r=R$.

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.

The force between two long, straight, and parallel conductors separated by a distance **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 F_2). The field due to I_1 at a distance **r** is

$$B_1 = \frac{\mu_0 I_1}{2\pi r}$$

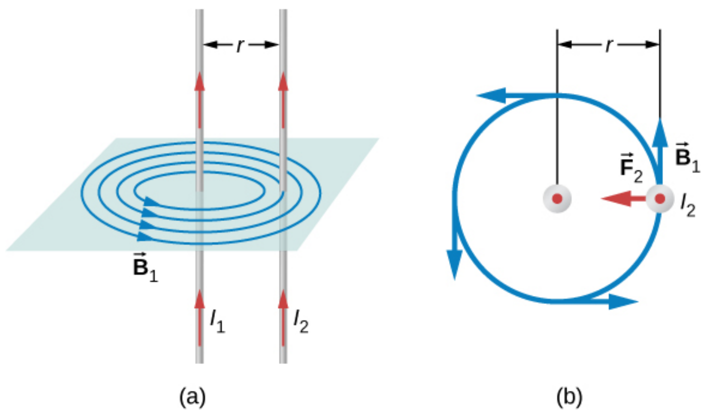


Figure 18.6.1: (a) The magnetic field produced by a long straight conductor is perpendicular to a parallel conductor, as indicated by right-hand rule (RHR)-2. (b) A view from above of the two wires shown in (a), with one magnetic field line shown for wire 1. RHR-1 shows that the force between the parallel 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.

This field is uniform from the wire 1 and perpendicular to it, so the force F_2 it exerts on a length **l** of wire 2 is given by $F = IlB\,sin\,\theta$ with $sin\,\theta = 1$:

$$F_2 = I_2 l B_1. \tag{18.6.1}$$

The forces on the wires are equal in magnitude, so we just write **F** for the magnitude of F_2 (Note that $\vec{F}_1 = -\vec{F}_2$.) Since the wires are very long, it is convenient to think in terms of **F/l**, the force per unit length. Substituting the expression for B_1 into Equation 18.6.1 and rearranging terms gives

Note

$$\frac{F}{l} = \frac{\mu_0 I_1 I_2}{2\pi r}. \tag{18.6.2}$$

The ratio **F/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?

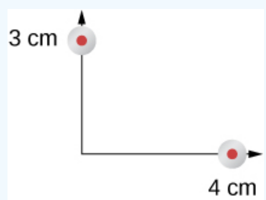


Figure 18.6.2: Two current-carrying wires at given locations with currents out of the page.

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:

$$r = \sqrt{(3.0\,cm)^2 + (4.0\,cm)^2} = 5.0\,cm.$$

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

$$\theta = tan^{-1}\left(\frac{3\,cm}{4\,cm}\right) = 36.9^o.$$

The unit vector for this is calculated by

$$cos(36.9^o)\hat{i} - sin(36.9^o)\hat{j} = 0.8\hat{i} - 0.6\hat{j}.$$

Therefore, the force per unit length from wire one on wire 2 is

$$\frac{\vec{F}}{l} = (1 \times 10^{-10}\ N/m) \times (0.8\hat{i} - 0.6\hat{j}) = (8 \times 10^{-11}\hat{i} - 6 \times 10^{-11}\hat{j})\ N/m.$$

The force per unit length from wire 2 on wire 1 is the negative of the previous answer:

$$\frac{\vec{F}}{l} = (-8 \times 10^{-11}\hat{i} + 6 \times 10^{-11}\hat{j}) N/m.$$

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 $9.23 \times 10^{-12}\ N/m$

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 and 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.

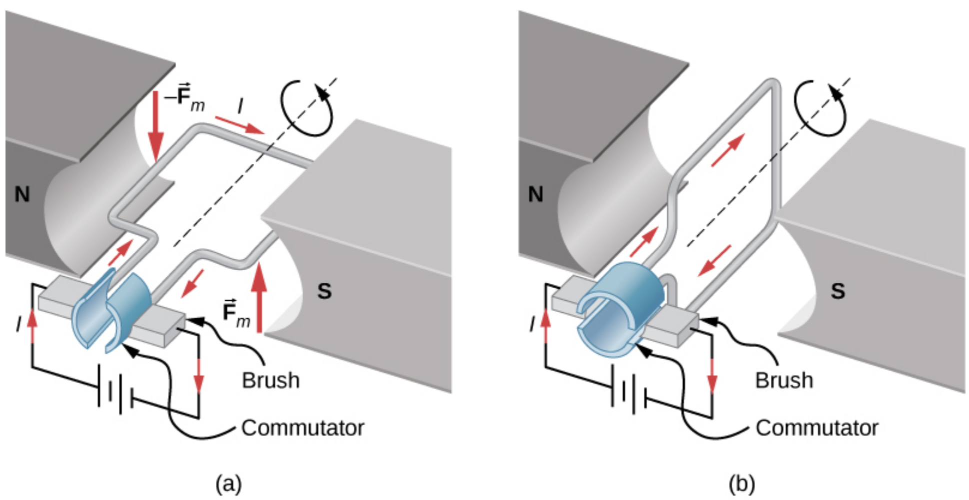


Figure 18.7.1: A simplified version of a dc electric motor. (a) The rectangular wire loop is placed in a magnetic field. The forces on the wires closest to the magnetic poles (N and S) are opposite in direction as determined by the right-hand rule-1. Therefore, the loop has a net torque and rotates to 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

$$\vec{F}_1 = IaB\sin(90^o - \theta)\hat{i} = IaB\cos\theta\hat{i}$$

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

$$\vec{F}_3 = -IaB\sin(90^o + \theta)\hat{i} = -IaB\cos\theta\hat{i}$$

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.

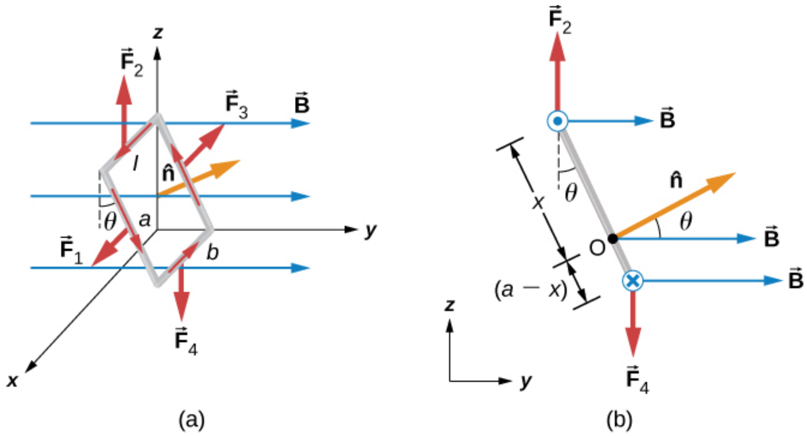


Figure 18.7.2: (a) A rectangular current loop in a uniform magnetic field is subjected to a net torque but not a net force. (b) A side view of the coil.

To find the net torque on the current loop shown in Figure $18.7.2a$, we first consider F_1 and 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 F_2 and 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 F_2 and F_4 are $x\ sin\,\theta$ and $(a-x)\ sin\,\theta$, respectively, so the net torque on the loop is

$$\sum \vec{\tau} = \vec{\tau}_1 + \vec{\tau}_2 + \vec{\tau}_3 + \vec{\tau}_4 = F_2 x\ sin\,\theta\hat{i} - F_4(a-x)\ sin\,(\theta)\hat{i}$$

$$-IbBx\ sin\,\theta\hat{i} - IbB(a-x) sin\,\theta\hat{i}.$$

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** $\boldsymbol{\mu}$**.** 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:

$$\tau = \vec{\mu} \times \vec{B} = \mu B \,sin\, \theta = (2.5 \times 10^{-6} A \cdot m^2)(0.50 T) sin(30^o) = 6.3 \times 10^{-7} N \cdot m$$

$$U = -\vec{\mu} \cdot \vec{B} = -\mu B cos \theta = -(2.5 \times 10^{-6} A \cdot m^2)(0.50 T) cos(30^o) = -1.1 \times 10^{-6} J.$$

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

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 **l** 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 **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:

$$eE = ev_dB \tag{18.8.1}$$

where **e** is the magnitude of the electron charge, v_d is the drift speed of the electrons, and **E** is the magnitude of the electric field created by the separated charge. Solving this for the drift speed results in

$$v_d = \frac{E}{B}. \tag{18.8.2}$$

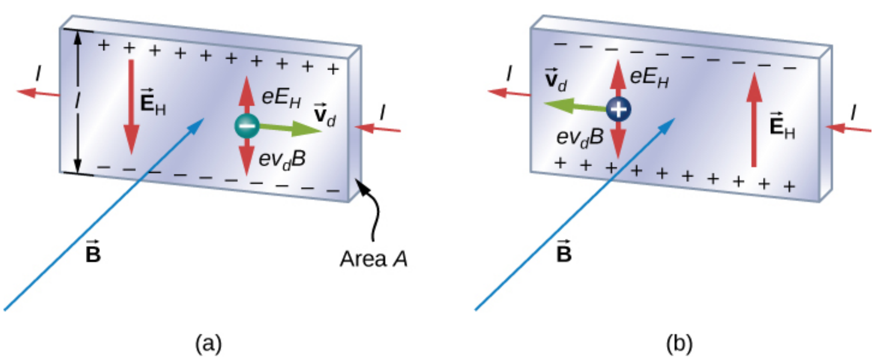


Figure 18.8.1: In the Hall effect, a potential difference between the top and bottom edges of the metal strip is produced when moving charge carriers are deflected by the magnetic field. (a) Hall effect for negative charge carriers; (b) Hall effect for positive charge carriers.

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 **I**, then from Current and Resistance, we know that

$$I = nev_dA \tag{18.8.3}$$

where **n** is the number of charge carriers per volume and **A** is the cross-sectional area of the strip. Combining the equations for v_d and **I** results in

$$I = ne\left(\frac{E}{B}\right)A. \tag{18.8.4}$$

The field **E** is related to the potential difference **V** between the edges of the strip by

$$E = \frac{V}{l}. \tag{18.8.5}$$

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

$$V = \frac{IBl}{neA} \tag{18.8.6}$$

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 $6.0 \times 10^3\ N/C$, 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:

$$v_d = \frac{E}{B} = \frac{6 \times 10^3 N/C}{2 \times 10^{-3} T} = 3 \times 10^6 m/s.$$

2. The acceleration is calculated from the net force from the magnetic field, equal to mass times acceleration. The magnitude of the acceleration is:

$$ma = qvB$$

$$a = \frac{qvB}{m} = \frac{(1.6 \times 10^{-19} C)(3 \times 10^6 m/s)(2 \times 10^{-3} T)}{0.1 \times 10^{-31} kg} = 1.1 \times 10^{15} m/s^2.$$

3. The radius of the path comes from a balance of the circular and magnetic forces, or Equation 18.8.2:

$$r = \frac{mv}{qB} = \frac{(9.1 \times 10^{-31} kg)(3 \times 10^6 m/s)}{(1.6 \times 10^{-19} C)(2 \times 10^{-3} T)} = 8.5 \times 10^{-3} m.$$

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.

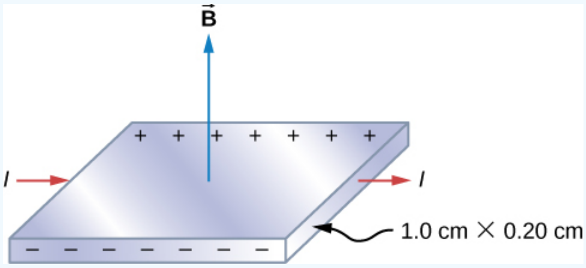


Figure 18.8.2: Finding the Hall potential in a silver ribbon in a magnetic field is shown.

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:

$$\begin{aligned} v &= \frac{IBl}{neA} \\ &= \frac{(100\,A)(1.5\,T)(1.0 \times 10^{-2} m)}{(5.9 \times 10^{28}/m^3)(1.6 \times 10^{-19} C)(2.0 \times 10^{-5} m^2)} \\ &= 7.9 \times 10^{-6} V. \end{aligned}$$

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 A and the Hall potential is

a. $4.0\,\mu V$ and
b. $6.0\,\mu V$?

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.

$$r=\frac{mv}{qB} \tag{18.9.1}$$

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.

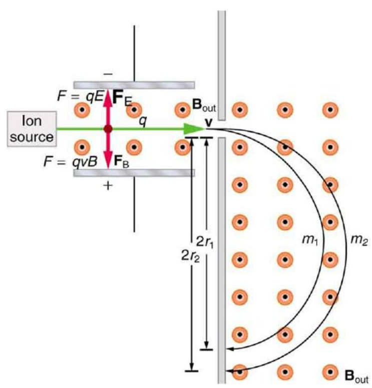


Figure 18.9.1: This mass spectrometer uses a velocity selector to fix v so that the radius of the path is proportional to mass.

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 $F=qE$ equals the magnetic force $F=qvB$, so that $qE=qvB$. Noting that q cancels, we see that

$$v=\frac{E}{B} \tag{18.9.2}$$

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.

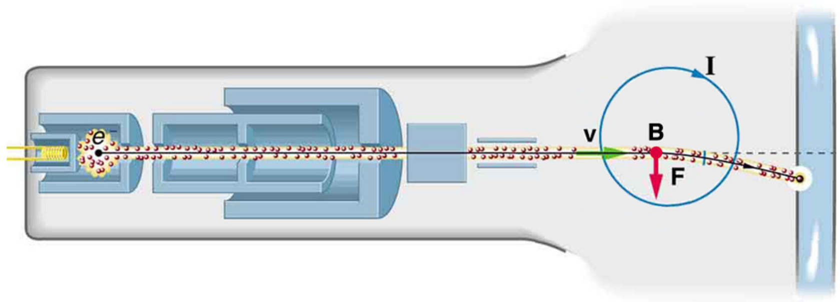


Figure 18.9.2: The cathode ray tube (CRT) is so named because rays of electrons originate at the cathode in the electron gun. Magnetic coils are used to steer the beam in many CRTs. In this case, the beam is moved down. Another pair of horizontal coils would steer the beam horizontally.

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 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

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 $4.2\,K\,(-269.2^oC)$, 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.

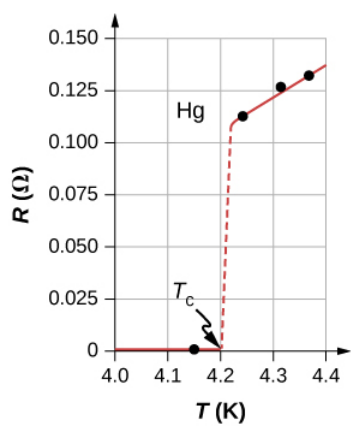


Figure 18.10.1: The resistance of a sample of mercury is zero at very low temperatures—it is a superconductor up to the temperature of about 4.2 K. Above that critical temperature, its resistance makes a sudden jump and then increases nearly linearly with temperature.

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 $T_c = 16\,K\,(-257^oC)$ and in 1953, vanadium-silicon was found to become superconductive at $T_c = 17.5\,K\,(-255.7^oC)$. 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 $T_c = 17.5\,K$ 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 $T_c = 92\,K\,(-181^oC)$. 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).

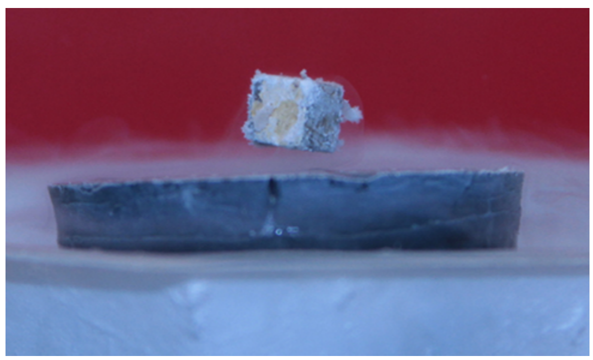


Figure 18.10.1: A small, strong magnet levitates over a superconductor cooled to liquid nitrogen temperature. The magnet levitates because the superconductor excludes magnetic fields.

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 $T_c(K)$	Critical Magnetic Field $\boldsymbol{H_c(T)}$	Type
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
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				

Material	Symbol or Formula	Critical Temperature $T_c(K)$	Critical Magnetic Field $H_c(T)$	Type
	$HgBa_2CaCu_2O_8$	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.

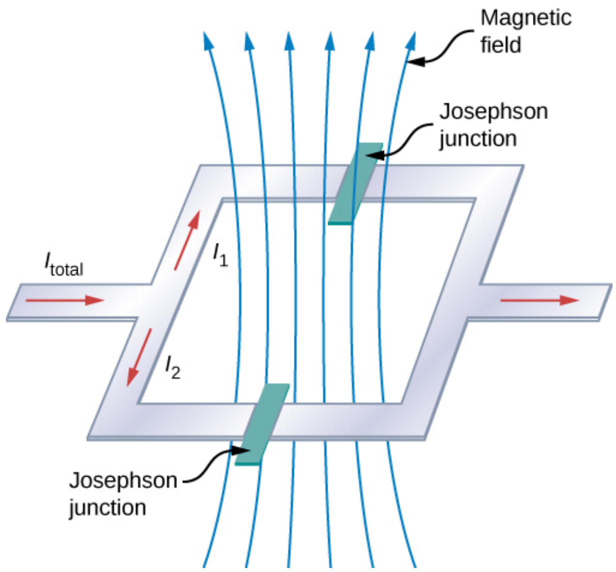


Figure 18.10.3: The SQUID (superconducting quantum interference device) uses a superconducting current loop and two Josephson junctions to detect magnetic fields as low as 10^{-14} (Earth's magnet field is on the order of $0.3 \times 10^{-5} T$).

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.

18.11: Conclusion

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 **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	$F = qvBsin\theta$
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.

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	$dB = \frac{\mu_0}{4\pi} \frac{I dl sin\theta}{r^2}$
Biot–Savart law	$\vec{B} = \frac{\mu_0}{4\pi} \int_{wire} \frac{I d\vec{l} \times \hat{r}}{r^2}$
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	$B=\frac{\mu_o NI}{2\pi r}$
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 $\mu_0=4\pi\times 10^{-7} T\cdot m/s$ 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

  $$B=\mu_0 nI \text{ (inside a solenoid)}$$

  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

  $$B=\frac{\mu_o NI}{2\pi r} \text{ (within the toroid)}$$

  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.

18.14: Current and Resistance (Summary)

Key Terms

ampere (amp)	SI unit for current; $1 A=1 C/s$
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, $1 \Omega=1 V/A$
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 $1A = 1C/s$.
- 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 ⋅ h).
- The total energy used over a time interval can be found by $E = \int P dt$.

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.

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 accidently 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?

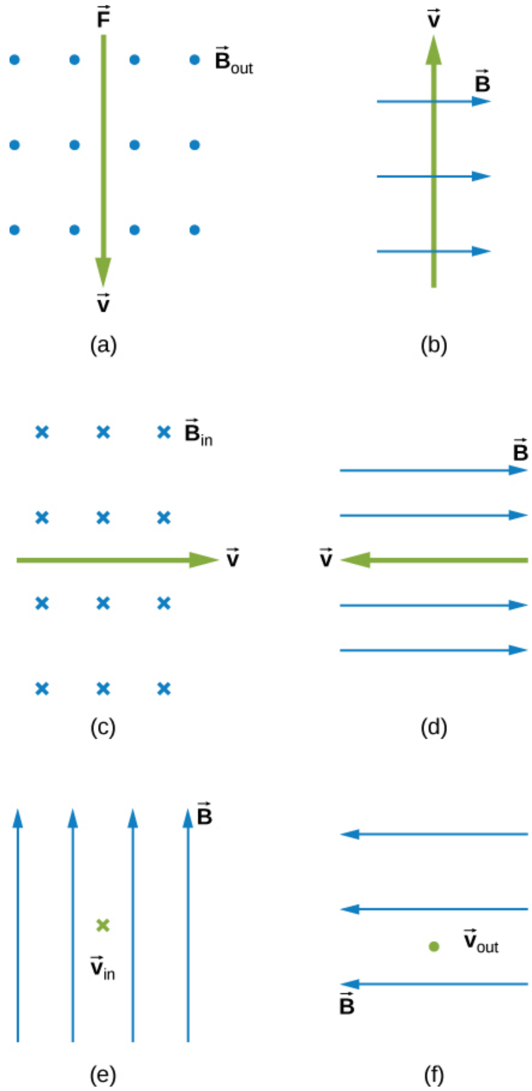
11.8 Applications of Magnetic Forces and Fields

14. Describe the primary function of the electric field and the magnetic field in a cyclotron.

Problems

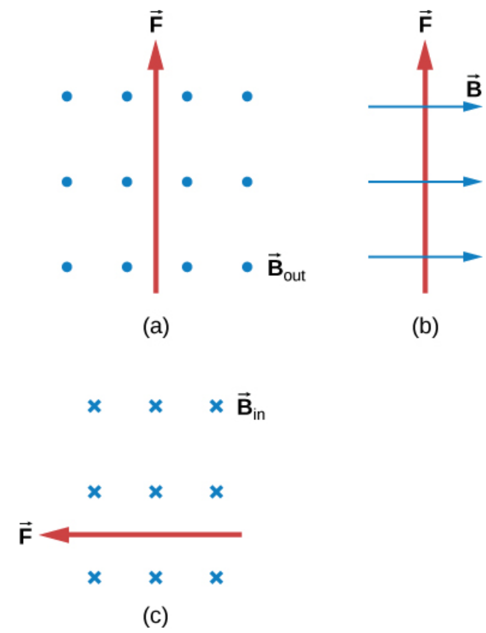
11.3 Magnetic Fields and Lines

15. What is the direction of the magnetic force on a positive charge that moves as shown in each of the six cases?



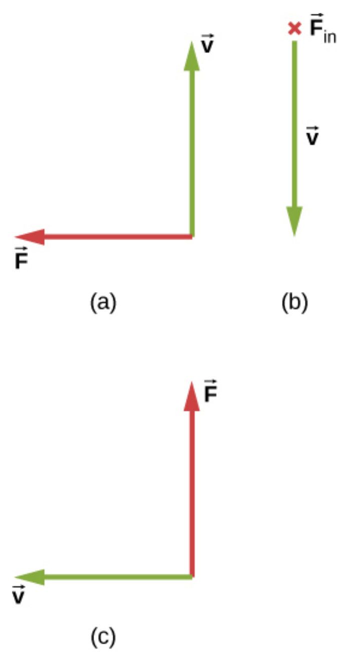
16. Repeat previous exercise for a negative charge.

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 $\vec{\mathbf{B}}$ is perpendicular to $\vec{\mathbf{v}}$?



20. Repeat previous exercise for a negative charge.

21. (a) Aircraft sometimes acquire small static charges. Suppose a supersonic jet has a 0.500-μC charge and flies due west at a speed of 660. m/s over Earth's south magnetic pole, where the $8.00 \times 10^{-5} - T$ 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 $5.00 \times 10^{7} m/s$ experiences a magnetic force of $1.70 \times 10^{-16} N$. What is the strength of the magnetic field if there is a 45º 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 $4.00 \times 10^3 m/s$ in a 1.25-T magnetic field experiences a magnetic force of $1.40 \times 10^{-16} N$. 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 $1.00 \times 10^{-12} N$. 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 $7.5 \times 10^6 m/s$ perpendicular to Earth's magnetic field at an altitude where the field strength is $1.0 \times 10^{-5} T$. 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 $5.0 \times 10^7 m/s$ 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 $2.66 \times 10^{-26} kg$ travels at $5.0 \times 10^6 m/s$ 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 $6.0 \times 10^7 m/s$, in a direction perpendicular to Earth's field, which has a strength of $5.0 \times 10^{-5} T$. (a) What strength electric field must be applied perpendicular to the Earth's field to make the electron moves 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 $6.00 \times 10^{-7} m/s$? (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 ($m = 6.64 \times 10^{-27} kg, q = 3.2 \times 10^{-19} C$) 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?

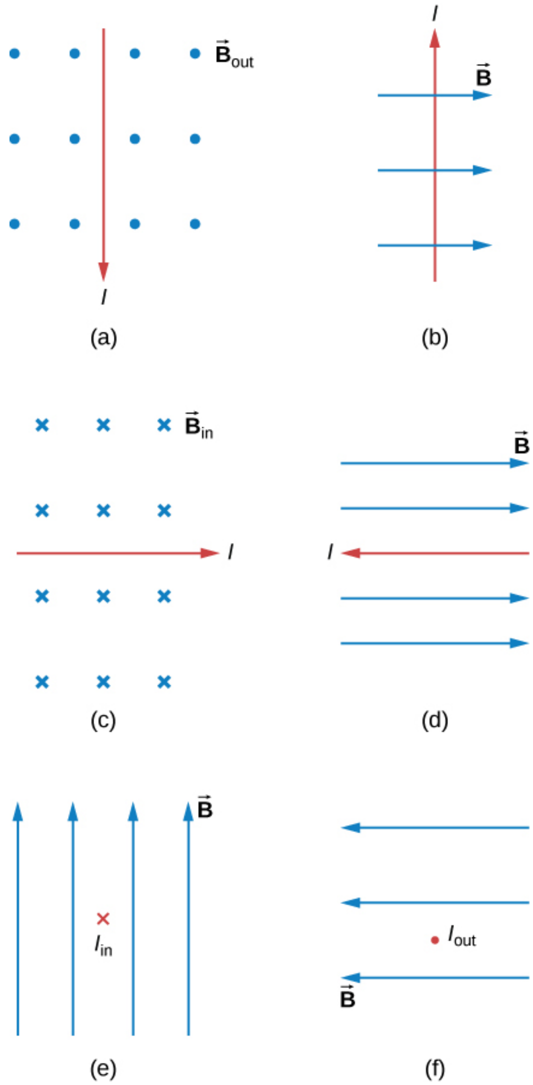
(b) What is the kinetic energy in electron-volts?

(c) Through what potential difference must the particle be accelerated in order to give it this kinetic energy?

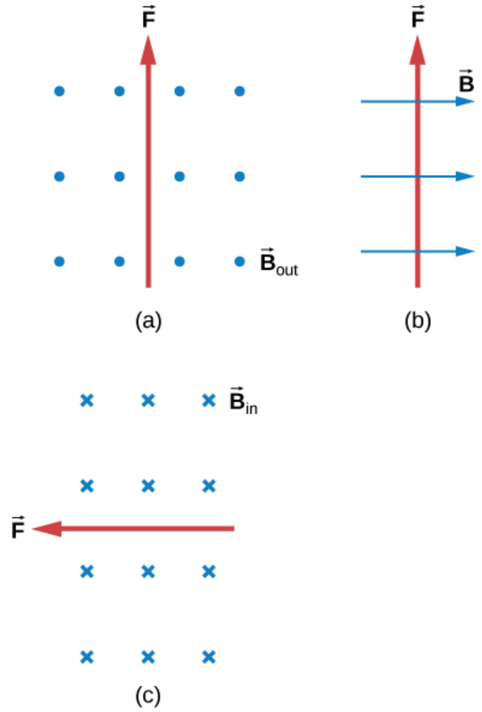
32. A particle of charge **q** and mass **m** is accelerated from rest through a potential difference **V**, after which it encounters a uniform magnetic field **B**. If the particle moves in a plane perpendicular to **B**, what is the radius of its circular orbit?

11.5 Magnetic Force on a Current-Carrying Conductor

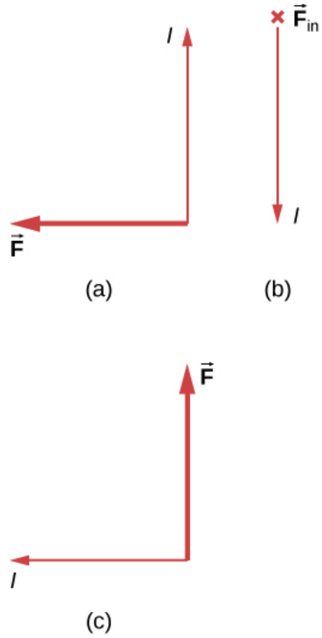
33. What is the direction of the magnetic force on the current in each of the six cases?



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 **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 30.0º 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 10.9º?

41. Find the current through a loop needed to create a maximum torque of **9.0N·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 **τ = NIAB** sin **θ**, 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?

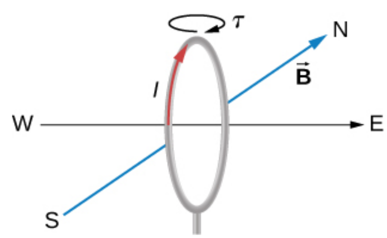
(b) 50.0% of maximum?

(c) 10.0% of maximum?

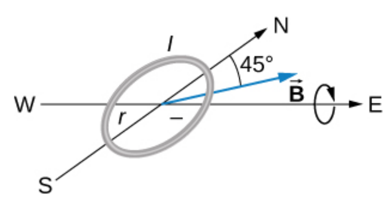
45. A proton has a magnetic field due to its spin. The field is similar to that created by a circular current loop $0.65 \times 10^{-15} m$ in radius with a current of $1.05 \times 10^4 A$. Find the maximum torque on a proton in a 2.50-T field. (This is a significant torque on a small particle.)

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

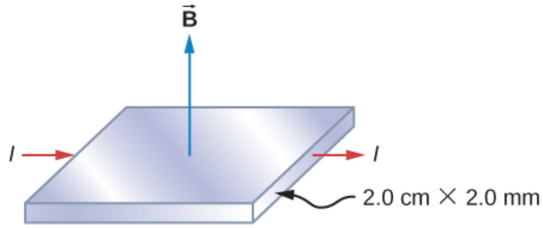
48. A strip of copper is placed in a uniform magnetic field of magnitude 2.5 T. The Hall electric field is measured to be $1.5 \times 10^{-3} V/m$.

(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 $_{s}O^{16}$ 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 **1.5μV** 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 **2μV** 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 $B_o = 0.750T$. 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 $3.0 \times 10^5 m/s$ 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.

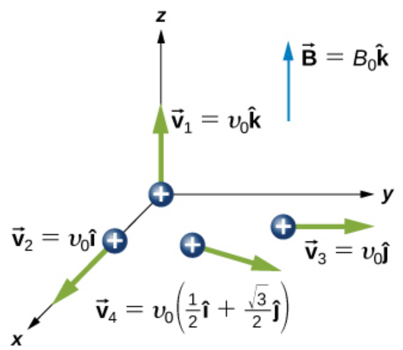
58. Calculate the magnetic force on a hypothetical particle of charge $1.0 \times 10^{-19} C$ moving with a velocity of $6.0 \times 10^4 \hat{i} m/s$ in a magnetic field of $1.2\hat{k} T$.

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 $\vec{v} = (4.0\hat{i}\,\hat{} + 3.0\hat{j}\,\hat{} + 2.0\hat{k}) \times 10^6 m/s$ 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 x 10^{-5} T north and a vertical component of 5.0 x 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 $2.0 \times 10^{-3} s$ to complete eight revolutions in a uniform magnetic field of magnitude $2.0 \times 10^{-2} T$. What is the mass of the ion?

72. A particle moving downward at a speed of $6.0 \times 10^6 m/s$ enters a uniform magnetic field that is horizontal and directed from east to west.

(a) If the particle is deflected initially to the north in a circular arc, is its charge positive or negative?

(b) If B = 0.25 T and the charge-to-mass ratio (q/m) of the particle is $4.0 \times 10^7 C/kg$, what is the radius of the path?

(c) What is the speed of the particle after it has moved in the field for $1.0 \times 10^{-5} s$? 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 $7.5 \times 10^{-2} T$ completes 10 revolutions in $3.47 \times 10^{-4} s$. 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 $\vec{v} = (4\hat{j} + 3\hat{k}) \times 10^6 m/s$ and travels in a helical path with a radius of 5.0 cm.

(a) What is the value of B?

(b) What is the time required for one trip around the helix?

(c) Where is the proton $5.0 \times 10^{-7} s$ after entering the field?

77. An electron moving at $5.0 \times 10^6 m/s$ enters a magnetic field that makes a 75^o angle with the x-axis of magnitude 0.20 T. Calculate the

(a) pitch and

(b) 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 60º with Earth's $5.5 \times 10^{-5} T$ field. What is the current when the wire experiences a force of $7.0 \times 10^{-3} N$?

(b) 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?

(b) What is the force on the wire if it is rotated to make an angle of 90º 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 $(3.0\hat{i} \times 4.0\hat{k}) \times 10^{-3} T$. 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 $8.0 \times 10^{-3} T$. Calculate the magnitude of the force on the section if the angle between the field and the direction of the current is

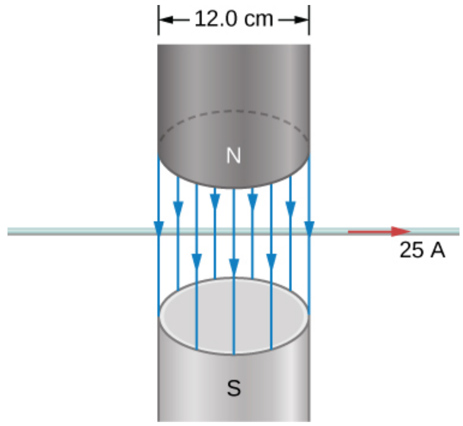
(a) 45°;

(b) 90°;

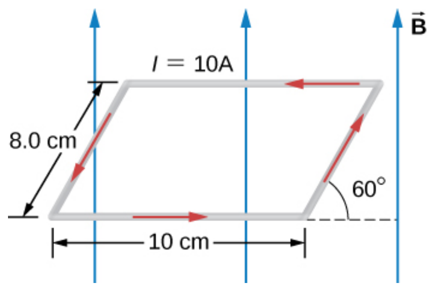
(c) 0°; or

(d) 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 A** and **B = 1.5 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 Am2?

(b) What is the maximum torque that the coil will experience in a uniform field of strength $5.0 \times 10^{-2} T$?

(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 **μV** 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 8.70×105m/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?

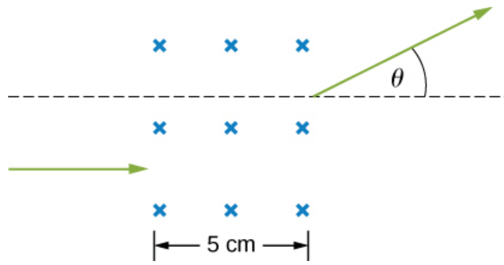
(c) Which premise is responsible?

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.

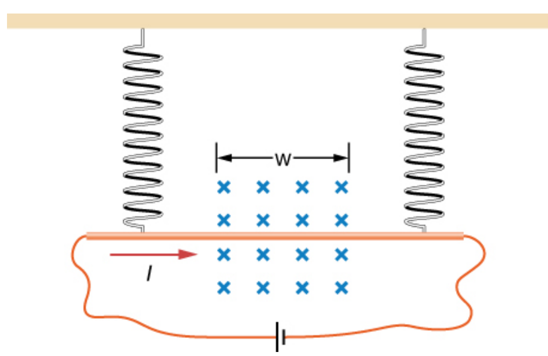
103. A proton of speed $v = 6 \times 10^5 m/s$ enters a region of uniform magnetic field of **B = 0.5 T** at an angle of $q = 30^\circ$ to the magnetic field. In the region of magnetic field proton describes a helical path with radius R and pitch p (distance between loops). Find **R** and **p**.

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 $4 \times 10^6 m/s$ 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.)

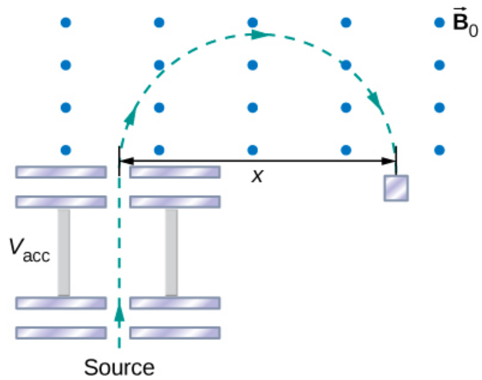


105. In a region a non-uniform magnetic field exists such that $B_x = 0, B_y = 0,$ and $B_z = ax$, where a is a constant. At some time **t**, a wire of length **L** is carrying a current **I** is located along the **x-axis** from origin to **x = L**. Find the magnetic force on the wire at this instant in time.

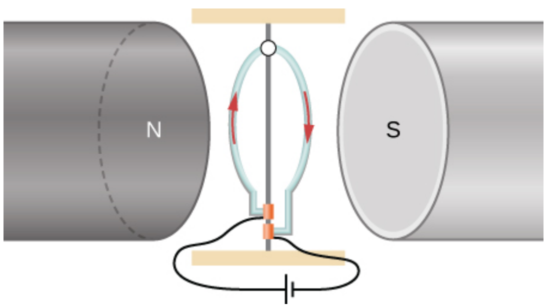
106. A copper rod of mass **m** and length **L** is hung from the ceiling using two springs of spring constant **k**. A uniform magnetic field of magnitude B_0 pointing perpendicular to the rod and spring (coming out of the page in the figure) exists in a region of space covering a length **w** of the 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 **B** in millitesla. Calculate the magnetic force on the segment of wire between **x = 2.0 m** and **x = 4.0 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 $\vec{B} = (2.0\hat{i} + 6.0\hat{j} + 8.0\hat{k}) \times 10^{-3} T$. 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?

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 $\vec{B}=0$ inside a long, hollow copper pipe that is carrying an electric current parallel to the axis. Is $\vec{B}=0$ outside the pipe?

12.8 Magnetism in Matter

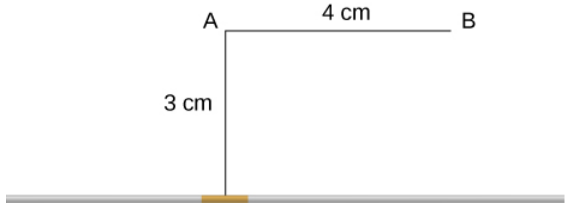
14. A diamagnetic material is brought close to a permanent magnet. What happens to the material?

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.

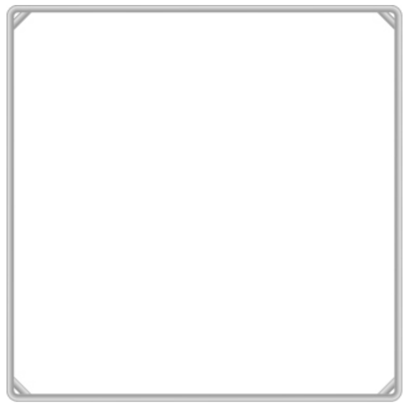
Problems

12.2 The Biot-Savart Law

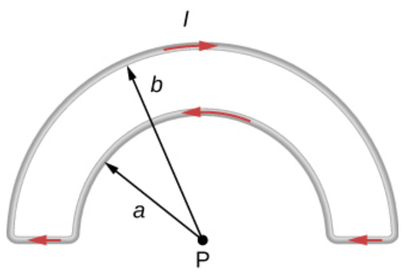
16. A 10-A current flows through the wire shown. What is the magnitude of the magnetic field due to a 0.5-mm segment of wire as measured at (a) point A and (b) point B?



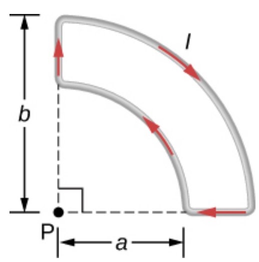
17. Ten amps flow through a square loop where each side is 20 cm in length. At each corner of the loop is a 0.01-cm segment that connects the longer wires as shown. Calculate the magnitude of the magnetic field at the center of the loop.



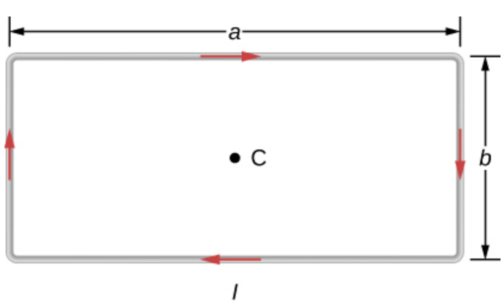
18. What is the magnetic field at P due to the current I in the wire shown?



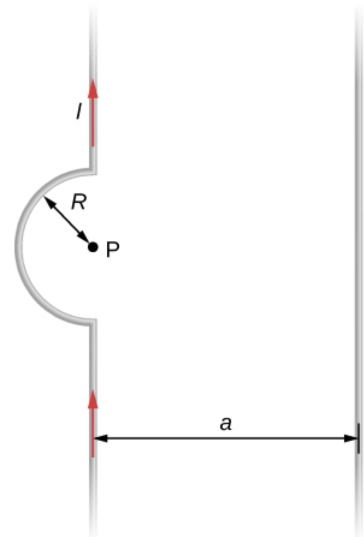
19. The accompanying figure shows a current loop consisting of two concentric circular arcs and two perpendicular radial lines. Determine the magnetic field at point P.



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.

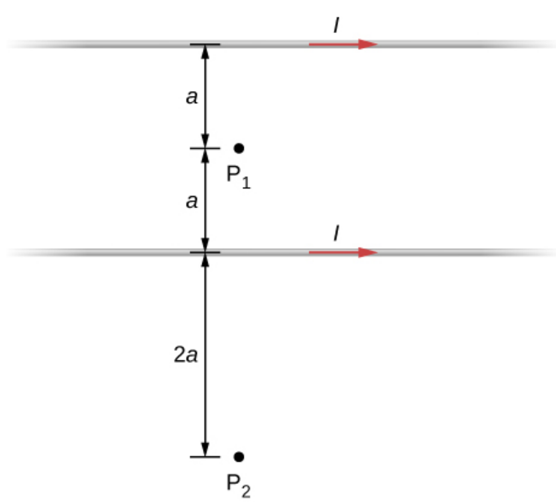
23. The magnitude of the magnetic field 50 cm from a long, thin, straight wire is **8.0μT**. What is the current through the long wire?

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 $4.0 \times 10^{-5} T$ 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 $I_1 = 10A$ and $I_2 = 20A$, 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.

(a) If each wire carries a current of 50 A in the same direction, what is the magnetic force per meter exerted on each wire?

(b) Does the force pull the wires together or push them apart?

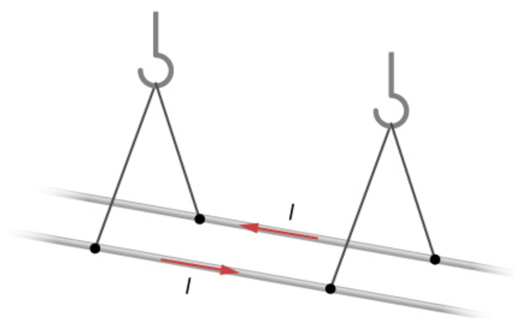
(c) What happens if the currents flow in opposite directions?

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.

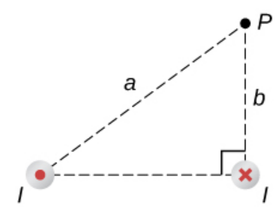
(a) If the two currents flow in opposite directions, what is the magnitude and direction of the force per unit length of one wire on the other?

(b) What is the magnitude and direction of the force per unit length if the currents flow in the same direction?

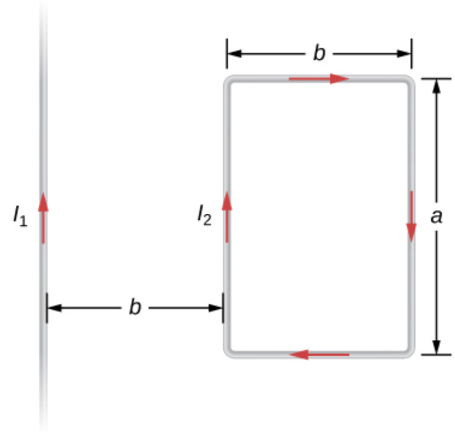
32. Two long, parallel wires are hung by cords of length 5.0 cm, as shown in the accompanying figure. Each wire has a mass per unit length of 30 g/m, and they carry the same current in opposite directions. What is the current if the cords hang at **6.0°** with respect to the vertical?



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.

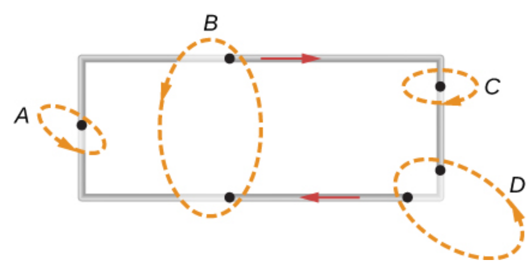
38. A circular loop of radius **R** carries a current **I**. At what distance along the axis of the loop is the magnetic field one-half its value at the center of the loop?

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?

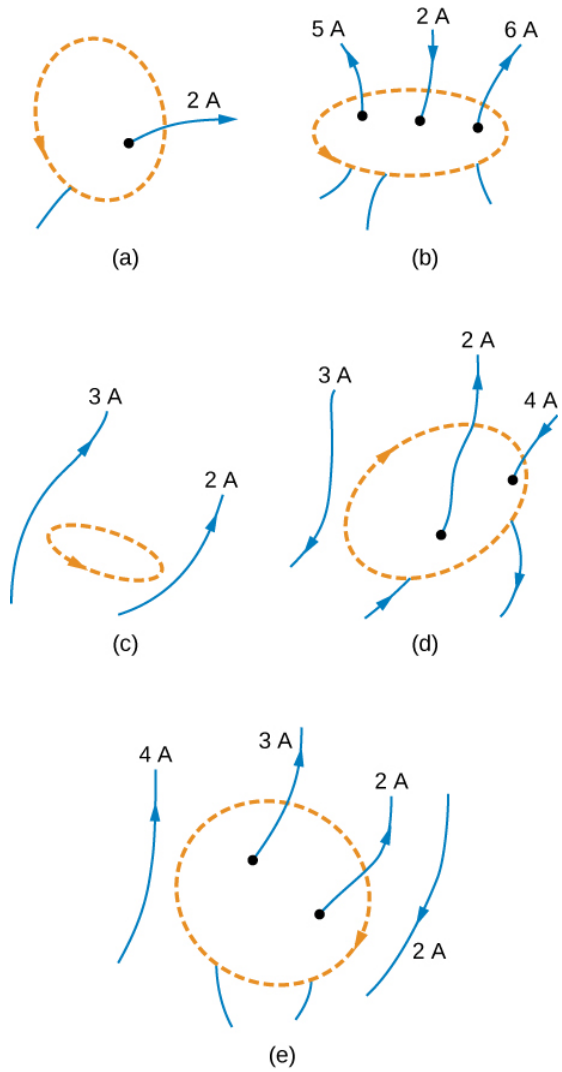
40. For the coils in the preceding problem, what is the magnetic field at the center of either coil?

12.6 Ampère's Law

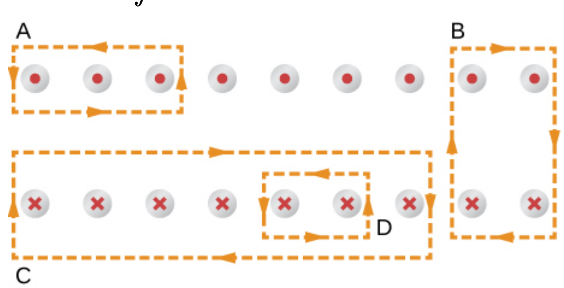
41. A current **I** flows around the rectangular loop shown in the accompanying figure. Evaluate $\oint \vec{B} \cdot \vec{dl}$ for the paths **A, B,C**, and **D.**



42. Evaluate $\oint \vec{B} \cdot \vec{dl}$ for each of the cases shown in the accompanying figure.



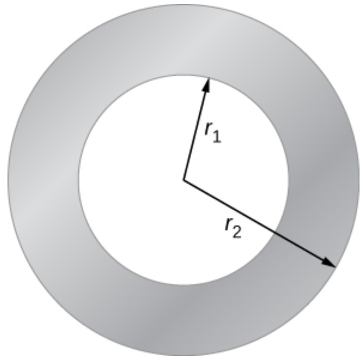
43. The coil whose lengthwise cross section is shown in the accompanying figure carries a current I and has N evenly spaced turns distributed along the length l. Evaluate $\oint \vec{B} \cdot \overrightarrow{dl}$ for the paths indicated.



44. A superconducting wire of diameter 0.25 cm carries a current of 1000 A. What is the magnetic field just outside the wire?

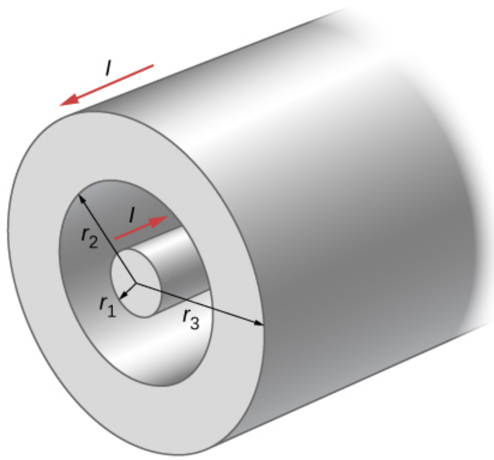
45. A long, straight wire of radius R carries a current I that is distributed uniformly over the cross-section of the wire. At what distance from the axis of the wire is the magnitude of the magnetic field a maximum?

46. The accompanying figure shows a cross-section of a long, hollow, cylindrical conductor of inner radius $r_1 = 3.0 cm$ and outer radius $r_2 = 5.0 cm$. A 50-A current distributed uniformly over the cross-section flows into the page. Calculate the magnetic field at $r = 2.0 cm$, $r = 4.0 cm$, and $r = 6.0 cm$.



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?

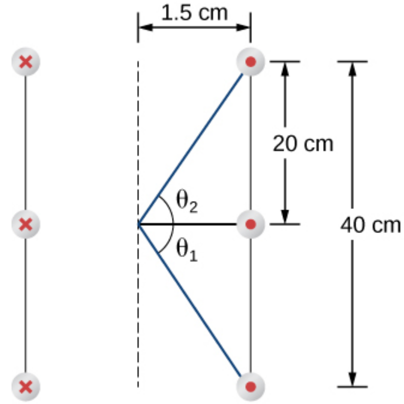
52. A solenoid is 40 cm long, has a diameter of 3.0 cm, and is wound with 500 turns. If the current through the windings is 4.0 A, what is the magnetic field at a point on the axis of the solenoid that is

(a) at the center of the solenoid,

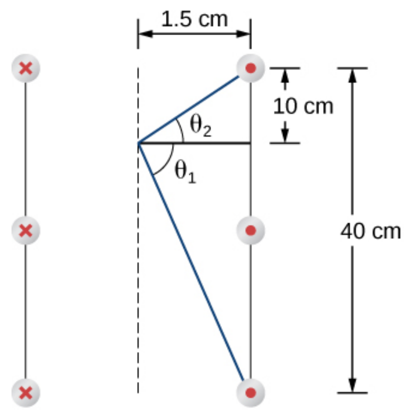
(b) 10.0 cm from one end of the solenoid, and

(c) 5.0 cm from one end of the solenoid?

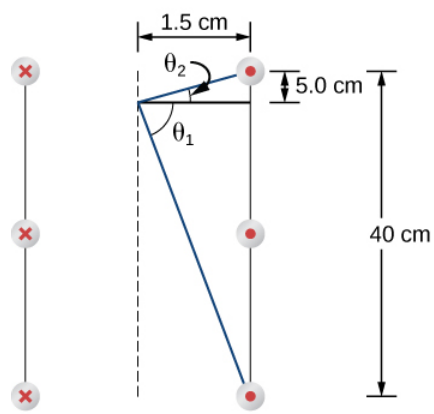
(d) Compare these answers with the infinite-solenoid case.



(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 $B=\mu_0 nI$ 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 $2.0\times10^5 m/s$, 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 **8.9cm**?

57. A toroid with a square cross section 3.0 cm × 3.0 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 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 $2.1 \times 10^{-23} A \cdot m^2$.

(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?

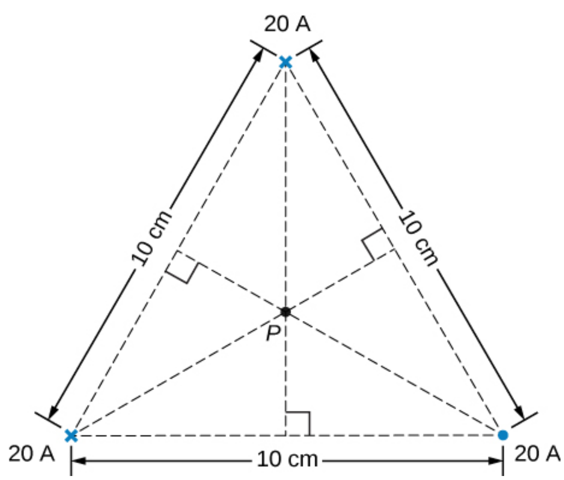
62. Suppose you wish to produce a 1.2-T magnetic field in a toroid with an iron core for which $\chi = 4.0 \times 10^3$. The toroid has a mean radius of 15 cm and is wound with 500 turns. What current is required?

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?

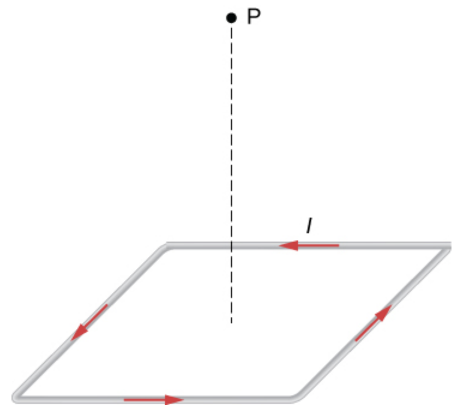
64. A solenoid with an iron core is 25 cm long and is wrapped with 100 turns of wire. When the current through the solenoid is 10 A, the magnetic field inside it is 2.0 T. For this current, what is the permeability of the iron? If the current is turned off and then restored to 10 A, will the magnetic field necessarily return to 2.0 T?

Additional Problems

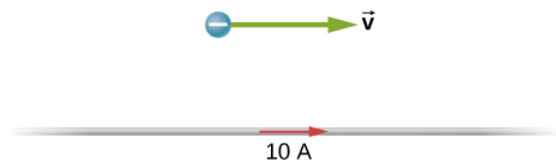
65. Three long, straight, parallel wires, all carrying 20 A, are positioned as shown in the accompanying figure. What is the magnitude of the magnetic field at the point **P**?



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)?



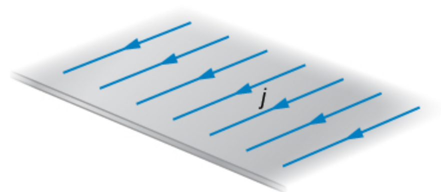
67. The accompanying figure shows a long, straight wire carrying a current of 10 A. What is the magnetic force on an electron at the instant it is 20 cm from the wire, traveling parallel to the wire with a speed of $2.0 \times 10^5 m/s$? Describe qualitatively the subsequent motion of the electron.



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.

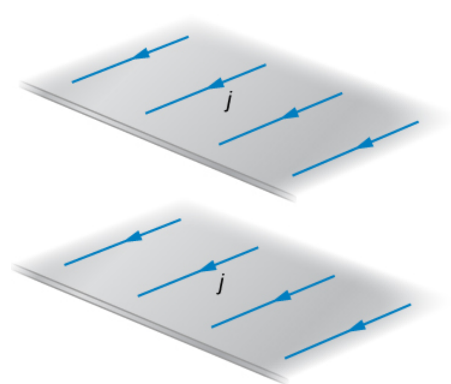
(a) Use the Biot-Savart law to show that $B = \mu_0 J/2$ on either side of the sheet. What is the direction of $\vec{B}$ on each side?

(b) Now use Ampère's law to calculate the field.

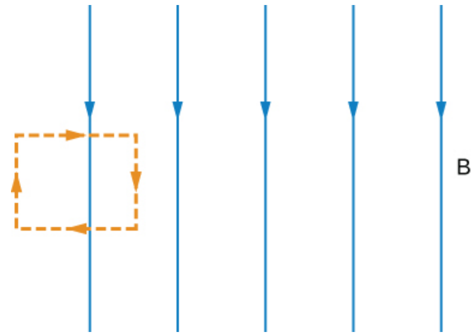


69. (a) Use the result of the previous problem to calculate the magnetic field between, above, and below the pair of infinite sheets shown in the accompanying figure.

(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?

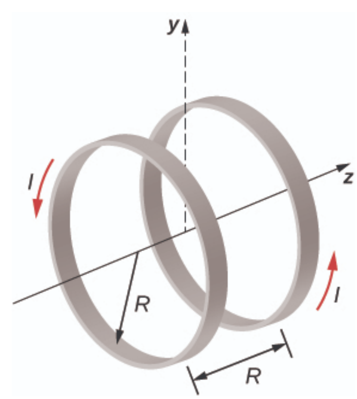
74. A wire element has $vecdl, I\vec{dl} = JAdl = Jdv$, where A and dv are the cross-sectional area and volume of the element, respectively. Use this, the Biot-Savart law, and $J = nev$ to show that the magnetic field of a moving point charge q is given by:

$$\vec{B} = \frac{\mu_0}{4\pi}\frac{qv\times\hat{r}}{r^2}.$$

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_dz2 are both zero at **z = 0.** (These vanishing derivatives demonstrate that the magnetic field varies only slightly near **z = 0**.)



76. A charge of **4.0μC** is distributed uniformly around a thin ring of insulating material. The ring has a radius of 0.20 m and rotates at $2.0\times10^4 rev/min$ 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?

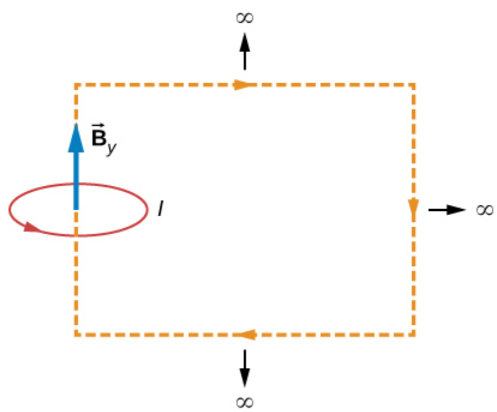
77. A thin, nonconducting disk of radius **R** is free to rotate around the axis that passes through its center and is perpendicular to the face of the disk. The disk is charged uniformly with a total charge **q**. If the disk rotates at a constant angular velocity **ω**, what is the magnetic field at 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.

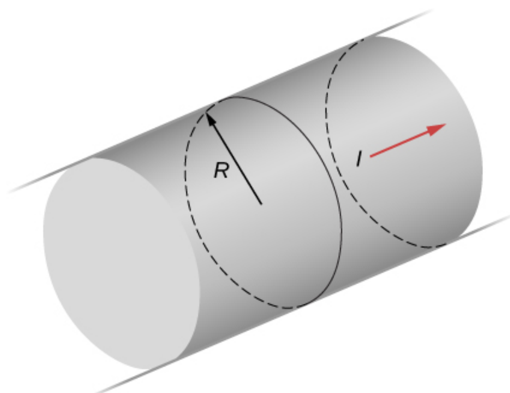
79. Consider the axial magnetic field $B_y = \mu_0 I R^2/2(y^2+R^2)^{3/2}$ of the circular current loop shown below.

(a) Evaluate $\int_{-a}^{a} B_y dy$. Also so show that $\lim_{a\to\infty}\int_{-a}^{a} B_y dy = \mu_0 I$.

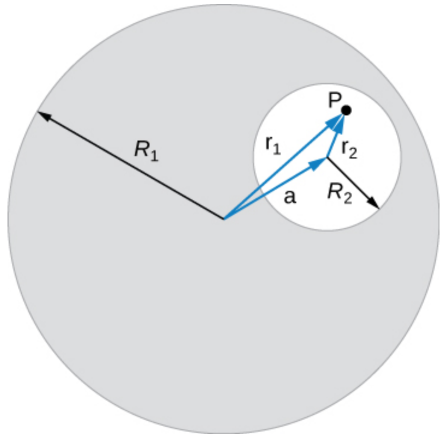
(b) Can you deduce this limit without evaluating the integral? (**Hint:**See the accompanying figure.)



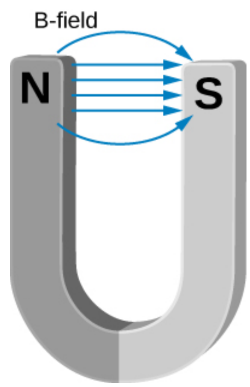
80. The current density in the long, cylindrical wire shown in the accompanying figure varies with distance r from the center of the wire according to $J = cr$, where c is a constant. (a) What is the current through the wire? (b) What is the magnetic field produced by this current for $r \le R$? For $r \ge R$?



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 $\vec{B} = \frac{1}{2}\mu_0 J_0 k \times a$, where $a = r_1 - r_2$ and $r_1 = r_1\hat{r_1}$ is the position of P relative to the center of the conductor and $2_{=} r_2\overrightarrow{r_2}$ is the position of **P** relative to the center of the cavity.

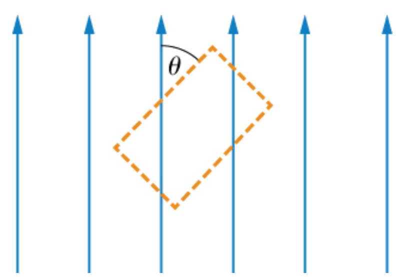


82. Between the two ends of a horseshoe magnet the field is uniform as shown in the diagram. As you move out to outside edges, the field bends. Show by Ampère's law that the field must bend and thereby the field weakens due to these bends.



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 $\vec{B} \cdot \vec{dl}$ for each side of the loop then find the entire $\oint \vec{B} \cdot \vec{dl}$. 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

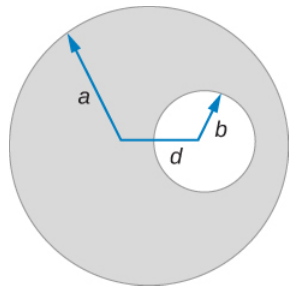
(b) at a point inside the wire. Write your answer in terms of the net current I through the wire.

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.)

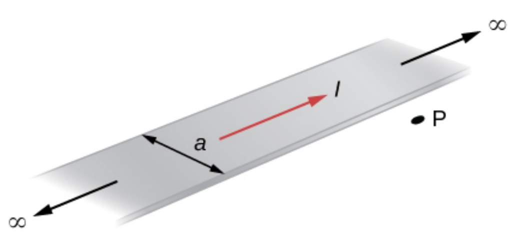


87. Magnetic field inside a torus. Consider a torus of rectangular cross-section with inner radius a and outer radius b. N turns of an insulated thin wire are wound evenly on the torus tightly all around the torus and connected to a battery producing a steady current I in the wire. Assume that the current 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 **r** from the axis.

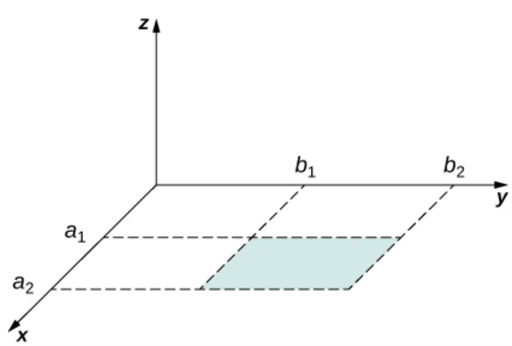
88. Two long coaxial copper tubes, each of length **L**, are connected to a battery of voltage **V**. The inner tube has inner radius **a** and outer radius **b**, and the outer tube has inner radius c and outer radius **d**. The tubes are then disconnected from the battery and rotated in the same direction 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 **r<a**, and (b) at a point between the tubes **b<r<c**, and (c) at a point outside the tubes **r>d**. (Hint: Think of copper tubes as a capacitor and find the charge density based on the voltage applied, **Q=VC,** \(\displaystyle C=\frac{2πε_0L}{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 $\mathbf{a \rightarrow 0}$.



90. A hypothetical current flowing in the z-direction creates the field $vecB = C[(x/y^2)\hat{i} + (1/y)\hat{j}]$ 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.

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. $A \cdot m^2 \cdot T = A \cdot m^2 . \dfrac{N}{A \cdot m} = N \cdot m$

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. $2.50 \times 10^{-} 2m$;

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. $9.8 \times 10^{-5} \hat{j}T$; 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. $0.31 Am^2$;

b. 0.16 Nm

87. $0.024 Am^2$

89. a. $0.16 Am^2$;

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. $R = (mvsin\theta)/qB; p = (\frac{2\pi m}{eB})vcos\theta$

105. $IaL^2/2$

107. $m = \frac{qB_0^2}{8V_{acc}}x^2$

109. 0.01 N

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 with 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. $B = \frac{\mu_o I}{8}\left(\frac{1}{a} - \frac{1}{b}\right)$ 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, $B = \frac{3\mu_o I}{8\pi a}$ 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. $B = \dfrac{\mu_o Ia}{2\pi b^2}$ into the page

35. 0.019 m

37. $6.28 \times 10^{-5} T$

39. $B = \dfrac{\mu_o IR^2}{\left(\left(\frac{d}{2}\right)^2 + R^2\right)^{3/2}}$

41. a. $\mu 0I$;

b. 0;

c. $\mu 0I$;

d. 0

43. a. $3\mu_0 I$;

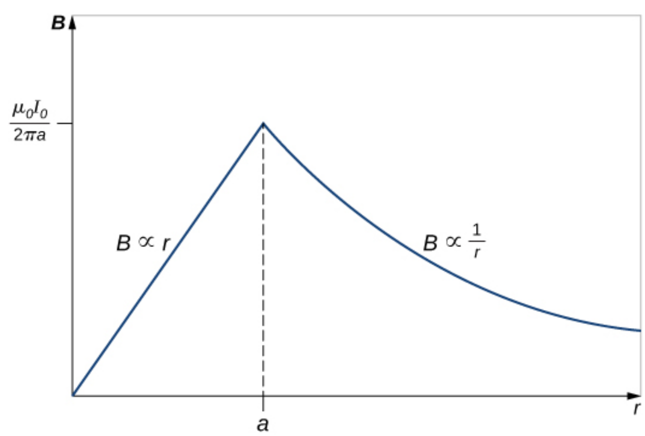
b. 0;

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. $B = \dfrac{1}{2}\mu_0 nI$

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

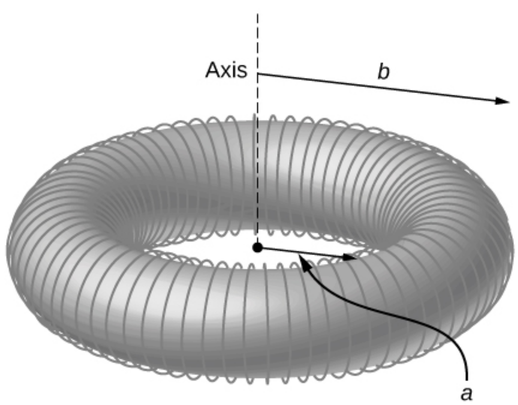
81. derivation

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. $B = \frac{\mu_0 \sigma \omega}{2} [\frac{2h^2 + R^2}{\sqrt{R^2 + h^0}}]$;

b. $B = 4.09 \times 10^{-5} T$, 82% of Earth's magnetic field

CHAPTER OVERVIEW

19: Alternating-Current (AC) Circuits

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.

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.

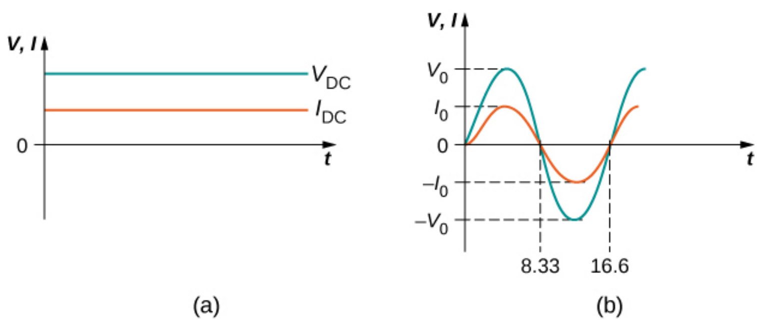


Figure 19.2.1: (a) The dc voltage and current are constant in time, once the current is established. (b) The voltage and current versus time are quite different for ac power. In this example, which shows 60-Hz ac power and time t in milliseconds, voltage and current are sinusoidal and are in phase for 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, $V_0 = 156\ V$ and $\omega = 120\pi\ rad/s$, whereas in Europe, $V_0 = 311\ V$ and $\omega = 100\pi\ rad/s$.

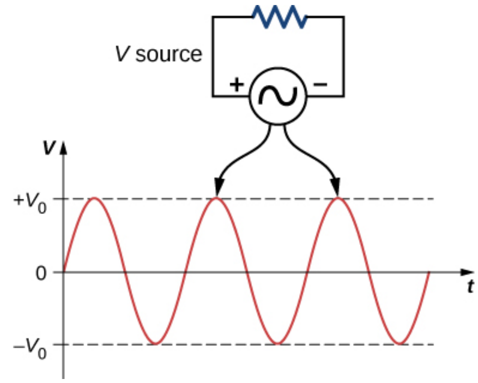


Figure 19.2.2: The potential difference **V** between the terminals of an ac voltage source fluctuates, so the source and the resistor have ac sine waves on top of each other. The mathematical expression for **v** is given by $v = V_0\, sin\, \omega t$.

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

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 $v = 0 \, V$ at a time of $t = 0 \, s$. 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.

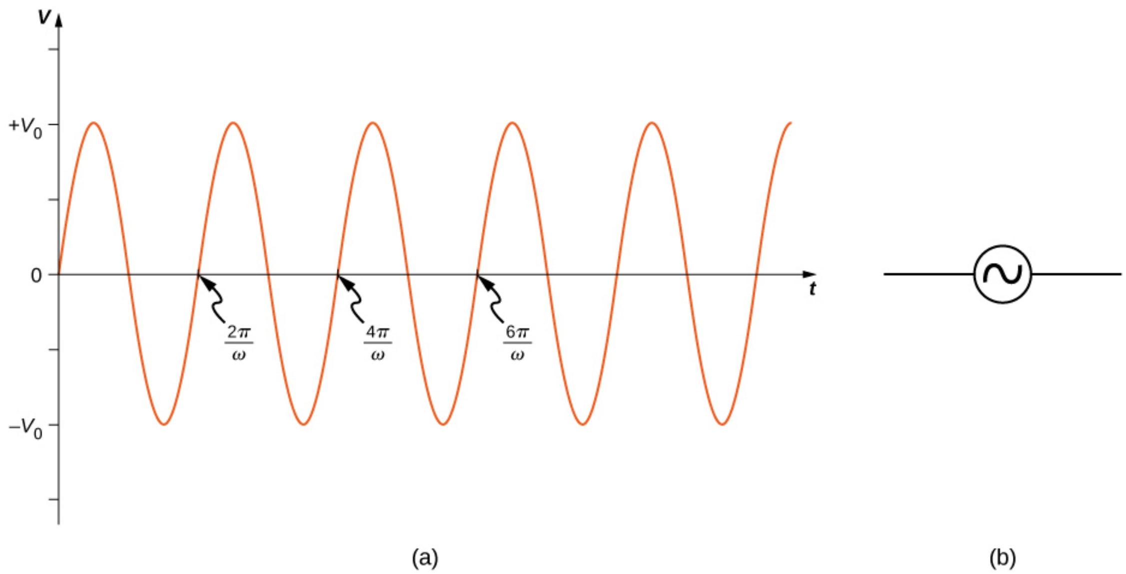


Figure 19.3.1: (a) The output $v(t) = V_0 \, sin \, \omega t$ of an ac generator. (b) Symbol used to represent an ac voltage source in a circuit diagram.

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

$$i_R(t) = \frac{v_R(t)}{R} = \frac{V_0}{R} \, sin \, \omega t = I_0 \, sin \, \omega t.$$

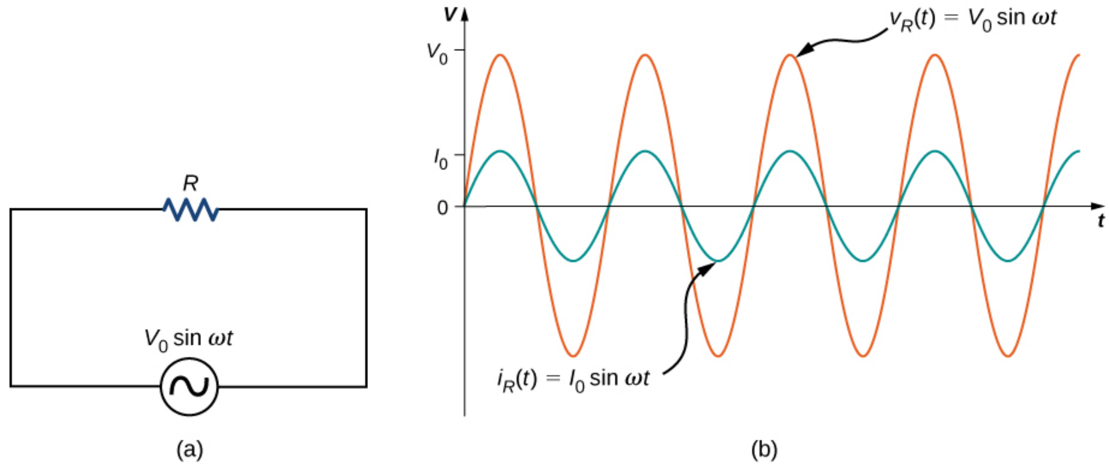


Figure 19.3.2: (a) A resistor connected across an ac voltage source. (b) The current $i_R(t)$ through the resistor and the voltage $v_R(t)$ across the resistor. The two quantities are in phase.

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.

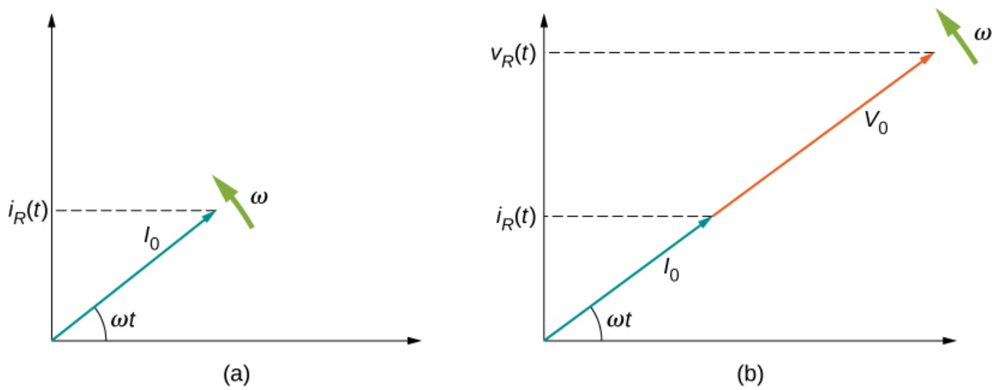


Figure 19.3.3: (a) The phasor diagram representing the current through the resistor of Figure 19.3.2 (b) The phasor diagram representing both $i_R(t)$ and $v_R(t)$.

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,

$$i_C(t) = \frac{dq(t)}{dt} = \omega C V_0 \cos \omega t = I_0 \cos \omega t,$$

where $I_0 = \omega C V_0$ 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:

$$\frac{V_0}{I_0} = \frac{1}{\omega C} = X_C. \tag{19.3.1}$$

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.

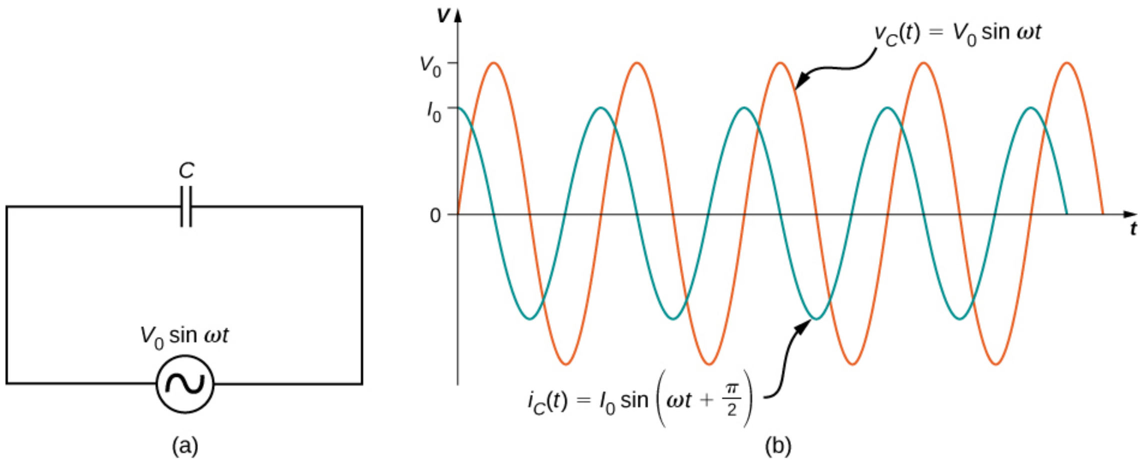


Figure 19.3.4: (a) A capacitor connected across an ac generator. (b) The current $i_C(t)$ through the capacitor and the voltage $v_C(t)$ across the capacitor. Notice that $i_C(t)$ leads $v_C(t)$ by $\pi/2$ rad.

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, with the current phasor leading by $\pi/2$ rad.

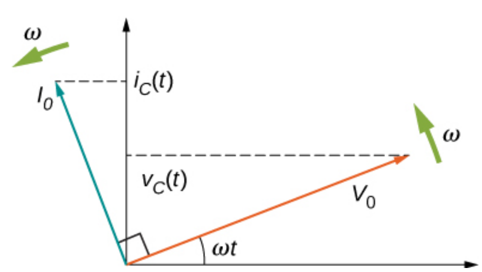


Figure 19.3.5: The current phasor leads the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

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

$$v_L(t) = V_0\, sin\, \omega t. \tag{19.3.2}$$

The emf across an inductor is equal to $\epsilon = -L(di_L/dt)$; however, the potential difference across the inductor is $v_L(t) = L di_L(t)/dt$, 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

$$\frac{di_L(t)}{dt} = \frac{V_0}{L} sin\, \omega t.$$

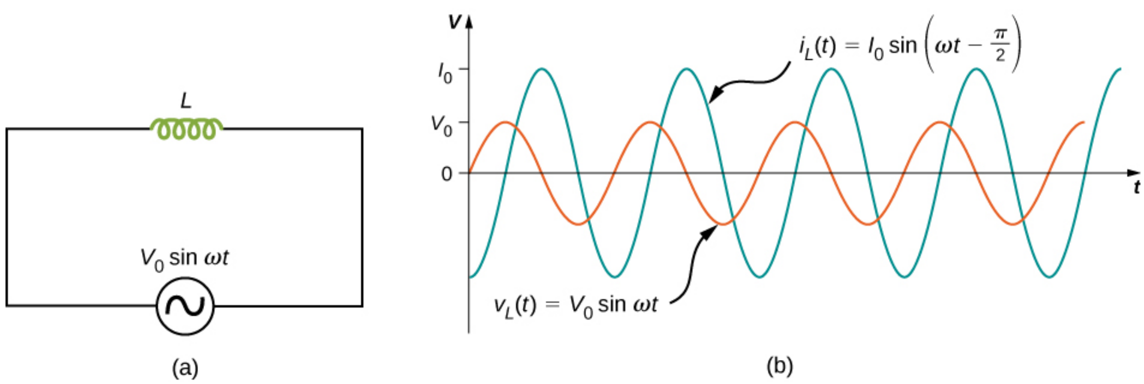


Figure 19.3.6: (a) An inductor connected across an ac generator. (b) The current $i_L(t)$ through the inductor and the voltage $v_L(t)$ across the inductor. Here $i_L(t)$ lags $v_L(t)$ by $\pi/2$ rad.

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

$$i_L(t) = -\frac{V_0}{\omega L} cos\, \omega t = \frac{V_0}{\omega L} sin\left(\omega t - \frac{\pi}{2}\right) = I_0\, sin\left(\omega t - \frac{\pi}{2}\right), \tag{19.3.3}$$

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
>
> $$\frac{V_0}{I_0} = \omega L = X_L. \tag{19.3.4}$$

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.

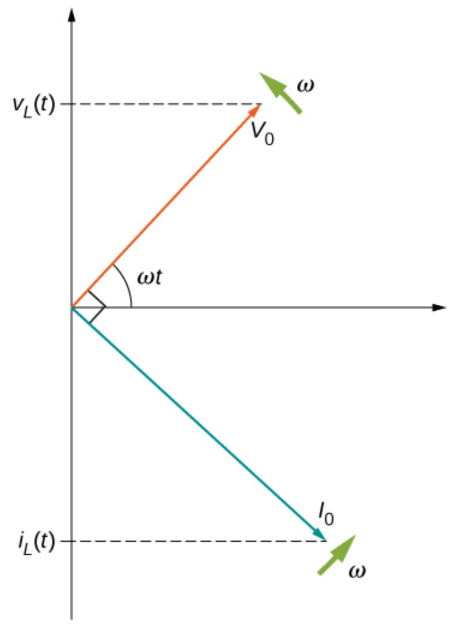


Figure 19.3.7: The current phasor lags the voltage phasor by $\pi/2$ rad as they both rotate with the same angular frequency.

> ✔ 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 $f = 60\,Hz$. Determine the voltages across and the currents through the circuit elements when the generator is connected to (a) a $100\,\Omega$ resistor, (b) a $10\,\mu F$ 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
>
> $$v(t) = V_0\,sin\,\omega t = (10\,V) sin\,120\pi t,$$
>
> where $\omega = 2\pi f = 120\pi\,rad/s$ is the angular frequency. Since **v(t)** is also the voltage across each of the elements, we have

$$v(t) = v_R(t) = v_C(t) = v_L(t) = (10\,V) sin\,120\pi t.$$

a. When $R = 100\,\Omega$, the amplitude of the current through the resistor is

$$I_0 = V_0/R = 10\,V/100\,\Omega = 0.10\,A,$$

so

$$i_R(t) = (0.10\,A) sin\,120\pi t.$$

b. From Equation 19.3.1, the capacitive reactance is

$$X_C = \frac{1}{\omega C} = \frac{1}{(120\pi\,rad/s)(10\times10^{-6}F)} = 265\,\Omega,$$

so the maximum value of the current is

$$I_0 = \frac{V_0}{X_C} = \frac{10\,V}{265\,\Omega} = 3.8\times10^{-2}\,A$$

and the instantaneous current is given by

$$i_C(t) = (3.8\times10^{-2}\,A)\,sin\,\left(120\pi t + \frac{\pi}{2}\right).$$

c. From Equation 19.3.4, the inductive reactance is

$$X_L = \omega L = (120\pi\,rad/s)(15\times10^{-3}H) = 5.7\,\Omega.$$

The maximum current is therefore

$$I_0 = \frac{10\,V}{5.7\,\Omega} = 1.8\,A$$

and the instantaneous current is

$$i_L(t) = (1.8\,A)\,sin\,\left(120\pi t - \frac{\pi}{2}\right).$$

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. $(20\,V)\,sin\,200\pi t$; $(0.20\,A)\,sin\,200\pi t$

b. $(20\,V)\,sin\,200\pi t$; $(0.13\,A)\,sin\,(200\pi t + \pi/2)$

c. $(20\,V)\,sin\,200\pi t$; $(2.1\,A)\,sin\,(200\pi t - \pi/2)$

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

$$v(t) = V_0 \sin \omega t.$$

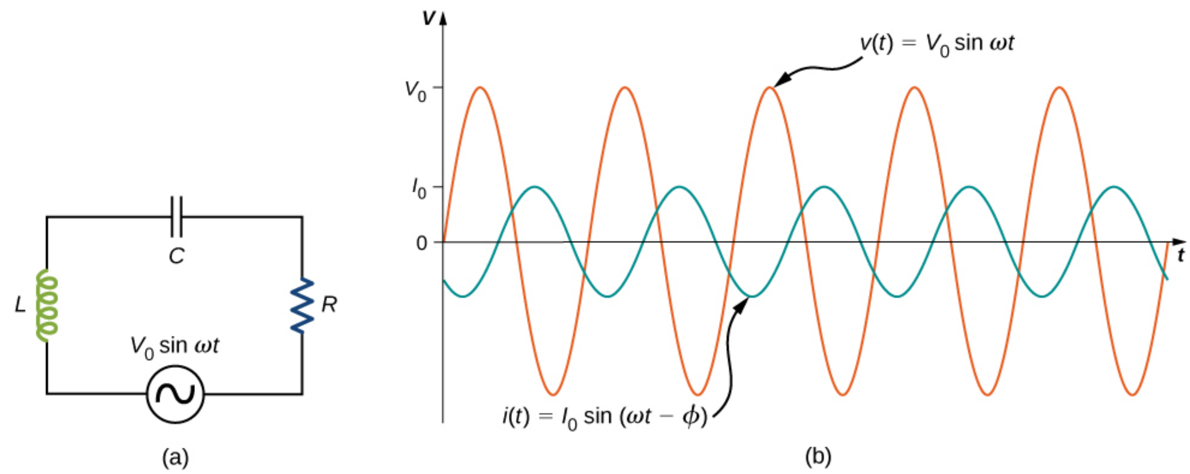


Figure 19.4.1: (a) An RLC series circuit. (b) A comparison of the generator output voltage and the current. The value of the phase difference ϕ depends on the values of R, C, and L.

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 $V_L = I_0 X_L$.

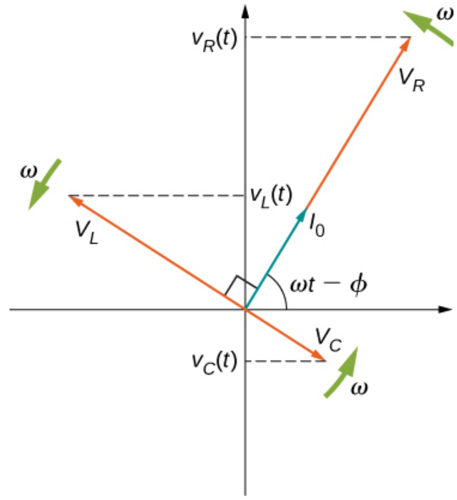


Figure 19.4.1.

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,

$$\phi = tan^{-1}\frac{V_L-V_C}{V_R}=tan^{-1}\frac{I_0X_L-I_0X_C}{I_0R},$$

and after cancellation of I_0, this becomes

$$\phi = tan^{-1}\frac{X_L-X_C}{R}. \tag{19.4.1}$$

Furthermore, from the Pythagorean theorem,

$$V_0=\sqrt{V_R^2+(V_L-V_C)^2}=\sqrt{(I_0R)^2+(I_0X_L-I_)X_C)^2}=I_0\sqrt{R^2+(X_L-X_C)^2}.$$

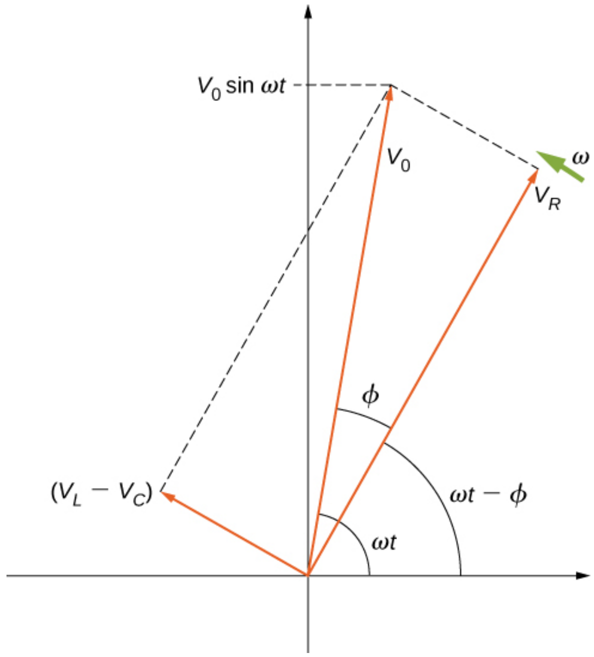


Figure 19.4.3: The resultant of the phasors for $v_L(t)$, $v_C(t)$, and $v_R(t)$ is equal to the phasor for $v_R(t) = V_0 \, sin \, \omega t$. The **i(t)** phasor (not shown) is aligned with the $v_R(t)$ phasor.

The current amplitude is therefore the ac version of Ohm's law:

$$I_0 = \frac{V_0}{\sqrt{R^2 + (X_L - X_C)^2}} = \frac{V_0}{Z}, \tag{19.4.2}$$

where

$$Z = \sqrt{R^2 + (X_L - X_C)^2} \tag{19.4.3}$$

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.

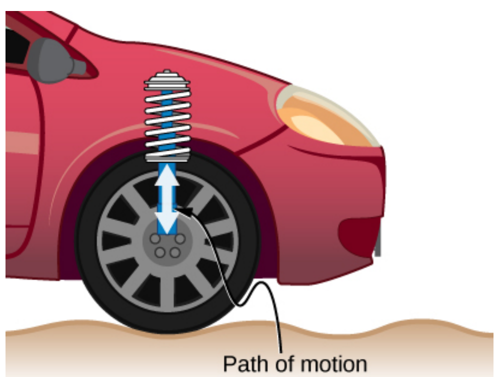


Figure 19.4.5: On a car, the shock absorber damps motion and dissipates energy. This is much like the resistance in an RLC circuit. The mass and spring determine the resonant frequency.

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

$$X_C = \frac{1}{\omega C} = \frac{1}{2\pi(200\,Hz)(8.00\times 10^{-4} F)} = 0.995\,\Omega.$$

2. From Equation 15.3.14, the inductive reactance is

$$X_L = \omega L = 2\pi(200\,Hz)(3.00\times 10^{-3} H) = 3.77\,\Omega.$$

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

$$\phi = tan^{-1}\frac{X_L - X_C}{R} = tan^{-1}\frac{2.77\,\Omega}{4.00\,\Omega} = 0.607\,rad.$$

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

$v_R = (V_0 R/Z)\, sin\,(\omega t - \phi)$; $v_C = (V_0 X_C/Z)\, sin\,(\omega t - \phi + \pi 2/) = -(V_0 X_C/Z)\, cos\,(\omega t - \phi)$;
$v_L = (V_0 X_L/Z)\, sin\,(\omega t - \phi - \pi/2) = (V_0 X_L/Z)\, cos\,(\omega t - \phi)$

- **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 *[Math Processing Error]*, where **I** is the current through the element and *[Math Processing Error]* is the voltage across it. Since the current and the voltage both depend on time in an ac circuit, the instantaneous power *[Math Processing Error]* is also time dependent. A plot of *[Math Processing Error]* for various circuit elements is shown in Figure *[Math Processing Error]*. For a resistor, *[Math Processing Error]* and *[Math Processing Error]* are in phase and therefore always have the same sign. For a capacitor or inductor, the relative signs of *[Math Processing Error]* and *[Math Processing Error]* vary over a cycle due to their phase differences. Consequently, *[Math Processing Error]* 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.

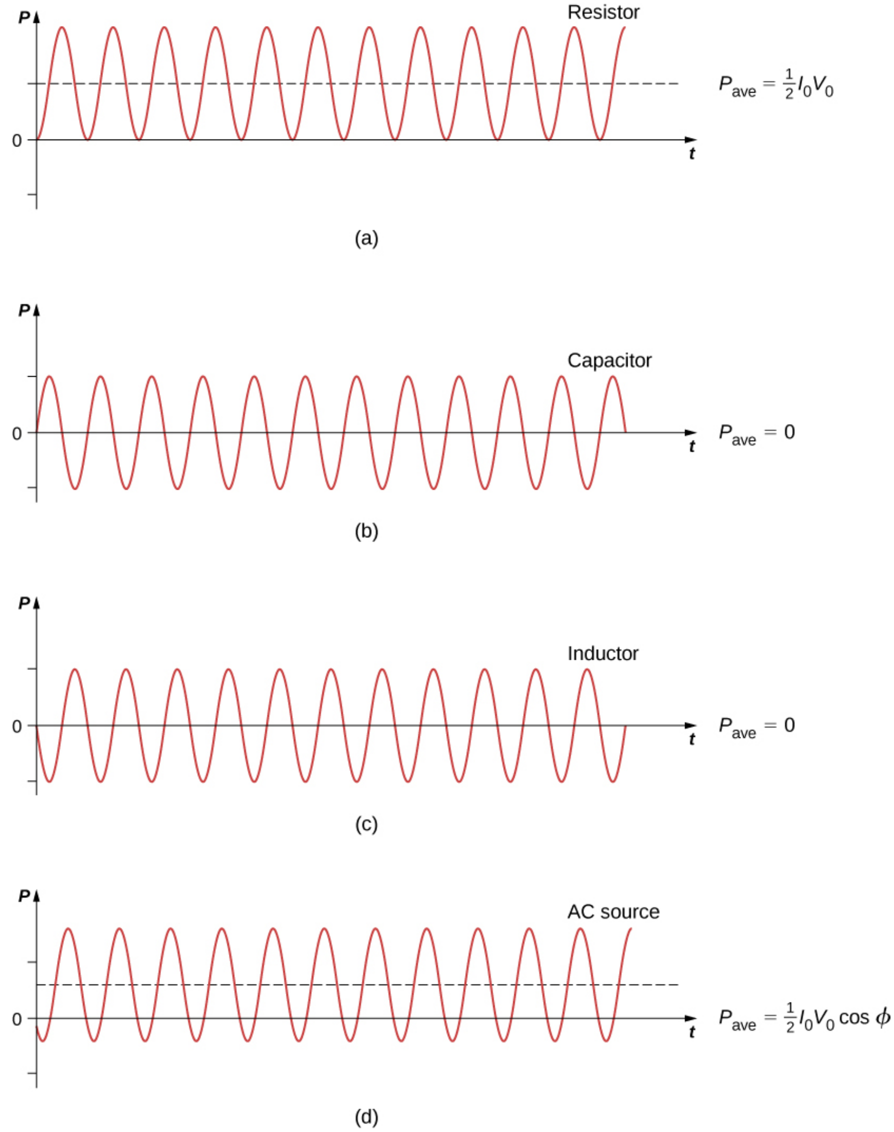


Figure *[Math Processing Error]*: Graph of instantaneous power for various circuit elements. (a) For the resistor, *[Math Processing Error]*, whereas for (b) the capacitor and (c) the inductor, *[Math Processing Error]*. (d) For the source, *[Math Processing Error]*, which may be positive, negative, or zero, depending on *[Math Processing Error]*.

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:

[Math Processing Error]

where *[Math Processing Error]* is the period of the oscillations. With the substitutions *[Math Processing Error]* and *[Math Processing Error]*, Equation *[Math Processing Error]* becomes

[Math Processing Error]

Using the trigonometric difference identity

[Math Processing Error]

we obtain

[Math Processing Error]

Evaluation of these two integrals yields

[Math Processing Error]

and

[Math Processing Error]

Hence, the average power associated with a circuit element is given by

[Math Processing Error]

In engineering applications, [Math Processing Error] 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, [Math Processing Error], so the average power dissipated is

[Math Processing Error]

A comparison of [Math Processing Error] and [Math Processing Error] is shown in Figure [Math Processing Error]. To make Equation [Math Processing Error] look like its dc counterpart, we use the rms values [Math Processing Error] and [Math Processing Error] of the current and the voltage. By definition, these are

[Math Processing Error]

and

[Math Processing Error]

where

[Math Processing Error]

and

[Math Processing Error]

With [Math Processing Error] and [Math Processing Error], we obtain

[Math Processing Error]

and

[Math Processing Error]

We may then write for the average power dissipated by a resistor,

[Math Processing Error]

This equation further emphasizes why the rms value is chosen in discussion rather than peak values. Both Equations [Math Processing Error] and [Math Processing Error] 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 [Math Processing Error]. 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, [Math Processing Error] and [Math Processing Error], respectively. Since [Math Processing Error], we find from Equation [Math Processing Error] that the average power dissipated by either of these elements is [Math Processing Error]. 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 [Math Processing Error] and [Math Processing Error] which show [Math Processing Error] oscillating sinusoidally about zero.

The phase angle for an ac generator may have any value. If [Math Processing Error], the generator produces power; if [Math Processing Error], it absorbs power. In terms of rms values, the average power of an ac generator is written as

[Math Processing Error]

For the generator in an **RLC** circuit,

[Math Processing Error] and

[Math Processing Error]

Hence the average power of the generator is

[Math Processing Error]

This can also be written as

[Math Processing Error]

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 *[Math Processing Error]*: Power Output of a Generator

An ac generator whose emf is given by

[Math Processing Error]

is connected to an **RLC** circuit for which *[Math Processing Error]*, and *[Math Processing Error]*.

a. What is the rms voltage across the generator?
b. What is the impedance of the circuit?
c. What is the average power output of the generator?

Strategy

The rms voltage is the amplitude of the voltage times *[Math Processing Error]*. The impedance of the circuit involves the resistance and the reactances of the capacitor and the inductor. The average power is calculated by Equation *[Math Processing Error]* because we have the impedance of the circuit *[Math Processing Error]*, the rms voltage *[Math Processing Error]*, and the resistance *[Math Processing Error]*.

Solution

1. Since *[Math Processing Error]*, the rms voltage across the generator is *[Math Processing Error]*
2. The impedance of the circuit is *[Math Processing Error]*
3. From Equation *[Math Processing Error]*, the average power transferred to the circuit is *[Math Processing Error]*

Significance

If the resistance is much larger than the reactance of the capacitor or inductor, the average power is a dc circuit equation of *[Math Processing Error]*, where **V** replaces the rms voltage.

Exercise *[Math Processing Error]*

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

[Math Processing Error]

Exercise *[Math Processing Error]*

Show that the rms voltages across a resistor, a capacitor, and an inductor in an ac circuit where the rms current is *[Math Processing Error]* are given by *[Math Processing Error]*, and *[Math Processing Error]*, respectively. Determine these values for the components of the **RLC** circuit of Equation *[Math Processing Error]*.

Answer

2.00 V; 10.01 V; 8.01 V

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,

[Math Processing Error]

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 *[Math Processing Error]* versus *[Math Processing Error]* is shown in Figure *[Math Processing Error]*.

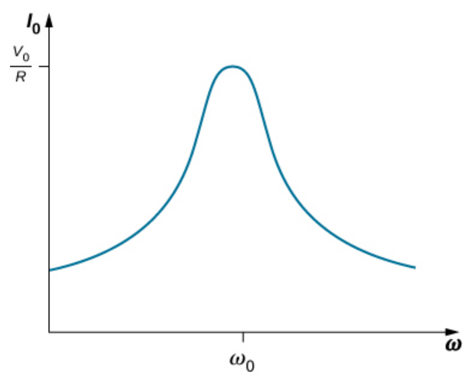


Figure *[Math Processing Error]*: At an RLC circuit's resonant frequency, *[Math Processing Error]*, the current amplitude is at its maximum value.

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

[Math Processing Error]

or

[Math Processing Error]

where we substituted **dq**(t)/**dt** for **i**(t). A comparison of Equation *[Math Processing Error]* 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** *[Math Processing Error]* 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 *[Math Processing Error]* at which the impedance **Z** in Equation *[Math Processing Error]* is a minimum, or when

[Math Processing Error]

and

[Math Processing Error]

This is the resonant angular frequency of the circuit. Substituting *[Math Processing Error]* into Equation 15.4.5, Equation 15.4.7, and Equation 15.4.8, we find that at resonance,

[Math Processing Error]

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, *[Math Processing Error]* reaches a maximum when *[Math Processing Error]*, which depends on the frequency, is a minimum, that is, when *[Math Processing Error]* and *[Math Processing Error]*. 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 *[Math Processing Error]*.

Figure *[Math Processing Error]* is a typical plot of *[Math Processing Error]* versus *[Math Processing Error]* in the region of maximum power output. The **bandwidth** *[Math Processing Error]* of the resonance peak is defined as the range of angular frequencies *[Math Processing Error]* over which the average power *[Math Processing Error]* is greater than one-half the maximum value of *[Math Processing Error]*. The sharpness of the peak is described by a dimensionless quantity known as the **quality factor Q** of the circuit. By definition,

[Math Processing Error]

where *[Math Processing Error]* is the resonant angular frequency. A high **Q** indicates a sharp resonance peak. We can give **Q** in terms of the circuit parameters as

[Math Processing Error]

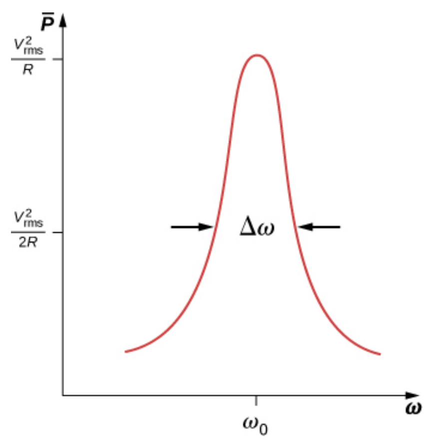


Figure *[Math Processing Error]*: Like the current, the average power transferred from an ac generator to an **RLC** circuit peaks at the resonant frequency.

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 *[Math Processing Error]*).

Figure *[Math Processing Error]*: 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 *[Math Processing Error]*: Resonance in an RLC Series Circuit

a. What is the resonant frequency of the circuit of Example 15.3.1?
b. 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 *[Math Processing Error]*, 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

1. The resonant frequency is found from Equation *[Math Processing Error]*: *[Math Processing Error]*
2. At resonance, the impedance of the circuit is purely resistive, and the current amplitude is *[Math Processing Error]*

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 *[Math Processing Error]*: Power Transfer in an RLC Series Circuit at Resonance

a. What is the resonant angular frequency of an **RLC** circuit with *[Math Processing Error]*, and *[Math Processing Error]*?
b. If an ac source of constant amplitude 4.00 V is set to this frequency, what is the average power transferred to the circuit?
c. Determine **Q** and the bandwidth of this circuit.

Strategy

The resonant angular frequency is calculated from Equation *[Math Processing Error]*. The average power is calculated from the rms voltage and the resistance in the circuit. The quality factor is calculated from Equation *[Math Processing Error]* and by knowing the resonant frequency. The bandwidth is calculated from Equation *[Math Processing Error]* and by knowing the quality factor.

Solution

1. The resonant angular frequency is *[Math Processing Error]*
2. At this frequency, the average power transferred to the circuit is a maximum. It is *[Math Processing Error]*
3. The quality factor of the circuit is *[Math Processing Error]* We then find for the bandwidth *[Math Processing Error]*

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 [Math Processing Error]

In the circuit of Figure 15.4.1, [Math Processing Error], and [Math Processing Error].

a. What is the resonant frequency?
b. What is the impedance of the circuit at resonance?
c. If the voltage amplitude is 10 V, what is **i(t)** at resonance?
d. 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

a. 160 Hz; b. [Math Processing Error]; c. [Math Processing Error]; d. 0.023 rad

Exercise [Math Processing Error]

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

a. halved; b. halved; c. same

Exercise [Math Processing Error]

The resonant angular frequency of an **RLC** series circuit is [Math Processing Error]. An ac source operating at this frequency transfers an average power of [Math Processing Error] to the circuit. The resistance of the circuit is [Math Processing Error]. Write an expression for the emf of the source.

Answer

[Math Processing Error]

19.7: AC Safety - Grounding and Bonding

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	*[Math Processing Error]*
AC current	*[Math Processing Error]*
capacitive reactance	*[Math Processing Error]*
rms voltage	*[Math Processing Error]*
rms current	*[Math Processing Error]*

inductive reactance	[Math Processing Error]
Phase angle of an RLC series circuit	[Math Processing Error]
AC version of Ohm's law	[Math Processing Error]
Impedance of an RLC series circuit	[Math Processing Error]
Average power associated with a circuit element	[Math Processing Error]
Average power dissipated by a resistor	[Math Processing Error]
Resonant angular frequency of a circuit	[Math Processing Error]
Quality factor of a circuit	[Math Processing Error]
Quality factor of a circuit in terms of the circuit parameters	[Math Processing Error]
Transformer equation with voltage	[Math Processing Error]
Transformer equation with current	[Math Processing Error]

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.

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 $5.0\text{-}\mu F$ 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** *[Math Processing Error]*. Write an expression for the current through the resistor.

22. A **25-μF** capacitor is connected to an emf given by *[Math Processing Error]*.

(a) What is the reactance of the capacitor?

(b) Write an expression for the current output of the source.

23. A 100-mH inductor is connected across the emf of the preceding problem.

(a) What is the reactance of the inductor?

(b) 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-μF5.0-μF capacitor, and a 10-μF10-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 *[Math Processing Error]* where *[Math Processing Error]* and *[Math Processing Error]*.

(a) What is the impedance of the circuit?

(b) What is the amplitude of the current through the resistor?

(c) Write an expression for the current through the resistor.

(d) 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 *[Math Processing Error]* where *[Math Processing Error]* and *[Math Processing Error]*; also, *[Math Processing Error]* and *[Math Processing Error]*.

(a) What is the impedance of the circuit?

(b) What is the amplitude of the current through the resistor?

(c) Write an expression for the current through the resistor.

(d) 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 *[Math Processing Error]* and *[Math Processing Error]*.

(a) What is the impedance of the circuit?

(b) What is the amplitude of the current from the source?

(c) If the emf of the source is given by *[Math Processing Error]*, how does the current vary with time?

(d) Repeat the calculations with i changed to **0.20μF**.

28. An **RLC** series circuit with *[Math Processing Error]* and *[Math Processing Error]* is driven by an ac source whose frequency and voltage amplitude are 500 Hz and 50 V, respectively.

(a) What is the impedance of the circuit?

(b) What is the amplitude of the current in the circuit?

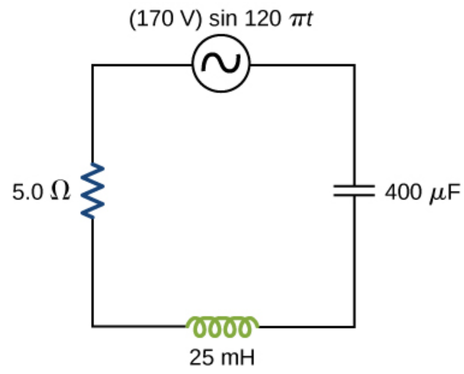
(c) What is the phase angle between the emf of the source and the current?

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 [Math Processing Error].



15.5 Power in an AC Circuit

30. The emf of an ac source is given by [Math Processing Error] where [Math Processing Error] and [Math Processing Error]. Calculate the average power output of the source if it is connected across

(a) a **20-μF** capacitor,

(b) a 20-mH inductor, and

(c) a **50-Ω** resistor.

31. Calculate the rms currents for an ac source is given by [Math Processing Error] where [Math Processing Error] and [Math Processing Error] when connected across

(a) a **20-μF** capacitor,

(b) a 20-mH inductor, and

(c) a **50-Ω** 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=500Ω**; and **L=0.20H**. Find the average power dissipated in the resistor for the following values for the capacitance:

(a) **C=2.0μF** and

(b) **C=0.20μF.**

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 [Math Processing Error], and **C=4.0μF**. (b) If **R** is changed to **300Ω**, what happens to the resonant angular frequency?

37. The resonant frequency of an **RLC** series circuit is [Math Processing Error]. 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?

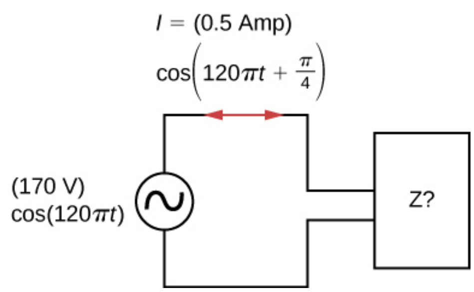
(d) What is the bandwidth of the circuit?

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 *[Math Processing Error]* where *[Math Processing Error]* and *[Math Processing Error]*. Find an expression that represents the output current of the source if it is connected across

(a) a **20-μF** capacitor,

(b) a 20-mH inductor, and

(c) a **50-Ω** resistor.

50. A **700-pF** 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-μF** 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?

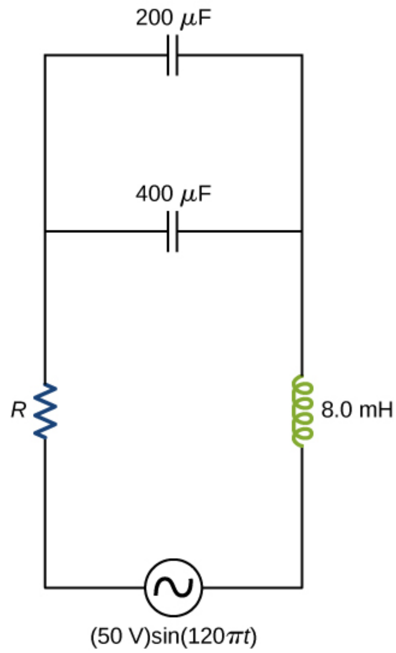
53. A 7.0-mH inductor is connected across a 60-Hz ac source whose voltage amplitude is 50 V.

(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, *[Math Processing Error]*, and *[Math Processing Error]*. 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 *[Math Processing Error]* 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 *[Math Processing Error]* 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 *[Math Processing Error]*.

(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-kΩ** 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.

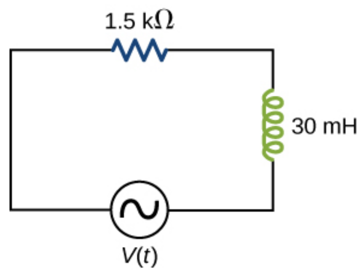
(b) Find the voltage drops across the resistor and inductor.

(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 *[Math Processing Error]*.

(a) What is the value of the resonance frequency *[Math Processing Error]*?

(b) What is the amplitude of the current if *[Math Processing Error]*?

(c) What is the phase constant of the current when *[Math Processing Error]*? 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 *[Math Processing Error]*.

(e) What is the power factor of the circuit when *[Math Processing Error]*?

(f) How much energy is used up by the resistor in 2.5 s when *[Math Processing Error]*?

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;

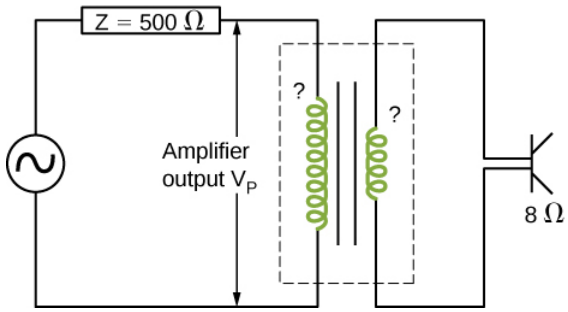
(c) 20-mH inductor with a frequency 6-Hz of the ac circuit;

(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

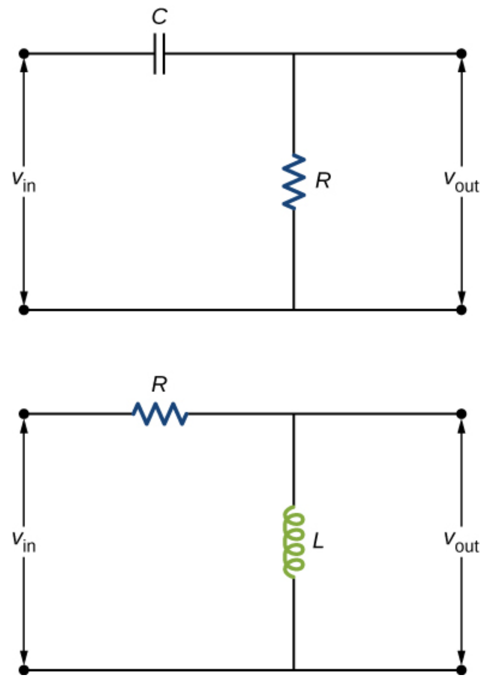
(c) inductor.

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 *[Math Processing Error]*, and the output voltage is vout.vout.

(a) Show that for the capacitor circuit, *[Math Processing Error]*, and for the inductor circuit, *[Math Processing Error]*.

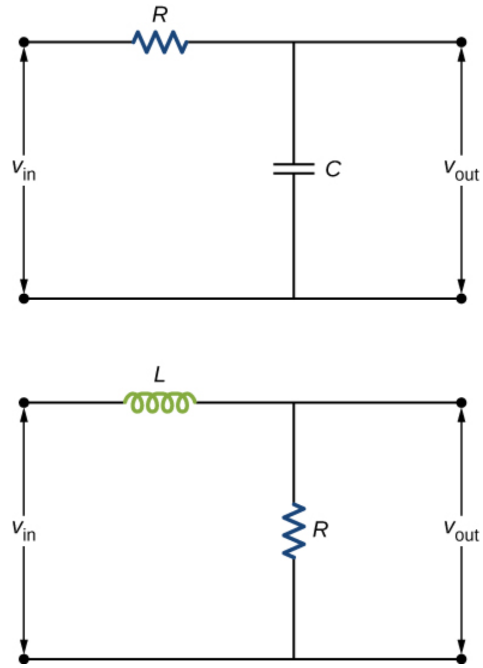
(b) Show that for high frequencies, *[Math Processing Error]* but for low frequencies, *[Math Processing Error]*.



72. The two circuits shown below act as crude low-pass filters. The input voltage to the circuits is *[Math Processing Error]*, and the output voltage is *[Math Processing Error]*.

(a) Show that for the capacitor circuit, *[Math Processing Error]*, and for the inductor circuit, *[Math Processing Error]*.

(b) Show that for low frequencies, *[Math Processing Error]*, but for high frequencies, *[Math Processing Error]*.



19.10: Alternating-Current Circuits (Answers)

Check Your Understanding

15.1. 10 ms

15.2. a. *[Math Processing Error]*;

b. *[Math Processing Error]*;

c. *[Math Processing Error]*

15.3. *[Math Processing Error]*

15.4. *[Math Processing Error]*

15.5. 2.00 V; 10.01 V; 8.01 V

15.6. a. 160 Hz;

b. *[Math Processing Error]*;

c. *[Math Processing Error]*;

d. 0.023 rad

15.7. a. halved;

b. halved;

c. same

15.8. *[Math Processing Error]*

15.9. a. 12:1;

b. 0.042 A;

c. *[Math Processing Error]*

Conceptual Questions

1. Angular frequency is *[Math Processing Error]* 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. *[Math Processing Error]*;

b. *[Math Processing Error]*;

c. *[Math Processing Error]*

17. a. *[Math Processing Error]*;

b. *[Math Processing Error]*;

c. *[Math Processing Error]*

19. 360 Hz

21. *[Math Processing Error]*

23. a. *[Math Processing Error]*;

b. *[Math Processing Error]*

25. a. *[Math Processing Error]*;

b. 0.16 A;

c. *[Math Processing Error]*;

d. *[Math Processing Error]*

27. a. *[Math Processing Error]*;

b. 0.15 A;

c. *[Math Processing Error]*;

d. *[Math Processing Error]*, 0.092 A, *[Math Processing Error]*

29. a. *[Math Processing Error]*;

b. *[Math Processing Error]*;

c. *[Math Processing Error]*

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. *[Math Processing Error]*

37. *[Math Processing Error]*

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. *[Math Processing Error]*

47. a. 41 turns;

b. 40.9 mA

Additional Problems

49. a. [Math Processing Error];

b. [Math Processing Error];

c. [Math Processing Error]

51. a. [Math Processing Error];

b. [Math Processing Error]

53. a. 19 A;

b. inductor leads by [Math Processing Error]

55. [Math Processing Error]

57. 36 W

59. a. [Math Processing Error];

b. [Math Processing Error]

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. [Math Processing Error];

b. 0.5 A;

c. [Math Processing Error], lagging;

d. [Math Processing Error];

e. 0.995;

f. 6.25 J

65. a. [Math Processing Error];

b. [Math Processing Error];

c. [Math Processing Error];

d. [Math Processing Error];

e. [Math Processing Error];

f. [Math Processing Error]

67. The units as written for inductive reactance Equation 15.8 are [Math Processing Error]H. Radians can be ignored in unit analysis. The Henry can be defined as [Math Processing Error]. Combining these together results in a unit of [Math Processing Error] for reactance.

69. a. 156 V;

b. 42 V;

c. 154 V

71. a. [Math Processing Error] and \(\displaystyle \frac{v_{out}}{v_{in}}=\frac{ωL}{\sqrt{R^2+ω^2L^2}};

b. [Math Processing Error] and [Math Processing Error]

- **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

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.

20.2: 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.

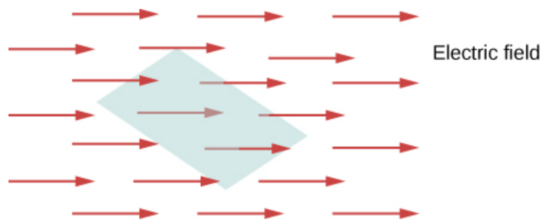


Figure 20.2.1: The flux of an electric field through the shaded area captures information about the "number" of electric field lines passing through the area. The numerical value of the electric flux depends on the magnitudes of the electric field and the area, as well as the relative orientation of 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 $\vec{E} = E\hat{y}$. 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 ($\mathrm{N \cdot m^2/C}$). 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.*

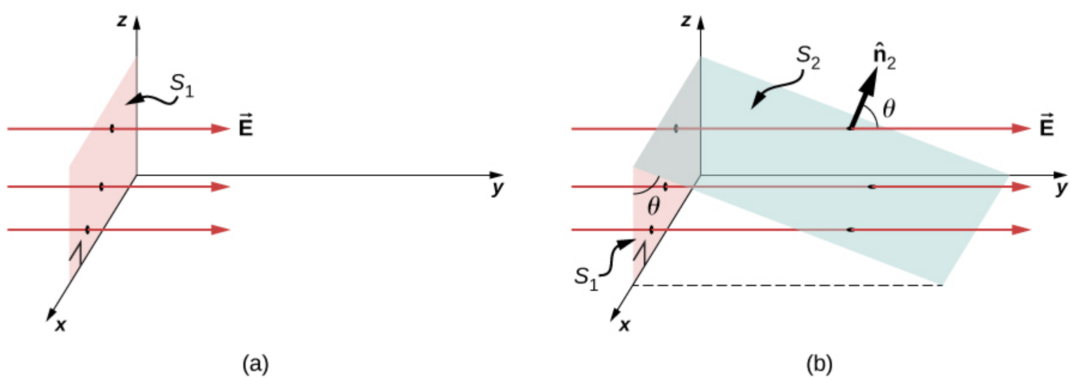


Figure 20.2.2: (a) A planar surface S_1 of area A_1 is perpendicular to the electric field $E\hat{j}$. N field lines cross surface S_1. (b) A surface S_2 of area A_2 whose projection onto the xz-plane is S_1. The same number of field lines cross each surface.

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 $\vec{A}$ 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 $20.2.3$.

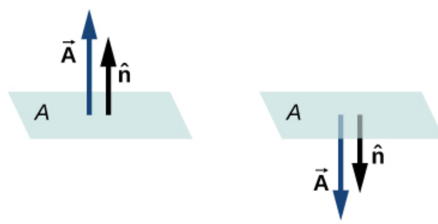


Figure $20.2.3$: The direction of the area vector of an open surface needs to be chosen; it could be either of the two cases displayed here. The area vector of a part of a closed surface is defined to point from the inside of the closed space to the outside. This rule gives a unique direction.

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.

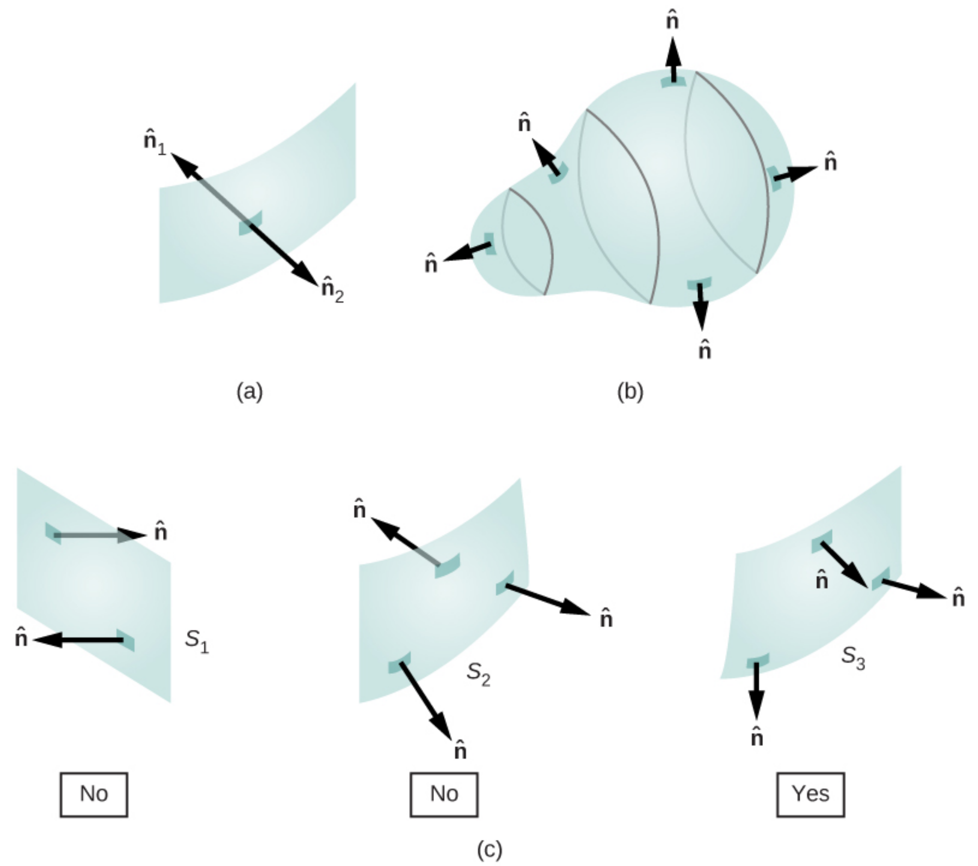


Figure 20.2.4: (a) Two potential normal vectors arise at every point on a surface. (b) The outward normal is used to calculate the flux through a closed surface. (c) Only S_3 has been given a consistent set of normal vectors that allows us to define the flux through the surface.

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?

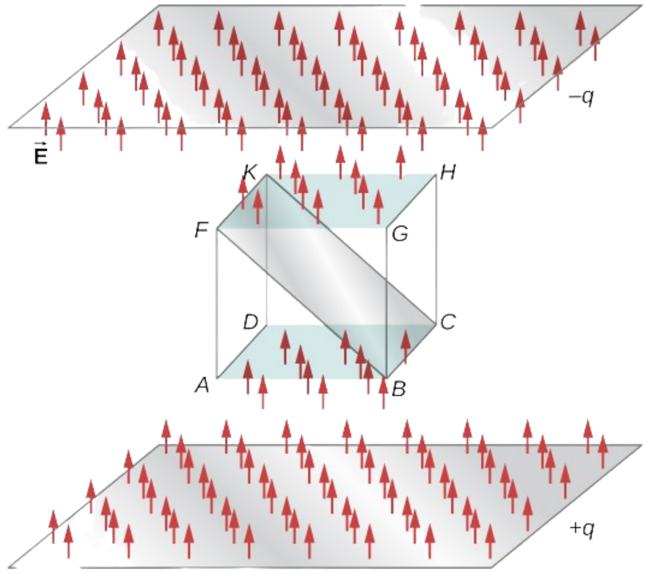


Figure 20.2.5: Electric flux through a cube, placed between two charged plates. Electric flux through the bottom face ($ABCD$) is negative, because $\vec{E}$ is in the opposite direction to the normal to the surface. The electric flux through the top face ($FGHK$) is positive, because the electric field 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.

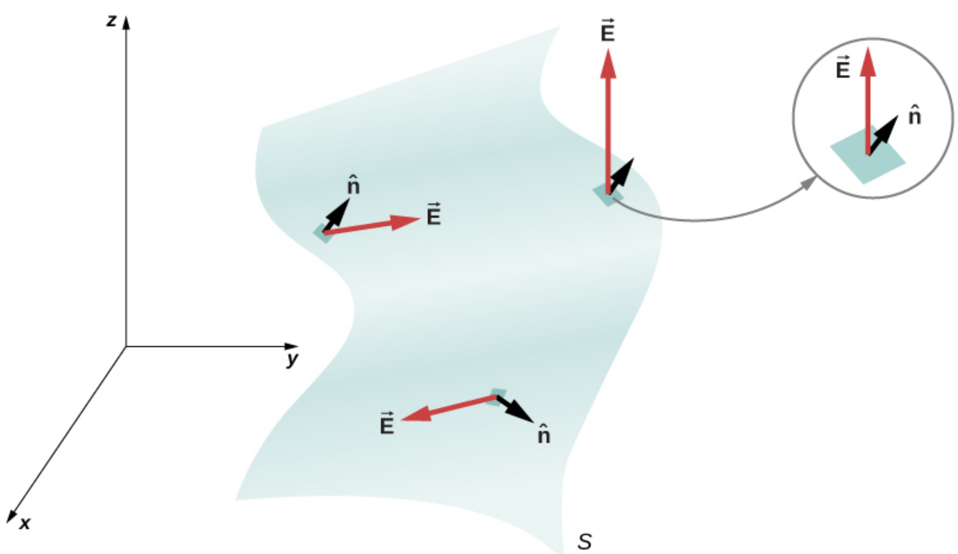


Figure 20.2.6: A surface is divided into patches to find the flux.

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 ith 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 ith 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 ith patch as

$$\Phi_i = \vec{E}_i \cdot \delta\vec{A}_i \;\; (i\text{th patch}).$$

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.

$$\Phi = \sum_{i=1}^{N} \Phi_i = \sum_{i=1}^{N} \vec{E}_i \cdot \delta\vec{A}_i \;\; (N \text{ patch estimate}).$$

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^o. 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,

$$\Phi = \int_S \vec{E} \cdot \hat{n}\, dA = \int_S \vec{E} \cdot d\vec{A} \;\; (\text{open surface}).$$

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

$$\Phi = \oint_S \vec{E} \cdot \hat{n}\, dA = \oint_S \vec{E} \cdot d\vec{A} \;\; (\text{closed surface})$$

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?

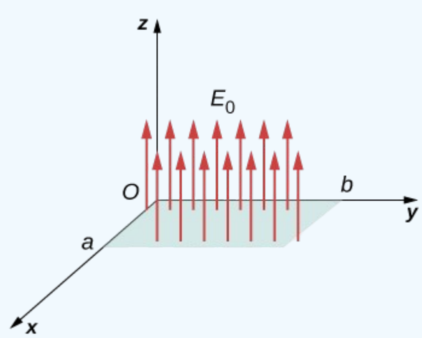


Figure 20.2.7: Calculating the flux of E_0 through a rectangular surface.

Strategy

Apply the definition of flux: $\Phi = \vec{E} \cdot \vec{A} \; (\text{uniform } \vec{E})$, 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 $20.2.8$). What is the net electric flux through a cube?

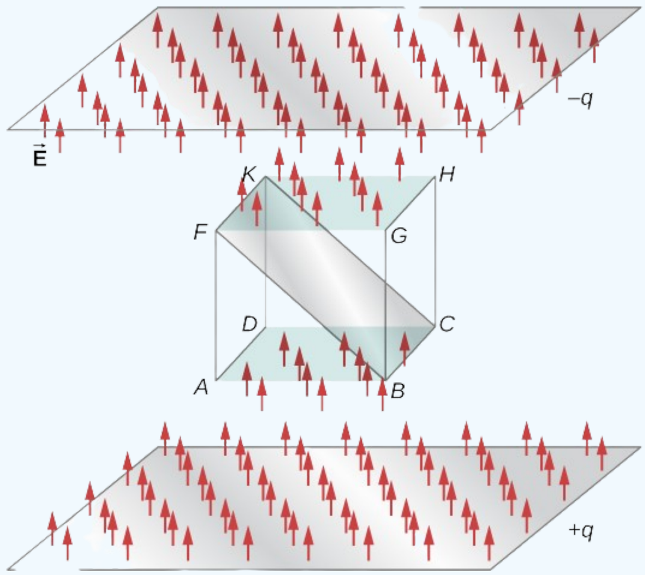


Figure 20.2.8: Calculating the flux of E_0 through a closed cubic surface.

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.

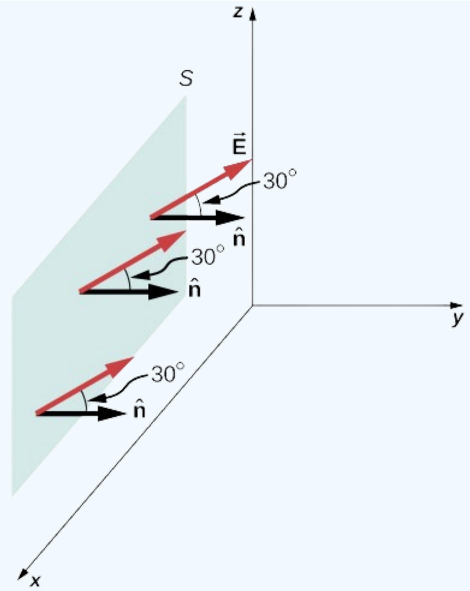


Figure 20.2.9: The electric field produces a net electric flux through the surface S.

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

$$\Phi=\int_S \vec{E}\cdot\hat{n}\,dA=EA\,\cos\theta$$

$$=(10\,\mathrm{N/C})(6.0\,\mathrm{m}^2)(\cos 30^\circ)=52\,\mathrm{N}\cdot\mathrm{m}^2/\mathrm{C}.$$

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 $20.2.10$?

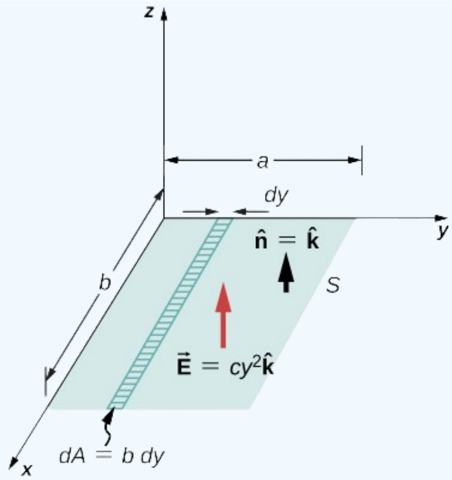


Figure 20.2.10: Since the electric field is not constant over the surface, an integration is necessary to determine the flux.

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

$$\begin{aligned}\Phi &= \int_S \vec{E} \cdot \hat{n} dA = \int_0^a (cy^2\hat{k}) \cdot \hat{k}(b\, dy) \\ &= cb \int_0^a y^2 dy = \frac{1}{3}a^3bc.\end{aligned}$$

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

$mab^2/2$

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 $20.3.1$.

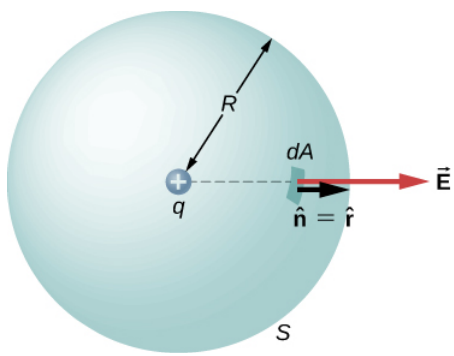


Figure 20.3.1: A closed spherical surface surrounding a point charge q.

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 $20.3.2$).

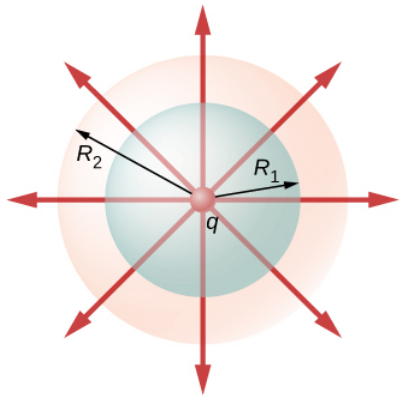


Figure $20.3.2$: Flux through spherical surfaces of radii R_1 and R_2 enclosing a charge q are equal, independent of the size of the surface, since all E-field lines that pierce one surface from the inside to outside direction also pierce the other surface in the same direction.

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 $20.3.3c$).

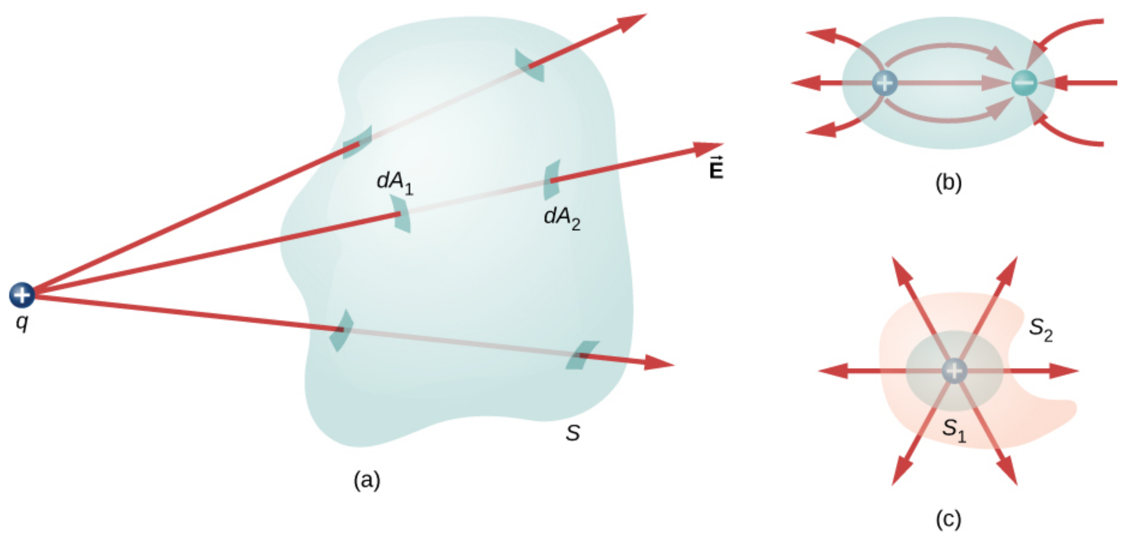


Figure $20.3.3$: Understanding the flux in terms of field lines. (a) The electric flux through a closed surface due to a charge outside that surface is zero. (b) Charges are enclosed, but because the net charge included is zero, the net flux through the closed surface is also zero. (c) The shape and 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.

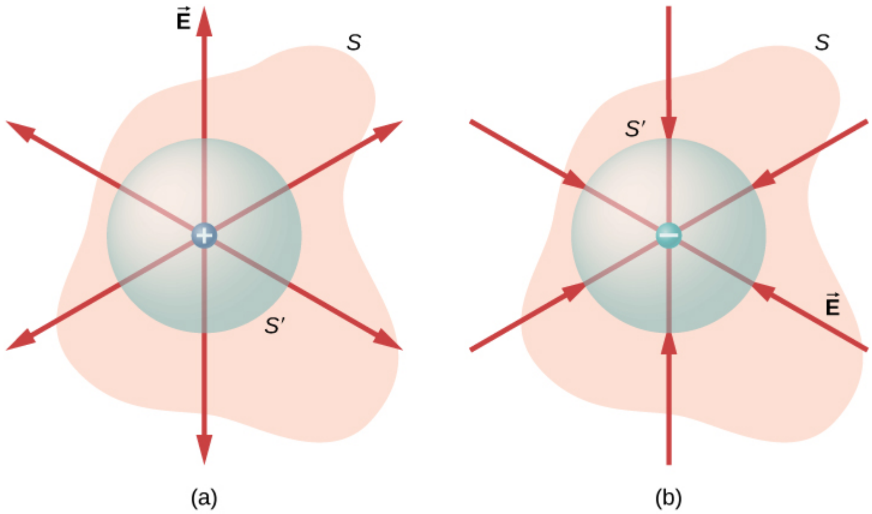


Figure 20.3.4: The electric flux through any closed surface surrounding a point charge q is given by Gauss's law. (a) Enclosed charge is positive. (b) Enclosed charge is negative.

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.

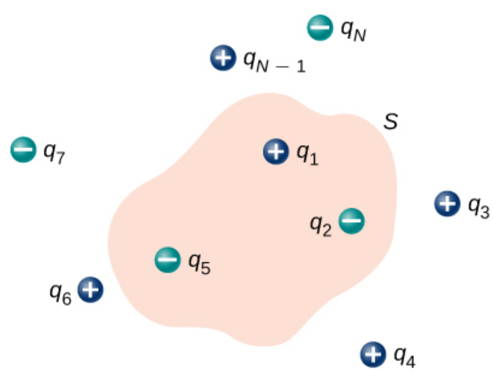


Figure 20.3.5: The flux through the Gaussian surface shown, due to the charge distribution, is $\Phi = (q_1 + q_2 + q_5)/\epsilon_0$.

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 $20.3.7$.

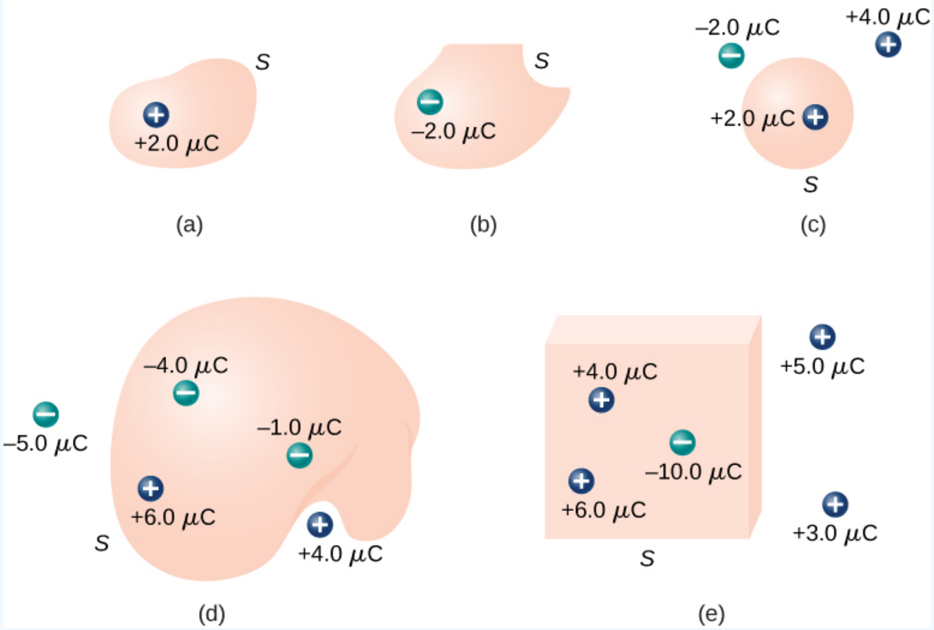


Figure 20.3.7: Various Gaussian surfaces and charges.

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. $\Phi = \frac{2.0\ \mu C}{\epsilon_0} = 2.3 \times 10^5\ \mathrm{N \cdot m^2/C}$.

b. $\Phi = \frac{-2.0\ \mu C}{\epsilon_0} = -2.3 \times 10^5\ \mathrm{N \cdot m^2/C}$.

c. $\Phi = \frac{2.0\ \mu C}{\epsilon_0} = 2.3 \times 10^5\ \mathrm{N \cdot m^2/C}$.

d. $\Phi = \frac{-4.0\ \mu C + 6.0\ \mu C - 1.0\ \mu C}{\epsilon_0} = 1.1 \times 10^5\ \mathrm{N \cdot m^2/C}$.

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

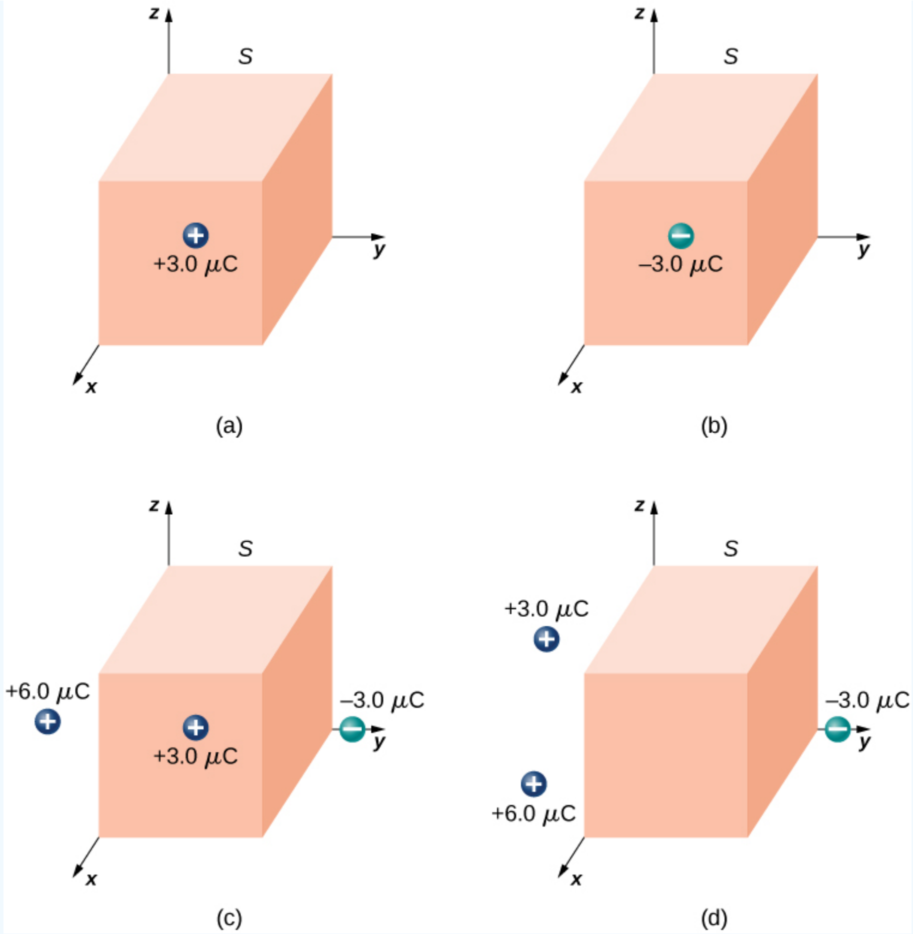


Figure 20.3.8: A cubical Gaussian surface with various charge distributions.

Answer a

$3.4 \times 10^5\ \mathrm{N \cdot m^2/C}$

Answer b

$-3.4 \times 10^5\ \mathrm{N \cdot m^2/C}$

Answer c

$3.4 \times 10^5\ \mathrm{N \cdot m^2/C}$

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.

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 directed counterclockwise and 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 **B** and the projection of **dl** onto the circle passing through $d\vec{l}$. If the radius of this particular circle is **r**, the projection is $r d\theta$, and

$$\vec{B} \cdot d\vec{l} = Br\, d\theta.$$

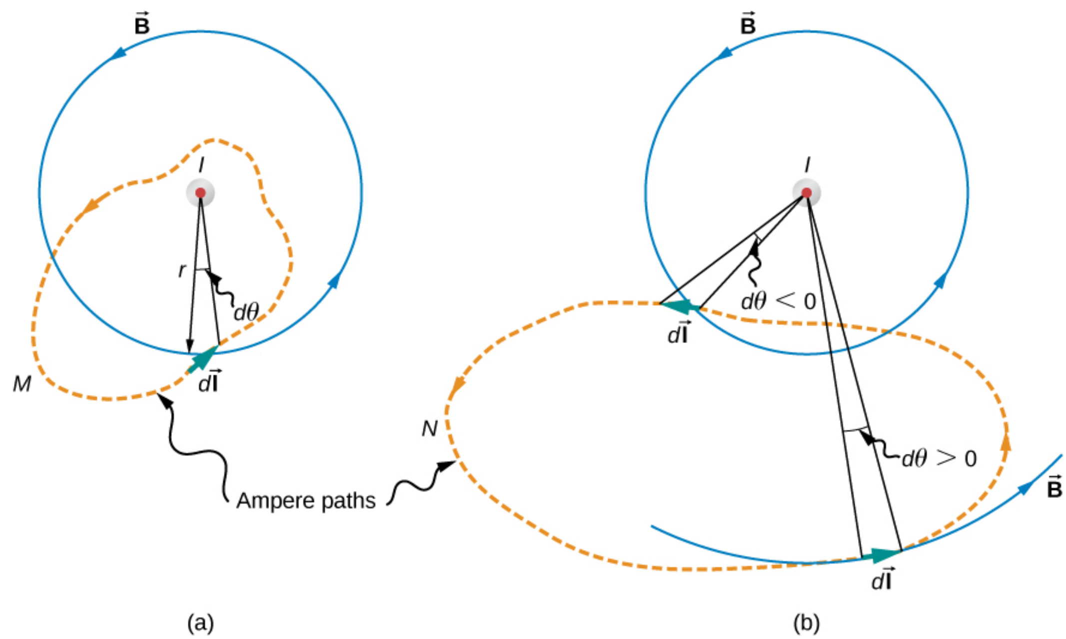


Figure 20.4.1: The current I of a long, straight wire is directed out of the page. The integral $\oint d\theta$ equals 2π and 0, respectively, for paths M and N.

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. Chose 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 $20.4.2$.

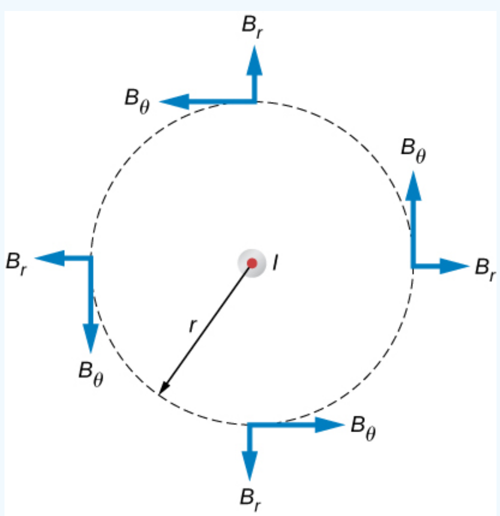


Figure 20.4.2: The possible components of the magnetic field **B** due to a current **I**, which is directed out of the page. The radial component is zero because the angle between the magnetic field and the path is at a right angle.

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 $\vec{B}_r \cdot d\vec{l} = 0$. 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.

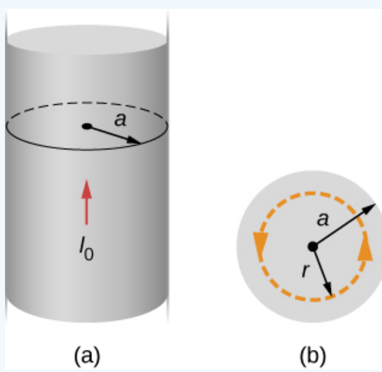


Figure 20.4.3: (a) A model of a current-carrying wire of radius **a** and current I_0. (b) A cross-section of the same wire showing the radius **a** and the Ampère's loop of radius **r**.

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 **B** with **r** is shown in Figure 20.4.4.

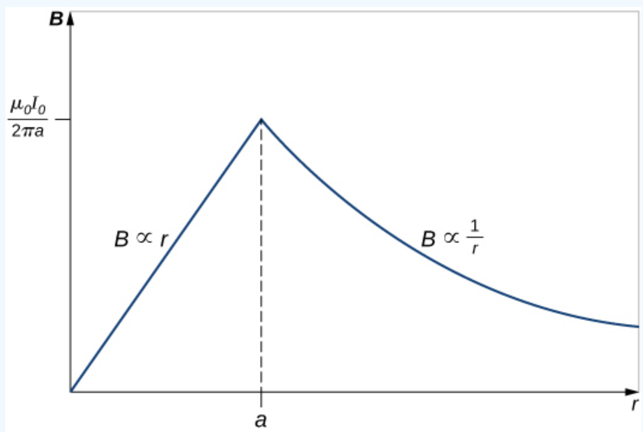


Figure 20.4.4: Variation of the magnetic field produced by a current I_0 in a long, straight wire of radius **a**.

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.

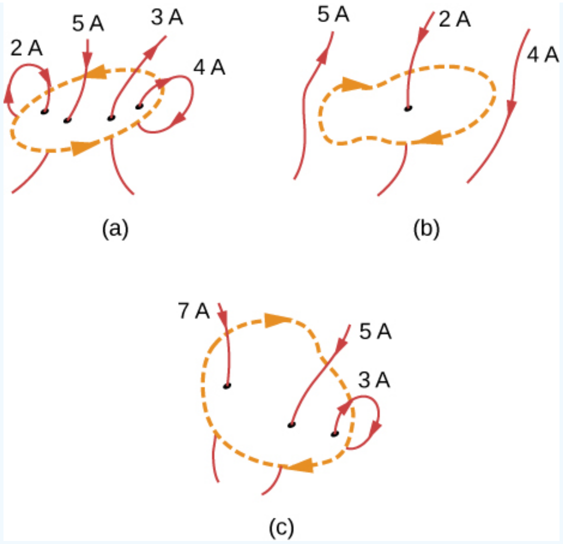


Figure 20.4.5: Current configurations and paths for Example 20.4.3.

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 $\oint \vec{B} \cdot d\vec{l} = \mu_0 (2\ A) = 2.51 \times 10^{-6} T \cdot m$.

(c) The right-hand rule shows us the current going downward through the loop is in the positive direction. There are $7A + 5A = 12A$ of current going downward and –3 A going upward. Therefore, the total current is 9 A and $\oint \vec{B} \cdot d\vec{l} = \mu_0 (9\ A) = 5.65 \times 10^{-6} T \cdot m$.

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 Ampèrian loop are very difficult because there is no symmetry, so this method would not be useful.

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:

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 I. \tag{20.5.1}$$

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 S_1 gives a nonzero value for the enclosed current **I**, whereas surface S_2 gives zero for the enclosed current because no current passes through it:

$$\underbrace{\oint_C \vec{B} \cdot d\vec{s} = \mu_0 I}_{\text{if surface } S_1 \text{is used}}$$

$$\underbrace{=0}_{\text{if surface } S_2 \text{is used}}$$

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.

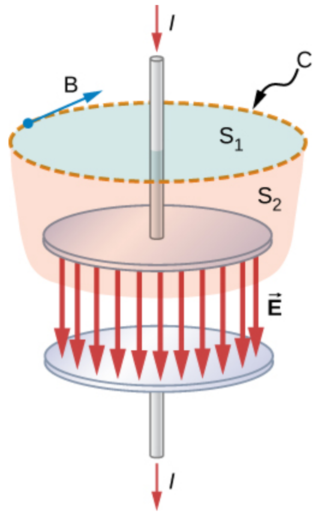


Figure 20.5.2: The currents through surface S_1 and surface S_2 are unequal, despite having the same boundary loop **C**.

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 **I**,

$$\boxed{\oint_S \vec{B} \cdot d\vec{s} = \mu_0 (I + I_d)} \tag{20.5.2}$$

where the displacement current is defined to be

$$\boxed{I_d = \epsilon_0 \frac{d\Phi_E}{dt}.} \tag{20.5.3}$$

Here ϵ_0 is the **permittivity of free space** and Φ_E is the electric flux, defined as

$$\Phi_E = \iint_{Surface\ S} \vec{E} \cdot d\vec{A}.$$

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 **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

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 I. \tag{20.5.4}$$

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

$$\oint_C \vec{B} \cdot d\vec{s} = \mu_0 \frac{d}{dt}\left[\epsilon_0 \iint_{Surface\ S_2} \vec{E} \cdot d\vec{A}\right].$$

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

$$\oint_{Surface\ S_1+S_2} \vec{E} \cdot d\vec{A} = \iint_{Surface\ S_1} \vec{E} \cdot d\vec{A} + \iint_{Surface\ S_2} \vec{E} \cdot d\vec{A} \tag{20.5.5}$$

$$= 0 + \iint_{Surface\ S_2} \vec{E} \cdot d\vec{A} \tag{20.5.6}$$

$$= \iint_{Surface\ S_2} \vec{E} \cdot d\vec{A}. \tag{20.5.7}$$

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 $t = 0$.

a. Find the displacement current between the capacitor plates at time **t**.
b. 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

a. The voltage between the plates at time **t** is given by

$$V_C = \frac{1}{C} Q(t) = V_0 \left(1 - e^{-t/RC}\right).$$

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

$$E_z(t) = \frac{V_0}{d}\left(1 - e^{-t/RC}\right).$$

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.

b. From the expression for V_C the charge on the capacitor is

$$Q(t) = CV_C = CV_0 \left(1 - e^{-t/RC}\right).$$

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

$$\oint \vec{E} \cdot d\vec{A} = \frac{Q_{in}}{\epsilon_0} \text{ (Gauss's law)} \tag{20.5.8}$$

$$\oint \vec{B} \cdot d\vec{A} = 0 \text{ (Gauss's law for magnetism)} \tag{20.5.9}$$

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d\Phi_m}{dt} \text{ (Faraday's law)} \tag{20.5.10}$$

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 I + \epsilon_0 I + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt} \text{ (Ampere-Maxwell law).} \tag{20.5.11}$$

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 **O**. This process may be visualized as the propagation of an electromagnetic wave through space.

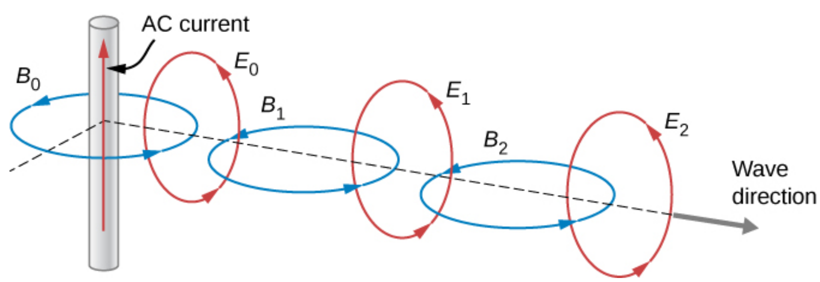


Figure $20.5.3$: How changing $\vec{E}$ and $\vec{B}$ fields propagate through space.

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.

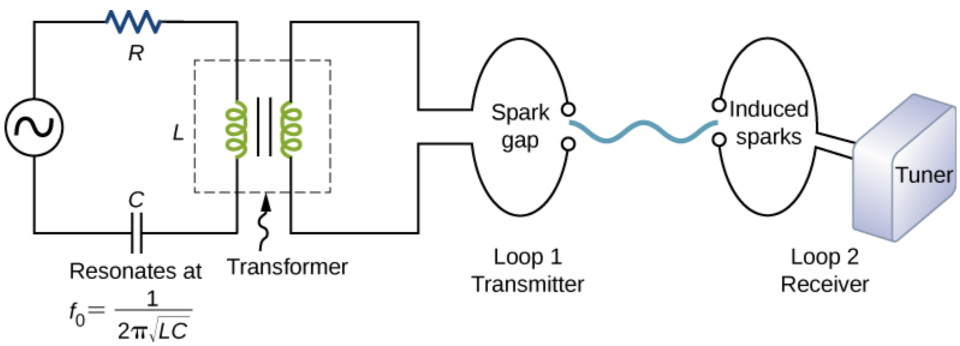


Figure $20.5.4$: The apparatus used by Hertz in 1887 to generate and detect electromagnetic waves.

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.

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

$$\text{Net flux} = -E_x(x,t)A + E_x(x+\Delta x,t)A = \frac{Q_{in}}{\epsilon_0} \tag{20.6.1}$$

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.

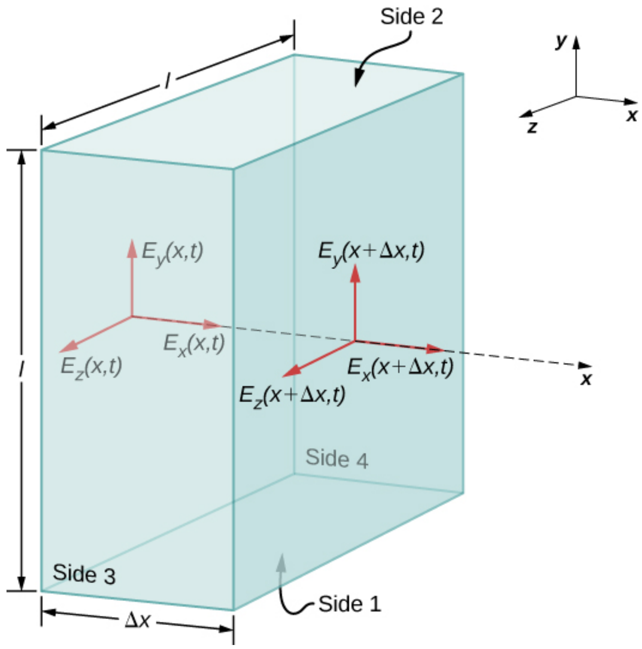


Figure 20.6.1: The surface of a rectangular box of dimensions $l \times l \times \Delta x$ is our Gaussian surface. The electric field shown is from an electromagnetic wave propagating along the x-axis.

A similar argument holds by substituting **E** for **B** and using Gauss's law for magnetism instead of Gauss's law for electric fields. This shows that the **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 **E** field from the changing **B** field, and for the **B** field from a changing **E** field. We then combine the two equations to show how the changing **E** and **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

$$\oint \vec{E} \cdot d\vec{s} = \frac{\partial E_y(x,t)}{\partial x}(l\Delta x).$$

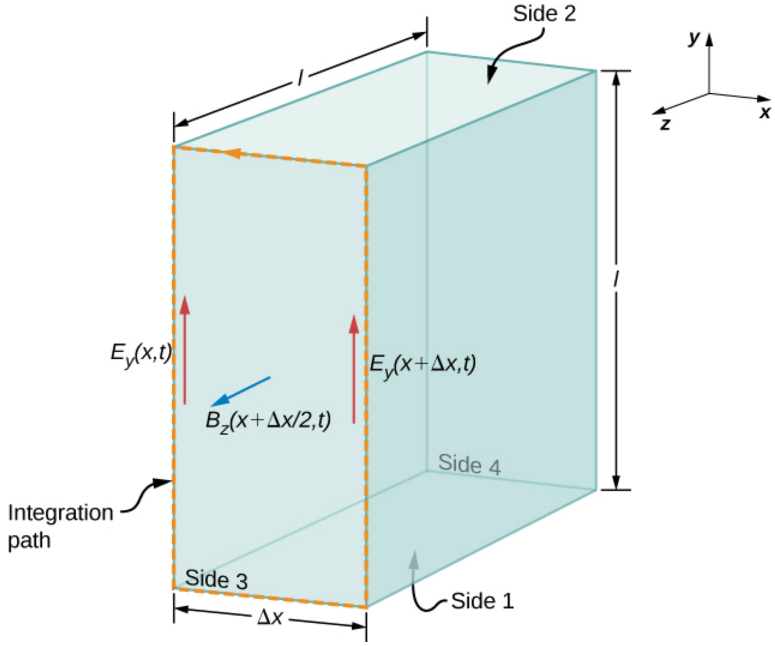


Figure 20.6.2: We apply Faraday's law to the front of the rectangle by evaluating $\oint \vec{E} \cdot d\vec{s}$ along the rectangular edge of Side 3 in the direction indicated, taking the B field crossing the face to be approximately its value in the middle of the area traversed.

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,

$$\oint_S \vec{B} \cdot \vec{n} dA = B_z\left(x+\frac{\Delta x}{2}, t\right)(l\Delta x). \tag{20.6.2}$$

From Faraday's law,

$$\oint \vec{E} \cdot d\vec{s} = -\frac{d}{dt}\int_S \vec{B} \cdot \vec{n} dA. \tag{20.6.3}$$

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

$$\frac{\partial E_y(x,t)}{\partial x} = -\frac{\partial B_z(x,t)}{\partial t}. \tag{20.6.4}$$

We could have applied Faraday's law instead to the top surface (numbered 2) in Figure $20.6.2$, to obtain the resulting equation

$$\frac{\partial B_z(x,t)}{\partial t} = -\frac{\partial E_y(x,t)}{\partial x}. \tag{20.6.5}$$

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,

$$\oint \vec{B} \cdot d\vec{s} = \mu_0 \epsilon_0 (d/dt) \int_S \vec{E} \cdot n\, da$$

to Surface 3, and then to Surface 2, yields the two equations

$$\frac{\partial E_y(x,t)}{\partial x} = -\epsilon_0 \mu_0 \frac{\partial E_z(x,t)}{\partial t}, \tag{20.6.6}$$

and

$$\frac{\partial B_z(x,t)}{\partial x} = -\epsilon_0 \mu_0 \frac{\partial E_y(x,t)}{\partial t}. \tag{20.6.7}$$

These equations describe the spatially dependent **B** field produced by the time-dependent **E** field.

We next combine the equations showing the changing **B** field producing an **E** field with the equation showing the changing **E** field producing a **B** field. Taking the derivative of Equation 20.6.4with respect to **x** and using Equation 20.6.13gives

$$\frac{\partial^2 E_y}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial E_y}{\partial x}\right) = -\frac{\partial}{\partial x}\left(\frac{\partial B_z}{\partial t}\right) = -\frac{\partial}{\partial t}\left(\frac{\partial B_z}{\partial x}\right) = \frac{\partial}{\partial t}\left(\epsilon_0 \mu_0 \frac{\partial E_y}{\partial t}\right)$$

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 **x – ct** for the wave traveling in the +**x**-direction, that is,

$$E_y(x,t) = f(\xi)\, where\, \xi = x - ct. \tag{20.6.8}$$

It is left as a mathematical exercise to show, using the chain rule for differentiation, that Equations 20.6.5and 20.6.6imply

$$1 = \epsilon_0 \mu_0 c^2.$$

The speed of the electromagnetic wave in free space is therefore given in terms of the permeability and the permittivity of free space by

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}}. \tag{20.6.9}$$

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 20.6.12and 20.6.11would 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 **E** field produced by the changing **B** field, (2) finding the **B** field produced by the changing **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:

$$E_y(x,t) = E_0 \cos(kx - \omega t). \tag{20.6.10}$$

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:

$$\omega = 2\pi f,\ f = \frac{1}{T},\ k = \frac{2\pi}{\lambda},\ and\ c = f\lambda = \omega/k.$$

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

$$\frac{\partial B_z}{\partial t} = -\frac{\partial E_y}{\partial x}$$

$$\frac{\partial B_z}{\partial t} = -\frac{\partial}{\partial x} E_0 \cos(kx - \omega t) = kE_0\, sin\,(kx - \omega t). \tag{20.6.11}$$

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

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

$$B_z(x,t) = B_0 \cos(kx - \omega t) \tag{20.6.12}$$

$$\frac{E_y}{B_z} = \frac{E_0}{B_0} = c. \tag{20.6.13}$$

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.

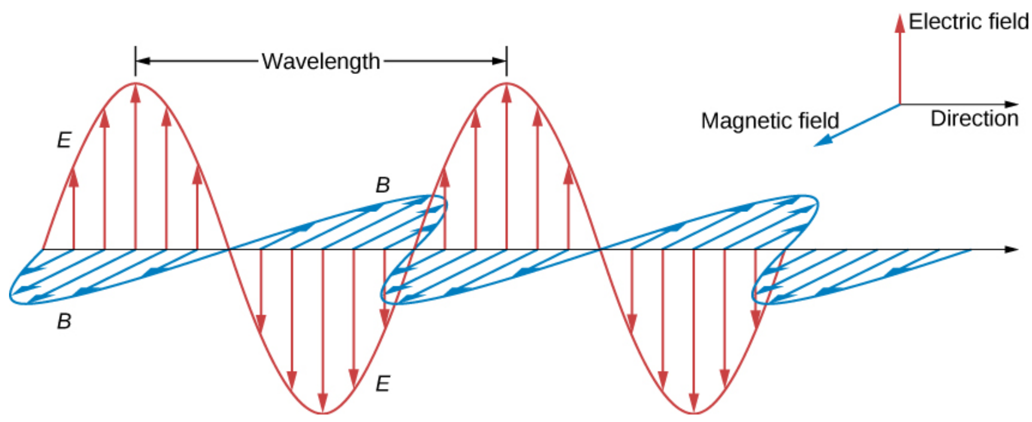


Figure 20.6.3: The plane wave solution of Maxwell's equations has the B field directly proportional to the E field at each point, with the relative directions shown.

✓ 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 20.6.10to 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

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

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:

a. the relative directions of wave propagation, of the **E** field, and of **B** field,
b. the speed of travel of the wave and how the speed depends on frequency, and
c. 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.

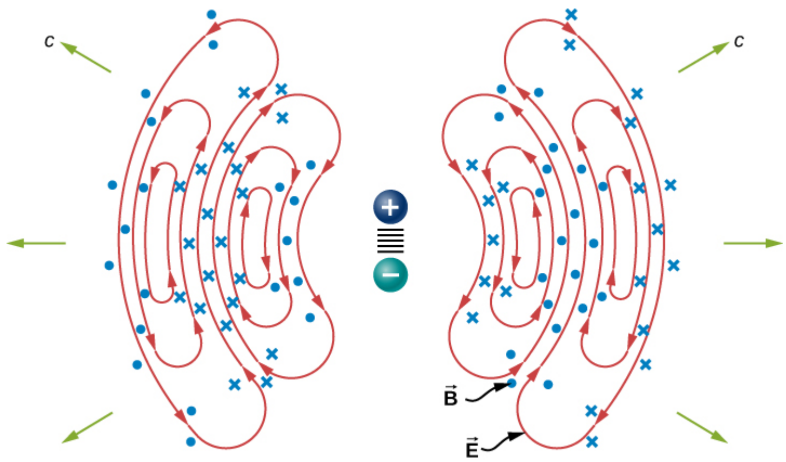


Figure 20.6.4: The oscillatory motion of the charges in a dipole antenna produces electromagnetic radiation.

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 tshe 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.

- **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.

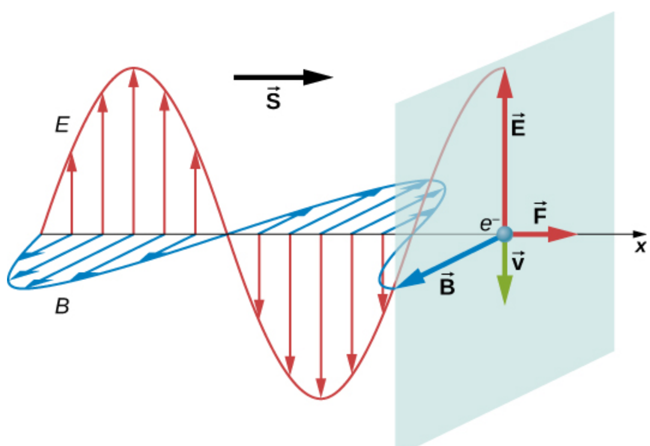


Figure 20.7.1: Electric and magnetic fields of an electromagnetic wave can combine to produce a force in the direction of propagation, as illustrated for the special case of electrons whose motion is highly damped by the resistance of a metal.

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.

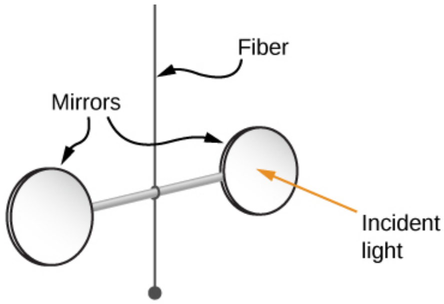


Figure 20.7.2: Simplified diagram of the central part of the apparatus Nichols and Hull used to precisely measure radiation pressure and confirm Maxwell's prediction.

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:

$$\underbrace{p_{rad} = u}_{\text{Perfect absorber}} . \tag{20.7.1}$$

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:

$$\underbrace{p_{rad} = 2u}_{\text{Perfect reflector}} . \tag{20.7.2}$$

We can confirm that the units are right:

$$[u] = \frac{J}{m^3} = \frac{N \cdot m}{m^3} = \frac{N}{m^2} = units\, of\, pressure.$$

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:

$$p = \langle p_{rad} \rangle = \begin{cases} I/c & \text{Perfect absorber} \\ 2I/c & \text{Perfect reflector} \end{cases} \tag{20.7.3}$$

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.

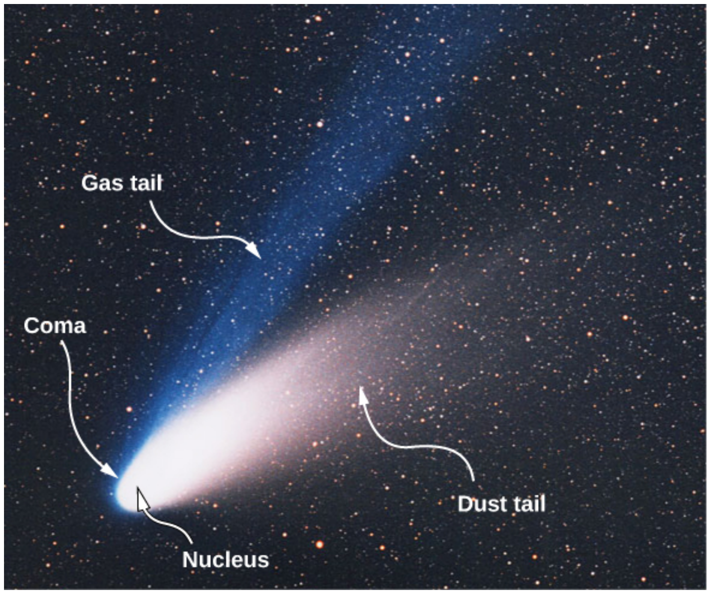


Figure 20.7.3: Evaporation of material being warmed by the Sun forms two tails, as shown in this photo of Comet Ison. (credit: modification of work by E. Slawik—ESO)

✓ 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 $3.8 \times 10^{26}\, W$.

a. Calculate the radiation pressure on the comet at this point in its orbit. Assume that the comet reflects all the incident light.
b. 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

$$F = pA$$
$$= (2.5 \times 10^{-5} N/m^2)(4.0 \times 10^{-2} m^2)$$
$$= 1.0 \times 10^{-6}\ N,$$

whereas the gravitational force of the Sun is

$$F_g = \frac{GMm}{r^2}$$
$$= \frac{(6.67 \times 10^{-11}\ N \cdot m^2/kg^2)(2.0 \times 10^{30} kg)(10\ kg)}{(9.0 \times 10^{10} m)^2}$$
$$= 0.16\ N.$$

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.

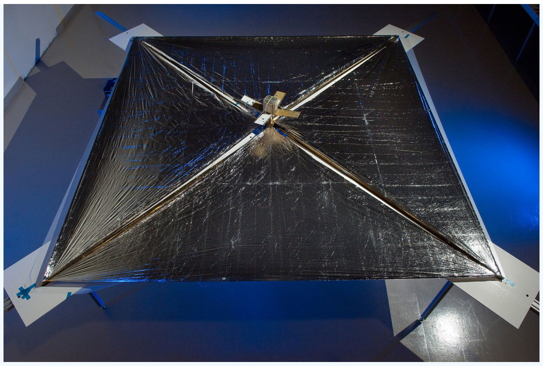
When the space program began in the 1960s, the idea started to receive serious attention from NASA. The most recent development in light propelled spacecraft has come from a citizen-funded group, the Planetary Society. It is currently testing the use of light sails to propel a small vehicle built from **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

$$F = pA = 2uA = \frac{2I}{c}A = \frac{2(1370\,W/m^2)(32\,m^2)}{(3.00\times 10^8 m/s)} = 2.92\times 10^{-4} N.$$

The resulting acceleration is

$$a = \frac{F}{m} = \frac{2.92\times 10^{-4} N}{5.0\,kg} = 5.8\times 10^{-5} m/s^2.$$

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.

CHAPTER OVERVIEW

21: Electrical Transmission Lines

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 lassociated 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.

21.2: 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.

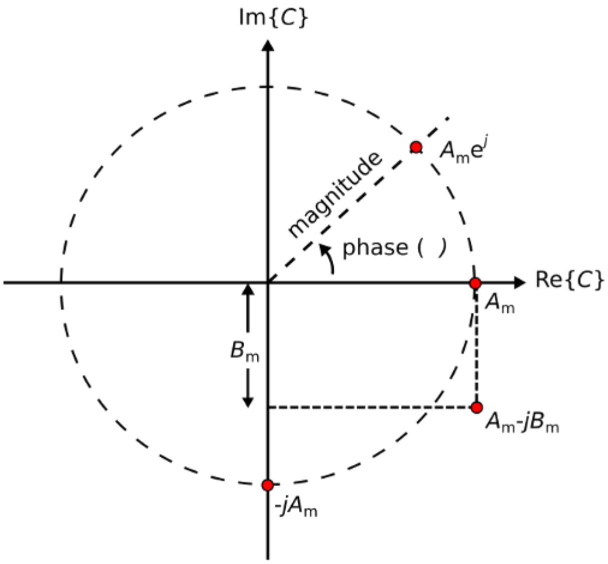


Figure $21.2.1$: Examples of phasors, displayed here as points in the real-imaginary plane.

Equation 21.2.1 is a completely general form for a physical (hence, real-valued) quantity varying sinusoidally with angular frequency $\omega = 2\pi f$

$$A(t;\omega) = A_m(\omega)\cos(\omega t + \psi(\omega)) \tag{21.2.1}$$

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 $\partial Am/\partial t = 0$; 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 $C(\omega)$:

$$C(\omega) = A_m(\omega)e^{j\psi(\omega)} \tag{21.2.2}$$

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:

$$\boxed{A(t;\omega) = \text{Re}\left\{C(\omega)e^{j\omega t}\right\}} \tag{21.2.3}$$

To see why this works, simply substitute the right hand side of Equation 21.2.2 into Equation 21.2.3. Then

$$
\begin{aligned}
A(t) &= \mathrm{Re}\left\{A_m(\omega)e^{j\psi(\omega)}e^{j\omega t}\right\} \\
&= \mathrm{Re}\left\{A_m(\omega)e^{j(\omega t+\psi(\omega))}\right\} \\
&= \mathrm{Re}\{A_m(\omega)[\cos(\omega t+\psi(\omega))+j\sin(\omega t+\psi(\omega))]\} \\
&= A_m(\omega)\cos(\omega t+\psi(\omega))
\end{aligned}
$$

as expected.

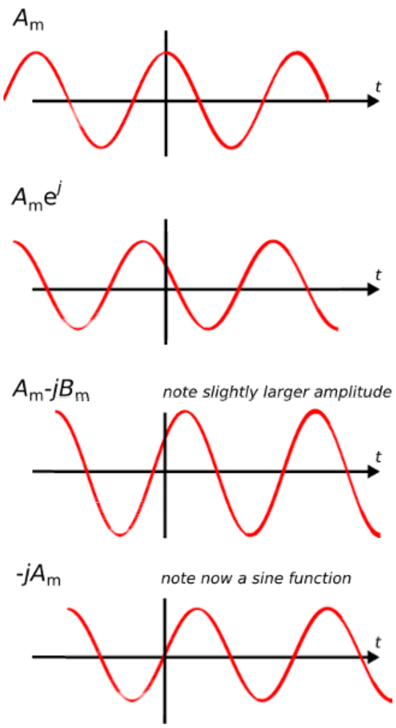


Figure 21.2.2 Sinusoids corresponding to the phasors shown in Figure 21.2.1

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, $\text{Re}\left\{C_1 e^{j\omega t}\right\} = \text{Re}\left\{C_2 e^{j\omega t}\right\}$ for all t. Then, $C_1 = C_2$.

Proof

Evaluating at $t=0$ we find $\text{Re}\{C_1\} = \text{Re}\{C_2\}$. 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

$$\text{Re}\left\{C_1 e^{j\omega t}\right\} = \text{Re}\{C_1 \cdot j\} = -\text{Im}\{C_1\}$$

and similarly

$$\text{Re}\left\{C_2 e^{j\omega t}\right\} = \text{Re}\{C_2 \cdot j\} = -\text{Im}\{C_2\}$$

therefore $\text{Im}\{C_1\} = \text{Im}\{C_2\}$. 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,

$$\mathcal{T}(\text{Re}\{C\}) = \text{Re}\{\mathcal{T}(C)\}. \tag{21.2.4}$$

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:

$$\begin{aligned} \text{Re}\{\mathcal{T}(C)\} &= \text{Re}\{\mathcal{T}(c_r + jc_i)\} \\ &= \text{Re}\{\mathcal{T}(c_r) + j\mathcal{T}(c_i)\} \\ &= \mathcal{T}(c_r) \\ &= \mathcal{T}(\text{Re}\{C\}) \end{aligned}$$

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:

$$\text{Re}\left\{\frac{\partial}{\partial\omega}C\right\} = \text{Re}\left\{\frac{\partial}{\partial\omega}(c_r + jc_i)\right\} = \frac{\partial}{\partial\omega}c_r$$

$$\frac{\partial}{\partial\omega}\text{Re}\{C\} = \frac{\partial}{\partial\omega}\text{Re}\{(c_r + jc_i)\} = \frac{\partial}{\partial\omega}c_r$$

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:

$$\int_{-\infty}^{+\infty} \mathrm{Re}\left\{C(\omega)e^{j\omega t}\right\}d\omega$$

We can go one step further using Claim 2:

$$\mathrm{Re}\left\{\int_{-\infty}^{+\infty} C(\omega)e^{j\omega t}d\omega\right\}$$

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.

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 $f = 10$ MHz over a length $l = 3$ cm, since l is only 0.1% of the wavelength $\lambda = c/f = 30$ 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 $f = 10$ MHz, $l = 3$ 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.

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.

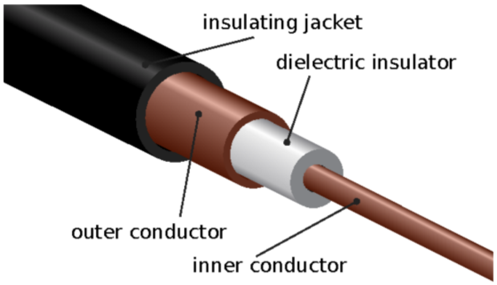


Figure $21.4.1$: Structure of a coaxial transmission line. (CC BY 3.0 (modified)).

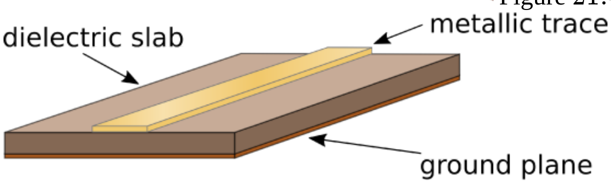


Figure $21.4.2$ Structure of a microstrip transmission line. (CC BY SA 3.0 (modified))

> 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.

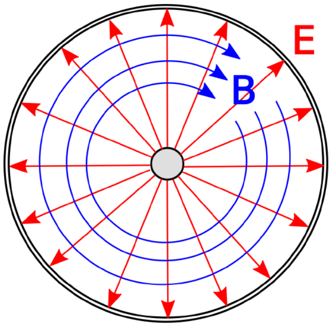


Figure $21.4.3$ Structure of the electric and magnetic fields within coaxial line. In this case, the wave is propagating away from the viewer.

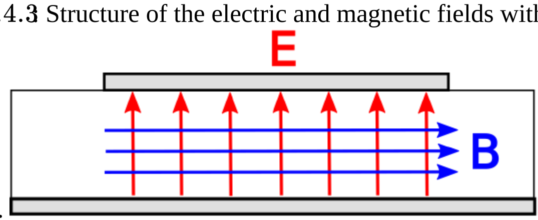


Figure $21.4.4$ Structure of the electric and magnetic fields within microstrip line. (The fields *outside* the line are possibly significant, complicated, and not shown.) In this case, the wave is propagating away from the viewer. (CC BY SA 3.0 Unported).

Figure $21.4.5$ A network of radio frequency waveguides in an air traffic control radar. (CC BY SA 2.0 Germany)

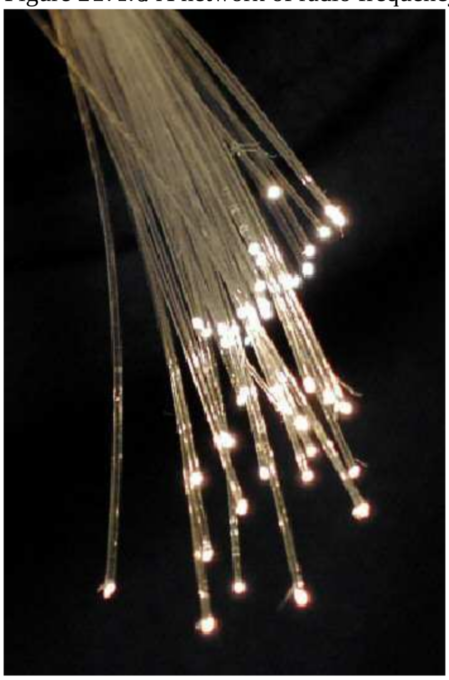


Figure $21.4.6$ Strands of optical fiber.

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.

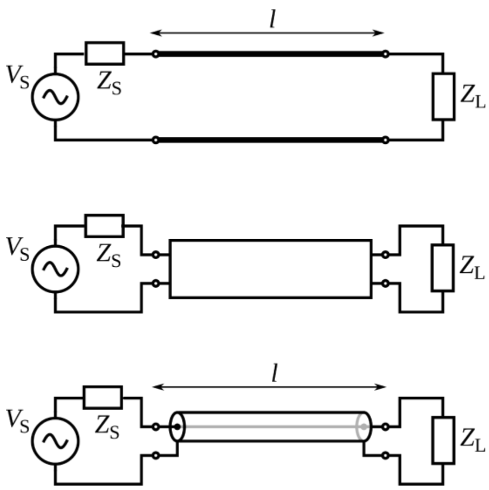


Figure $21.5.1$: Symbols representing transmission lines: Top: As a generic two-conductor direct connection. Middle: As a generic two-port “black box.” Bottom: As a coaxial cable. © CC BY SA 3.0 Unported (modified)

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.↩

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.

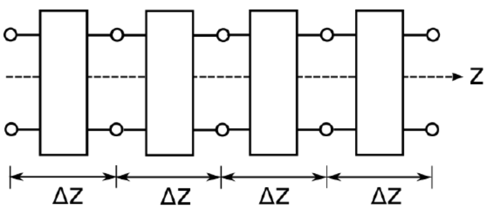


Figure $21.6.1$: Interpretation of a transmission line as a cascade of discrete series-connected two-ports.

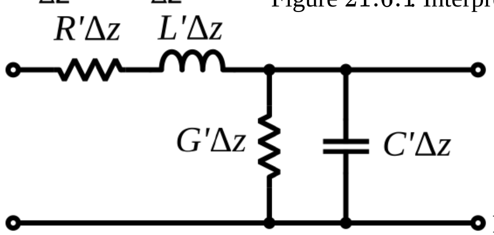


Figure $21.6.1$: Lumped-element equivalent circuit model for each of the two-ports in Figure $21.6.2$ (CC BY SA 3.0 Unported (modified))

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.

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 $21.7.1$.

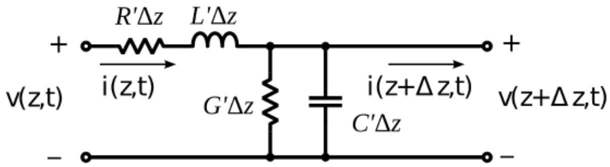


Figure $21.7.1$: Lumped-element equivalent circuit transmission line model, annotated with sign conventions for potentials and currents. (© CC BY SA 3.0 Unported (modified))

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:

$$\begin{gathered} v(z,t) - (R'\Delta z)\, i(z,t) - (L'\Delta z)\, \frac{\partial}{\partial t} i(z,t) \\ - v(z+\Delta z,t) = 0 \end{gathered}$$

Moving terms referring to current to the right side of the equation and then dividing through by Δz, we obtain

$$\begin{gathered} -\frac{v(z+\Delta z,t) - v(z,t)}{\Delta z} = \\ R'\, i(z,t) + L'\, \frac{\partial}{\partial t} i(z,t) \end{gathered}$$

Then taking the limit as $\Delta z \to 0$:

$$\boxed{-\frac{\partial}{\partial z} v(z,t) = R'\, i(z,t) + L'\, \frac{\partial}{\partial t} i(z,t)} \tag{21.7.1}$$

Applying Kirchoff's current law at the right port, we obtain:

$$\begin{gathered} i(z,t) - (G'\Delta z)\, v(z+\Delta z,t) - (C'\Delta z)\, \frac{\partial}{\partial t} v(z+\Delta z,t) \\ - i(z+\Delta z,t) = 0 \end{gathered}$$

Moving terms referring to potential to the right side of the equation and then dividing through by Δz, we obtain

$$\begin{gathered} -\frac{i(z+\Delta z,t) - i(z,t)}{\Delta z} = \\ G'\, v(z+\Delta z,t) + C'\, \frac{\partial}{\partial t} v(z+\Delta z,t) \end{gathered}$$

Taking the limit as $\Delta z \to 0$:

$$\boxed{-\frac{\partial}{\partial z} i(z,t) = G'\, v(z,t) + C'\, \frac{\partial}{\partial t} v(z,t)} \tag{21.7.2}$$

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 $\widetilde{V}(z)$ and $\tilde{I}(z)$ through the usual relationship:

$$v(z,t) = \text{Re}\left\{\widetilde{V}(z)\, e^{j\omega t}\right\}$$
$$i(z,t) = \text{Re}\left\{\tilde{I}\,(z)\, e^{j\omega t}\right\}$$

Now we see:

$$\begin{aligned}\frac{\partial}{\partial z} v(z,t) &= \frac{\partial}{\partial z} \text{Re}\left\{\widetilde{V}(z)\, e^{j\omega t}\right\} \\ &= \text{Re}\left\{\left[\frac{\partial}{\partial z}\widetilde{V}(z)\right]\, e^{j\omega t}\right\}\end{aligned}$$

In other words, $\partial v(z,t)/\partial z$ expressed in phasor representation is simply $\partial\widetilde{V}(z)/\partial z$; and

$$\begin{aligned}\frac{\partial}{\partial t} i(z,t) &= \frac{\partial}{\partial t} \text{Re}\left\{\tilde{I}\,(z)\, e^{j\omega t}\right\} \\ &= \text{Re}\left\{\frac{\partial}{\partial t}\left[\tilde{I}\,(z) e^{j\omega t}\right]\right\} \\ &= \text{Re}\left\{\left[j\omega\tilde{I}\,(z)\right]\, e^{j\omega t}\right\}\end{aligned}$$

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:

$$\boxed{-\frac{\partial}{\partial z}\widetilde{V}(z) = [R' + j\omega L']\;\tilde{I}\,(z)} \tag{21.7.3}$$

Following the same procedure, Equation 21.7.2 expressed in phasor representation is found to be:

$$\boxed{-\frac{\partial}{\partial z}\tilde{I}\,(z) = [G' + j\omega C']\;\widetilde{V}(z)} \tag{21.7.4}$$

> 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.↩

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 $\widetilde{V}(z)$ 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

$$-\frac{\partial}{\partial z}\widetilde{V}(z)=[R'+j\omega L']\ \tilde{I}(z) \tag{21.8.1}$$

$$-\frac{\partial}{\partial z}\tilde{I}(z)=[G'+j\omega C']\ \widetilde{V}(z) \tag{21.8.2}$$

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:

$$-\frac{\partial^2}{\partial z^2}\widetilde{V}(z)=[R'+j\omega L']\ \frac{\partial}{\partial z}\tilde{I}(z)$$

Then using Equation 21.8.2 to eliminate $\tilde{I}(z)$, we obtain

$$-\frac{\partial^2}{\partial z^2}\widetilde{V}(z)=-[R'+j\omega L'][G'+j\omega C']\ \widetilde{V}(z)$$

This equation is normally written as follows:

$$\boxed{\frac{\partial^2}{\partial z^2}\widetilde{V}(z)-\gamma^2\ \widetilde{V}(z)=0} \tag{21.8.3}$$

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*:

$$\gamma\triangleq\sqrt{(R'+j\omega L')(G'+j\omega C')} \tag{21.8.4}$$

> 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

$$\boxed{\frac{\partial^2}{\partial z^2}\tilde{I}(z)-\gamma^2\ \tilde{I}(z)=0} \tag{21.8.5}$$

> Equations 21.8.3 and 21.8.5 are the *wave equations* for $\widetilde{V}(z)$ and $\tilde{I}(z)$, respectively.

Note that both $\widetilde{V}(z)$ and $\tilde{I}(z)$ satisfy the *same* linear homogeneous differential equation. This does *not* mean that $\widetilde{V}(z)$ and $\tilde{I}(z)$ are equal. Rather, it means that $\widetilde{V}(z)$ and $\tilde{I}(z)$ can differ by no more than a multiplicative constant. Since $\widetilde{V}(z)$ 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

$$\widetilde{V}(z)=V_0^+e^{-\gamma z}+V_0^-e^{+\gamma z} \tag{21.8.6}$$

$$\tilde{I}(z)=I_0^+e^{-\gamma z}+I_0^-e^{+\gamma z} \tag{21.8.7}$$

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 $\widetilde{V}(z)$ 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^-.

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

$$\frac{\partial^2}{\partial z^2}\widetilde{V}(z) - \gamma^2\,\widetilde{V}(z) = 0 \tag{21.9.1}$$

where

$$\gamma \triangleq \sqrt{(R' + j\omega L')\,(G' + j\omega C')} \tag{21.9.2}$$

Equation 21.9.1 relates the potential phasor $\widetilde{V}(z)$ 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:

$$\frac{\partial^2}{\partial z^2}\tilde{I}\,(z) - \gamma^2\,\tilde{I}\,(z) = 0 \tag{21.9.3}$$

Since both $\widetilde{V}(z)$ and $\tilde{I}\,(z)$ satisfy the *same* linear homogeneous differential equation, they may differ by no more than a multiplicative constant. Since $\widetilde{V}(z)$ 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

$$\widetilde{V}(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \tag{21.9.4}$$

$$\tilde{I}\,(z) = I_0^+ e^{-\gamma z} + I_0^- e^{+\gamma z} \tag{21.9.5}$$

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 $\widetilde{V}(z)$ and $\tilde{I}\,(z)$ at some position(s) along the line. Also, we will make use of the telegrapher's equations (Section 3.5):

$$-\frac{\partial}{\partial z}\widetilde{V}(z) = [R' + j\omega L']\;\tilde{I}\,(z) \tag{21.9.6}$$

$$-\frac{\partial}{\partial z}\tilde{I}\,(z) = [G' + j\omega C']\;\widetilde{V}(z) \tag{21.9.7}$$

We begin by differentiating Equation 21.9.4 with respect to z, which yields

$$\frac{\partial}{\partial z}\widetilde{V}(z) = -\gamma\left[V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}\right]$$

Now we use this to eliminate $\partial\widetilde{V}(z)/\partial z$ in Equation 21.9.6, yielding

$$\gamma\left[V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}\right] = [R' + j\omega L']\;\tilde{I}\,(z)$$

Solving the above equation for $\tilde{I}\,(z)$ yields:

$$\tilde{I}\,(z) = \frac{\gamma}{R' + j\omega L'}\left[V_0^+ e^{-\gamma z} - V_0^- e^{+\gamma z}\right]$$

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

$$Z_0 = \frac{R' + j\omega L'}{\gamma} \tag{21.9.8}$$

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 $\widetilde{V}(z)$ 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*.

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

$$\widetilde{V}(z) = V_0^+ e^{-\gamma z} + V_0^- e^{+\gamma z} \tag{21.10.1}$$

$$\tilde{I}(z) = I_0^+ e^{-\gamma z} + I_0^- e^{+\gamma z} \tag{21.10.2}$$

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.,

$$\alpha \triangleq \text{Re}\{\gamma\}$$

$$\beta \triangleq \text{Im}\{\gamma\}$$

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:

$$\text{Re}\left\{e^{\pm\gamma z} e^{j\omega t}\right\} = e^{\pm\alpha z}\cos(\omega t \pm \beta z)$$

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.

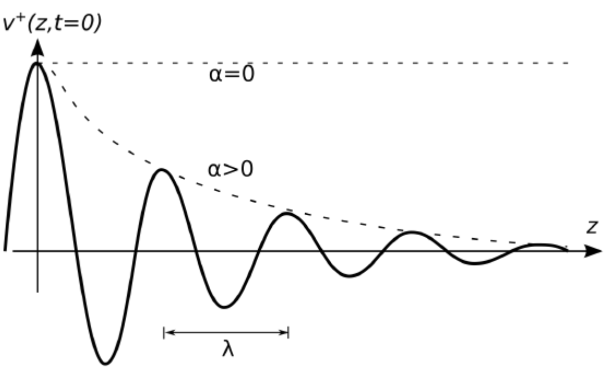


Figure $21.10.1$ The potential $v^+(z,t)$ of the wave traveling in the $+z$ direction at $t=0$ for $\psi=0$.

Let's define $\widetilde{V}^+(z)$ and $\tilde{I}^+(z)$ to be the potential and current associated with a wave propagating in the $+z$ direction. Then:

$$\widetilde{V}^+(z) \triangleq V_0^+ e^{-\gamma z}$$

or equivalently in the time domain:

$$\begin{aligned} v^+(z,t) &= \text{Re}\left\{\widetilde{V}^+(z)e^{j\omega t}\right\} \\ &= \text{Re}\left\{V_0^+ e^{-\gamma z} e^{j\omega t}\right\} \\ &= \left|V_0^+\right| e^{-\alpha z}\cos(\omega t - \beta z + \psi) \end{aligned}$$

where ψ is the phase of V_0^+. Figure $21.10.1$shows $v^+(z,t)$. From fundamental wave theory we recognize

> $\beta \triangleq \text{Im}\{\gamma\}$ (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

$$\boxed{\lambda = \frac{2\pi}{\beta}}$$

Also we recognize:

> $\alpha \triangleq \text{Re}\{\gamma\}$ (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:

$$\boxed{v_p = \frac{\omega}{\beta}}$$

Having previously noted that $\beta = 2\pi/\lambda$, the above expression also yields the expected result

$$\boxed{v_p = \lambda f}$$

> 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 $\widetilde{V}^-(z)$ 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 $\widetilde{V}(z)$ 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:

$$\widetilde{V}(z) = \widetilde{V}^{+}(z) + \widetilde{V}^{-}(z)$$

$$\tilde{I}(z) = \tilde{I}^{+}(z) + \tilde{I}^{-}(z)$$

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.

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:

$$R' \ll \omega L' \tag{21.11.1}$$

$$G' \ll \omega C' \tag{21.11.2}$$

When these conditions are met, the propagation constant simplifies as follows:

$$\begin{aligned} \gamma &\approx \sqrt{(j\omega L')(j\omega C')} \\ &= \sqrt{-\omega^2 L'C'} \\ &= j\omega\sqrt{L'C'} \end{aligned}$$

and subsequently

$$\alpha \triangleq \operatorname{Re}\{\gamma\} \approx 0 \qquad \text{(low-loss approx.)} \tag{21.11.3}$$

$$\beta \triangleq \operatorname{Im}\{\gamma\} \approx \omega\sqrt{L'C'} \qquad \text{(low-loss approx.)} \tag{21.11.4}$$

$$v_p = \omega/\beta \approx \frac{1}{\sqrt{L'C'}} \qquad \text{(low-loss approx.)} \tag{21.11.5}$$

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 21.11.1and 21.11.2and 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 21.11.1and 21.11.2 In this case, the low-loss expression for β is used, but α might not be approximated as zero.

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?

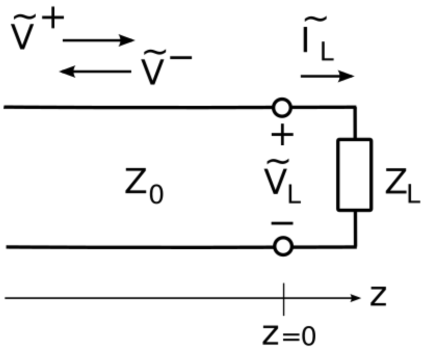


Figure 21.12.1: A wave arriving from the left incident on a termination located at $z=0$.

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

$$\widetilde{V}^{+}(z) = V_0^{+} e^{-j\beta z}$$

where V_0^{+} is determined by the source of the wave, and so is effectively a "given." Any reflected wave must have the form

$$\widetilde{V}^{-}(z) = V_0^{-} e^{+j\beta z}$$

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

$$Z_L \triangleq \frac{\widetilde{V}_L}{\tilde{I}_L} \tag{21.12.1}$$

where $\widetilde{V}_L$ 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,

$$\widetilde{V}^{+}(0) + \widetilde{V}^{-}(0) = \widetilde{V}_L \tag{21.12.2}$$

$$\tilde{I}^{+}(0) + \tilde{I}^{-}(0) = \tilde{I}_L \tag{21.12.3}$$

where $\tilde{I}^{+}(z)$ and $\tilde{I}^{-}(z)$ are the currents associated with $\widetilde{V}^{+}(z)$ and $\widetilde{V}^{-}(z)$, respectively. Since the voltage and current are related by Z_0, Equation 21.12.3 may be rewritten as follows:

$$\frac{\widetilde{V}^{+}(0)}{Z_0} - \frac{\widetilde{V}^{-}(0)}{Z_0} = \tilde{I}_L \tag{21.12.4}$$

Evaluating the left sides of Equations 21.12.2 and 21.12.4 at $z=0$, we find:

$$V_0^{+} + V_0^{-} = \widetilde{V}_L$$

$$\frac{V_0^{+}}{Z_0} - \frac{V_0^{-}}{Z_0} = \tilde{I}_L$$

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

$$V_0^- = \Gamma V_0^+ \text{ , where}$$

$$\boxed{\Gamma \triangleq \frac{Z_L - Z_0}{Z_L + Z_0}} \tag{21.12.5}$$

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 21.12.5, 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 $Z_L \to \infty$, and subsequently $\Gamma \to +1$. 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 $Z_L/Z_0\to\infty$ (i.e., an open circuit) or when $Z_L/Z_0=0$ (a short circuit); the result in either case is $|\Gamma|=1$. Thus,

$$\boxed{0\leq|\Gamma|\leq 1}$$

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

$$\begin{aligned}\widetilde{V}(z) &= V_0^+ e^{-j\beta z} + \Gamma V_0^+ e^{+j\beta z} \\ &= V_0^+ \left(e^{-j\beta z} + \Gamma e^{+j\beta z}\right)\end{aligned}$$

The magnitude of $\widetilde{V}(z)$ is most easily found by first finding $|\widetilde{V}(z)|^2$, which is:

$$\begin{aligned}&\tilde{V}(z)\tilde{V}^*(z) \\ &= |V_0^+|^2 \left(e^{-j\beta z} + \Gamma e^{+j\beta z}\right)\left(e^{-j\beta z} + \Gamma e^{+j\beta z}\right)^* \\ &= |V_0^+|^2 \left(e^{-j\beta z} + \Gamma e^{+j\beta z}\right)\left(e^{+j\beta z} + \Gamma^* e^{-j\beta z}\right) \\ &= |V_0^+|^2 \left(1 + |\Gamma|^2 + \Gamma e^{+j2\beta z} + \Gamma^* e^{-j2\beta z}\right)\end{aligned}$$

Let ϕ be the phase of Γ; i.e.,

$$\Gamma = |\Gamma|\, e^{j\phi}$$

Then, continuing from the previous expression:

$$\begin{aligned}&|V_0^+|^2 \left(1 + |\Gamma|^2 + |\Gamma| e^{+j(2\beta z+\phi)} + |\Gamma| e^{-j(2\beta z+\phi)}\right) \\ &= |V_0^+|^2 \left(1 + |\Gamma|^2 + |\Gamma| \left[e^{+j(2\beta z+\phi)} + e^{-j(2\beta z+\phi)}\right]\right)\end{aligned}$$

The quantity in square brackets can be reduced to a cosine function using the identity

$$\cos\theta = \frac{1}{2}\left[e^{j\theta} + e^{-j\theta}\right]$$

yielding:

$$|V_0^+|^2 \left[1 + |\Gamma|^2 + 2\,|\Gamma| \cos(2\beta z + \phi)\right]$$

Recall that this is $|\widetilde{V}(z)|^2$. $|\widetilde{V}(z)|$ is therefore the square root of the above expression:

$$\left|\tilde{V}(z)\right| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2\,|\Gamma| \cos(2\beta z + \phi)}$$

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:

$$\left|\tilde{I}(z)\right| = \frac{|V_0^+|}{Z_0} \sqrt{1 + |\Gamma|^2 - 2\,|\Gamma| \cos(2\beta z + \phi)}$$

Again we find the result is a standing wave.

Now let us consider the outcome for a few special cases.

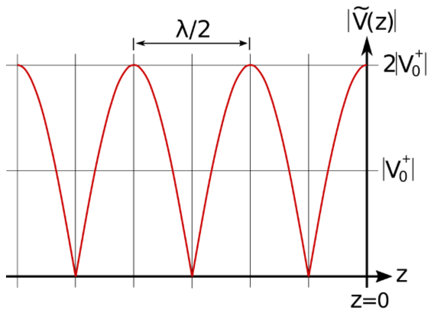
Matched load. When the impedance of the termination of the transmission line, Z_L, is equal to the characteristic impedance of the transmission line, Z_0, $\Gamma = 0$ and there is no reflection. In this case, the above expressions reduce to $|\widetilde{V}(z)| = |V_0^+|$ and $|\tilde{I}(z)| = |V_0^+|/Z_0$, as expected.

Open or Short-Circuit. In this case, $\Gamma = \pm 1$ and we find:

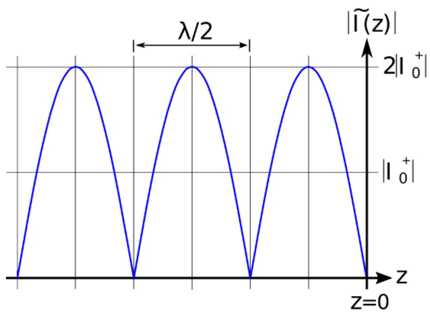
$$\left|\widetilde{V}(z)\right| = |V_0^+| \sqrt{2 + 2\cos(2\beta z + \phi)}$$

$$\left|\tilde{I}(z)\right| = \frac{|V_0^+|}{Z_0} \sqrt{2 - 2\cos(2\beta z + \phi)}$$

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 $21.13.2$

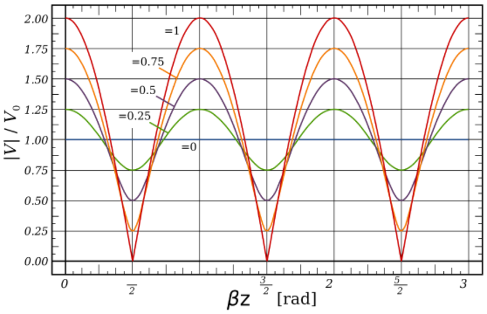


Figure $21.13.2$ Standing waves associated with loads exhibiting various reflection coefficients. In this figure the incident wave arrives from the right.

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:

$$\boxed{\text{SWR} \triangleq \frac{\text{maximum } |\widetilde{V}|}{\text{minimum } |\widetilde{V}|}}$$

SWR can be calculated using a simple expression, which we shall now derive. In Section 3.13, we found that:

$$\left|\widetilde{V}(z)\right| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2\,|\Gamma| \cos(2\beta z + \phi)}$$

The maximum value occurs when the cosine factor is equal to $+1$, yielding:

$$\max \left|\widetilde{V}\right| = |V_0^+| \sqrt{1 + |\Gamma|^2 + 2\,|\Gamma|}$$

Note that the argument of the square root operator is equal to $(1 + |\Gamma|)^2$; therefore:

$$\max \left|\widetilde{V}\right| = |V_0^+|\,(1 + |\Gamma|)$$

Similarly, the minimum value is achieved when the cosine factor is equal to -1, yielding:

$$\min \left|\widetilde{V}\right| = |V_0^+| \sqrt{1 + |\Gamma|^2 - 2\,|\Gamma|}$$

So:

$$\min \left|\widetilde{V}\right| = |V_0^+|\,(1 - |\Gamma|)$$

Therefore:

$$\boxed{\text{SWR} = \frac{1 + |\Gamma|}{1 - |\Gamma|}} \tag{21.14.1}$$

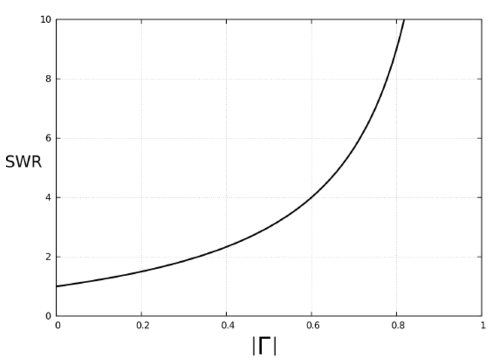


Figure 21.14.1: Relationship between SWR and $|\Gamma|$.

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:

$$|\Gamma| = \frac{\text{SWR} - 1}{\text{SWR} + 1} \tag{21.14.2}$$

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 $|\Gamma| = 1/2$.

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.

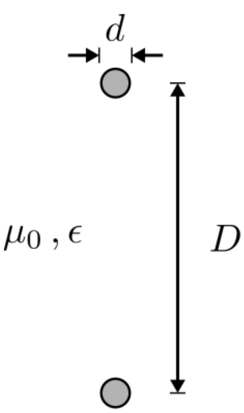


Figure $21.15.2$ Parallel wire transmission line structure and design parameters. (CC BY SA 4.0; C. Wang)

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 $21.15.3$

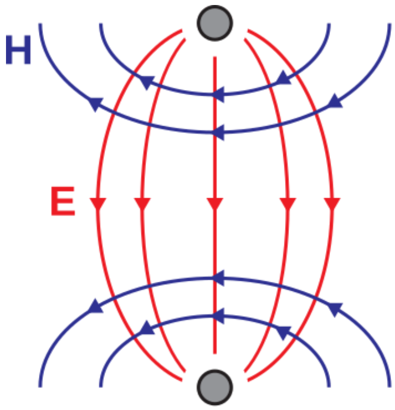


Figure 21.15.3 Structure of the electric and magnetic fields for a cross-section of parallel wire line. In this case, the wave is propogating away from the viewer. (CC BY SA 4.0; S. Lally)

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

$$\boxed{Z_0 \approx \frac{1}{\pi} \frac{\eta_0}{\sqrt{\epsilon_r}} \ln(2D/d)} \tag{21.15.2}$$

> 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 $300\ \Omega$ twin-lead. Although implementations vary, the wire diameter is usually about 1 mm and 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 21.15.2gives $Z_0 \approx 298\ \Omega$, 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.↩

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.

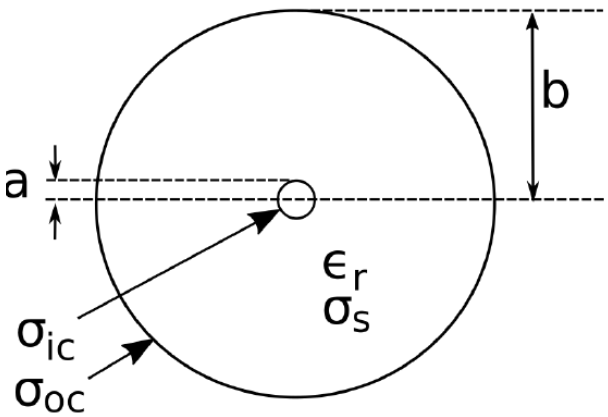


Figure $21.16.1$: Parameters defining the design of a coaxial cable.

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:

$$\delta_{ic} = \sqrt{2/\omega\mu\sigma_{ic}} \tag{21.16.2}$$

$$\delta_{oc} = \sqrt{2/\omega\mu\sigma_{oc}} \tag{21.16.3}$$

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

$$\begin{aligned} R' &\approx \frac{1}{2\pi\sqrt{2/\omega\mu_0}}\left[\frac{1}{a\sqrt{\sigma_{ic}}}\right] \\ &= \frac{1}{2\pi\delta_{ic}\sigma_{ic}}\ \frac{1}{a} \quad \text{(Case~I)} \end{aligned} \tag{21.16.4}$$

In the second case, "Case II," we assume $\sigma_{oc} = \sigma_{ic}$. In this case, we have

$$\begin{aligned} R' &\approx \frac{1}{2\pi\sqrt{2/\omega\mu_0}}\left[\frac{1}{a\sqrt{\sigma_{ic}}} + \frac{1}{b\sqrt{\sigma_{ic}}}\right] \\ &= \frac{1}{2\pi\delta_{ic}\sigma_{ic}}\ \left[\frac{1}{a} + \frac{1}{b}\right] \quad \text{(Case~II)} \end{aligned} \tag{21.16.5}$$

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:

$$\gamma \triangleq \sqrt{(R' + j\omega L')(G' + j\omega C')} = \alpha + j\beta \tag{21.16.6}$$

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

$$\alpha_R \approx K_R \frac{R'}{Z_0} \tag{21.16.7}$$

where Z_0 is the characteristic impedance

$$Z_0 \approx \frac{\eta_0}{2\pi} \frac{1}{\sqrt{\epsilon_r}} \ln \frac{b}{a} \quad \text{(low loss)} \tag{21.16.8}$$

and where K_R is a unitless constant to be determined. The justification for Equation 21.16.7is 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

$$\alpha_G \approx K_G G' Z_0 \tag{21.16.9}$$

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 $a \cong 0.292$ mm, $b \cong 1.855$ mm, $\sigma_{ic} \cong 2.28 \times 10^7$ S/m, $\sigma_s \cong 5.9 \times 10^{-5}$ S/m, and $\epsilon_r \cong 2.25$. 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 $C = 0$.

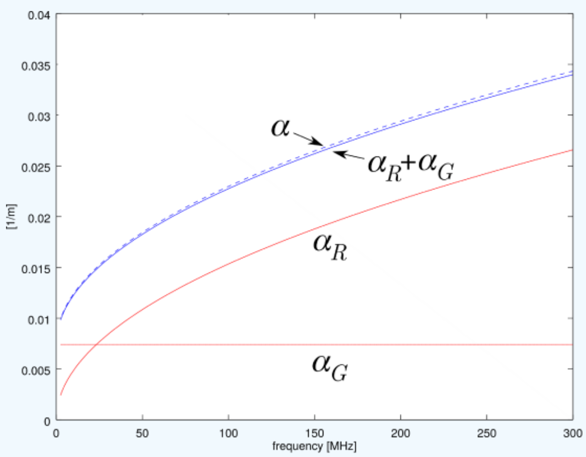


Figure 21.16.2 Comparison of $\alpha = \text{Re}\{\gamma\}$ to α_R, α_G, and $\alpha_R + \alpha_G$ for $K_R = K_G = 1/2$. The result for α has been multiplied by 1.01; otherwise the curves would be too close to tell apart.

Figure 21.16.2shows the components α_G and α_R computed for the particular choice $K_R = K_G = 1/2$. The figure also shows $\alpha_G + \alpha_R$, 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 $\alpha \approx \alpha_G + \alpha_R$ where $\alpha_G \triangleq R'/2Z_0$ and $\alpha_R \triangleq G' Z_0/2$.

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:

$$\begin{aligned} \alpha_G &\triangleq \frac{1}{2} G' Z_0 \\ &\approx \frac{1}{2} \cdot \frac{2\pi\sigma_s}{\ln(b/a)} \cdot \frac{1}{2\pi} \frac{\eta_0}{\sqrt{\epsilon_r}} \ln(b/a) \\ &= \frac{\eta_0}{2} \frac{\sigma_s}{\sqrt{\epsilon_r}} \end{aligned} \tag{21.16.10}$$

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:

$$\begin{aligned}\alpha_R &\triangleq \frac{R'}{2Z_0} \\ &= \frac{1}{2}\frac{(1/2\pi\delta_{ic}\sigma_{ic})\,[1/a+C/b]}{(1/2\pi)\left(\eta_0/\sqrt{\epsilon_r}\right)\ln(b/a)} \\ &= \frac{\sqrt{\epsilon_r}}{2\eta_0\delta_{ic}\sigma_{ic}}\cdot\frac{[1/a+C/b]}{\ln(b/a)}\end{aligned} \tag{21.16.11}$$

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 21.16.12can 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, $\epsilon_r \cong 2.25$ and $C = 0$, indicating $Z_0 \approx 39.9\ \Omega$ is optimum for attenuation. The actual characteristic impedance of Z_0 is about $75\ \Omega$, 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:

$$\begin{aligned}\frac{\partial}{\partial a}\left[a^2 \ln\left(\frac{b}{a}\right)\right] = &\left[\frac{\partial}{\partial a}a^2\right]\ln\left(\frac{b}{a}\right) \\ &+a^2\left[\frac{\partial}{\partial a}\ln\left(\frac{b}{a}\right)\right]\end{aligned} \tag{21.16.15}$$

Note

$$\frac{\partial}{\partial a}a^2 = 2a$$

and

$$\begin{aligned}\frac{\partial}{\partial a}\ln\left(\frac{b}{a}\right) &= \frac{\partial}{\partial a}[\ln(b) - \ln(a)] \\ &= -\frac{\partial}{\partial a}\ln(a) \\ &= -\frac{1}{a}\end{aligned} \tag{21.16.16}$$

So:

$$\begin{aligned}\frac{\partial}{\partial a}\left[a^2 \ln\left(\frac{b}{a}\right)\right] &= [2a]\ln\left(\frac{b}{a}\right) + a^2\left[-\frac{1}{a}\right] \\ &= \boxed{2a\ln\left(\frac{b}{a}\right) - a}\end{aligned} \tag{21.16.17}$$

This result is substituted for $a^2 \ln(b/a)$ in Equation 21.16.12 to obtain Equation 21.16.13.

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:

$$V(a) = V_0 \quad \rightarrow \quad A \ln a + B = V_0 \tag{21.17.1}$$

$$V(b) = 0 \quad \rightarrow \quad A \ln b + B = 0 \tag{21.17.2}$$

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

$$\mathbf{E} = -\hat{\rho}\frac{\partial}{\partial \rho} V \tag{21.17.3}$$

$$= -\hat{\rho}\frac{\partial}{\partial \rho}\left[\frac{-V_0}{\ln(b/a)} \ln \rho + \frac{V_0 \ln(b)}{\ln(b/a)}\right] \tag{21.17.4}$$

$$= +\hat{\rho}\frac{V_0}{\rho \ln(b/a)} \tag{21.17.5}$$

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

$$Z_0 \approx \frac{1}{2\pi}\frac{\eta_0}{\sqrt{\epsilon_r}} \ln\left(\frac{b}{a}\right) \tag{21.17.6}$$

Therefore, the maximum safe power is

$$P_{max} = \frac{V_0^2}{2Z_0} \tag{21.17.7}$$

$$\approx \frac{E_{pk}^2\, a^2 \ln^2(b/a)}{2\cdot(1/2\pi)\left(\eta_0/\sqrt{\epsilon_r}\right)\ln(b/a)} \tag{21.17.8}$$

$$= \frac{\pi E_{pk}^2}{\eta_0/\sqrt{\epsilon_r}}\, a^2 \ln(b/a) \tag{21.17.9}$$

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:

$$\frac{\partial}{\partial a} P_{max} = \frac{\pi E_{pk}^2}{\eta_0/\sqrt{\epsilon_r}}\ \left[2a\ln(b/a) - a\right] \tag{21.17.10}$$

For the above expression to be zero, it must be true that $2\ln(b/a) - 1 = 0$. Solving for b/a, we obtain:

$$\frac{b}{a} = \sqrt{e} \cong 1.65 \tag{21.17.11}$$

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 $30\ \Omega$. 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:

$$\frac{\partial}{\partial a}\left[\frac{1/a+C/b}{\ln(b/a)}\right] = \left[\frac{\partial}{\partial a}\left(\frac{1}{a}+\frac{C}{b}\right)\right]\ln^{-1}\left(\frac{b}{a}\right) + \left(\frac{1}{a}+\frac{C}{b}\right)\left[\frac{\partial}{\partial a}\ln^{-1}\left(\frac{b}{a}\right)\right] \tag{21.17.12}$$

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:

$$\begin{aligned}\frac{\partial}{\partial a}v^{-1} &= \left[\frac{\partial}{\partial v}v^{-1}\right]\left[\frac{\partial v}{\partial u}\right]\left[\frac{\partial u}{\partial a}\right] \\ &= \left[-v^{-2}\right]\left[\frac{1}{u}\right]\left[-ba^{-2}\right] \\ &= \left[-\ln^{-2}\left(\frac{b}{a}\right)\right]\left[\frac{a}{b}\right]\left[-ba^{-2}\right] \\ &= \frac{1}{a}\ln^{-2}\left(\frac{b}{a}\right)\end{aligned} \tag{21.17.13}$$

So:

$$\begin{aligned}\frac{\partial}{\partial a}\left[\frac{1/a+C/b}{\ln(b/a)}\right] = &\left[-\frac{1}{a^2}\right]\ln^{-1}\left(\frac{b}{a}\right) \\ &+\left(\frac{1}{a}+\frac{C}{b}\right)\left[\frac{1}{a}\ln^{-2}\left(\frac{b}{a}\right)\right]\end{aligned} \tag{21.17.14}$$

This result is substituted in Equation 21.17.9to obtain Equation 21.17.10.

21.18: Why 50 Ohms?

The quantity $50\ \Omega$ 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 $50\ \Omega$? The short answer is "nothing." In fact, other standard impedances are in common use – prominent among these is $75\ \Omega$. 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 $300\ \Omega$ 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 $50\ \Omega$ 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 $60\ \Omega$ to $77\ \Omega$ is optimum for air-filled cable, but more like $40\ \Omega$ to $50\ \Omega$ for cables using a plastic spacer material having typical $\epsilon_r \approx 2.25$. Thus, $50\ \Omega$ 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 $30\ \Omega$. This is significantly less than the $60\ \Omega$ to $77\ \Omega$ that minimizes attenuation in air-filled cables. So, $50\ \Omega$ can be viewed as a compromise between minimizing attenuation and maximizing power handling in air-filled coaxial cables.

Although the preceding arguments justify $50\ \Omega$ as a standard value, one can also see how one might make a case for $75\ \Omega$ as a secondary standard value, especially for applications where attenuation is the primary consideration.

Values of $50\ \Omega$ and $75\ \Omega$ also offer some convenience when connecting RF devices to antennas. For example, $75\ \Omega$ 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 $50\ \Omega$. In fact, we see that if we desire a single characteristic impedance that is equally convenient for applications involving either type of antenna, then $50\ \Omega$ 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 $300\ \Omega$ 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 $75\ \Omega$. Thus, we have an additional application for $75\ \Omega$ coaxial line.

Finally, note that it is quite simple to implement microstrip transmission line having characteristic impedance in the range $30\ \Omega$ to $75\ \Omega$. For example, $50\ \Omega$ 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.

21.19: Conclusion

CHAPTER OVERVIEW

22: Generation and Detection of Electromagnetic Waves

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.

22.2: 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 $22.2.1$.

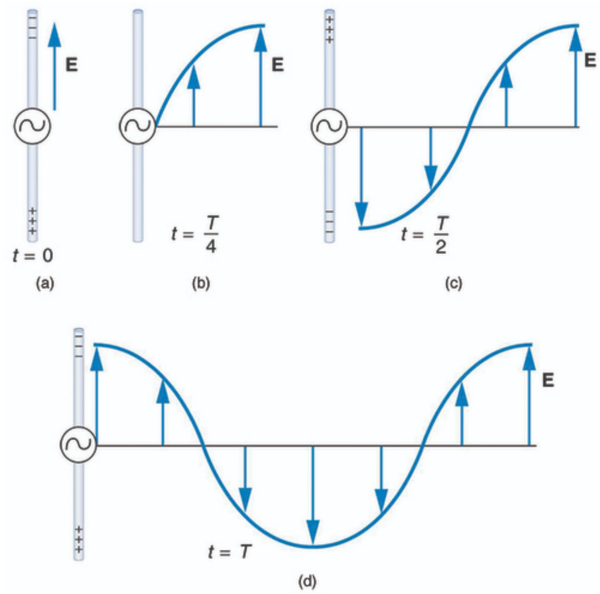


Figure $22.2.1$: This long straight gray wire with an AC generator at its center becomes a broadcast antenna for electromagnetic waves. Shown here are the charge distributions at four different times. The electric field ($textbf{E}$) propagates away from the antenna at the speed of light, forming part 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.

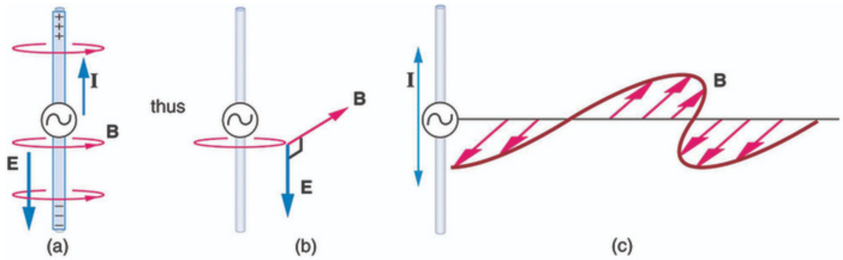


Figure $22.2.2$: (a) The current in the antenna produces the circular magnetic field lines. The current (I) produces the separation of charge along the wire, which in turn creates the electric field as shown. (b) The electric and magnetic fields (E and B) near the wire are perpendicular; they are shown 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**.

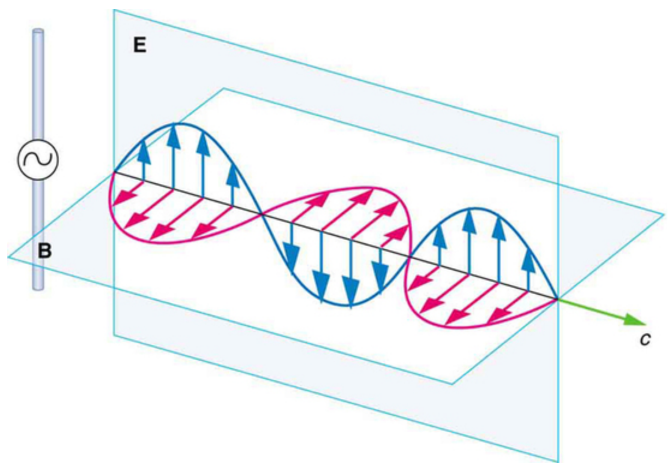


Figure $22.2.3$: A part of the electromagnetic wave sent out from the antenna at one instant in time. The electric and magnetic fields ($textbf{E}$ and $textbf{B}$) are in phase, and they are perpendicular to one another and the direction of propagation. For clarity, the waves are shown only along one 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 $\boldsymbol{E}$-Field and $\boldsymbol{B}$-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,

$$\frac{E}{B}=c \tag{22.2.1}$$

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 $1000 V/m$?

Strategy:

To find the B-field strength, we rearrange the Equation 22.2.1 to solve for B, yielding

$$B=\frac{E}{c}. \tag{22.2.2}$$

Solution:

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

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

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.

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

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

which implies that the magnetic field B is very weak relative to the electric field E.

Glossary

electric field

a vector quantity (**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 (**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

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.

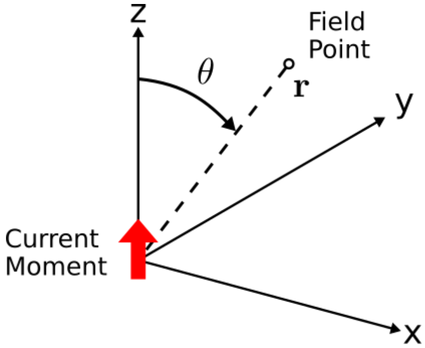


Figure $22.3.1$: A $+\hat{\mathbf{z}}$-directed current moment located at the origin. (CC BY-SA 4.0; C. Wang)

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:

$$\Delta\mathbf{J}(\mathbf{r}) = \hat{\mathbf{z}}\, I\, \Delta l\, \delta(\mathbf{r}) \tag{22.3.1}$$

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 function[1] defined as follows:

$$\delta(\mathbf{r}) \triangleq 0 \quad \text{for} \quad \mathbf{r} \neq 0; \quad \text{and} \tag{22.3.2}$$

$$\int_{\mathcal{V}} \delta(\mathbf{r})\, dv \triangleq 1 \tag{22.3.3}$$

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 $\mathrm{A/m}^2$, 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

$$\Delta\widetilde{\mathbf{J}}(\mathbf{r}) = \hat{\mathbf{z}}\, \tilde{I}\ \Delta l\, \delta(\mathbf{r}) \tag{22.3.4}$$

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 $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ due to $\Delta\widetilde{\mathbf{J}}(\mathbf{r})$?" Without doing any math, we know quite a bit about $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$. For example:

- Since electric fields are proportional to the currents that give rise to them, we expect $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ to be proportional to $\left|\tilde{I}\ \Delta l\right|$.

- If we are sufficiently far from the origin, we expect $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ to be approximately proportional to $1/r$ where $r \triangleq |\mathbf{r}|$ 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 $\left|\Delta\widetilde{\mathbf{E}}(\mathbf{r})\right|^2$, $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ 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 $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ to change approximately at rate β where β is the phase propagation constant $2\pi/\lambda$. Since we expect spherical phasefronts, $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ should therefore contain the factor $e^{-j\beta r}$.
- Ampere's law indicates that a $\hat{\mathbf{z}}$-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., $+\hat{\mathbf{r}}$. Therefore, $\Delta\widetilde{\mathbf{E}}(z=0)$ points in the $-\hat{\mathbf{z}}$ direction. The same principle applies outside of the $z=0$ plane, so in general we expect $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ to point in the $\hat{\theta}$ direction.
- We expect $\Delta\widetilde{\mathbf{E}}(\mathbf{r})=0$ along the z axis. Subsequently $\left|\Delta\widetilde{\mathbf{E}}(\hat{\mathbf{r}})\right|$ must increase from zero at $\theta=0$ and return to zero at $\theta=\pi$. The symmetry of the problem suggests $\left|\Delta\widetilde{\mathbf{E}}(\hat{\mathbf{r}})\right|$ is maximum at $\theta=\pi/2$. This magnitude must vary in the simplest possible way, leading us to conclude that $\Delta\widetilde{\mathbf{E}}(\hat{\mathbf{r}})$ is proportional to $\sin\theta$. Furthermore, the radial symmetry of the problem means that $\Delta\widetilde{\mathbf{E}}(\hat{\mathbf{r}})$ should not depend at all on ϕ.

Putting these ideas together, we conclude that the radiated electric field has the following form:

$$\Delta\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} C\left(\tilde{I}\ \Delta l\right)(\sin\theta)\frac{e^{-j\beta r}}{r}$$

where C is a constant which accounts for all of the constants of proportionality identified in the preceding analysis. Since the units of $\Delta\widetilde{\mathbf{E}}(\mathbf{r})$ are V/m, the units of C must be Ω/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:

$$\Delta\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta}\frac{j\eta\beta}{4\pi}\left(\tilde{I}\ \Delta l\right)(\sin\theta)\frac{e^{-j\beta r}}{r}$$

Additional evidence that this solution is correct comes from the fact that it satisfies the wave equation $\nabla^2\Delta\widetilde{\mathbf{E}}(\mathbf{r})+\beta^2\Delta\widetilde{\mathbf{E}}(\mathbf{r})=0$.[3]

Note that the expression we have obtained for the radiated electric field is approximate (hence the "$\approx$"). 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.↩

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 $22.4.1$.

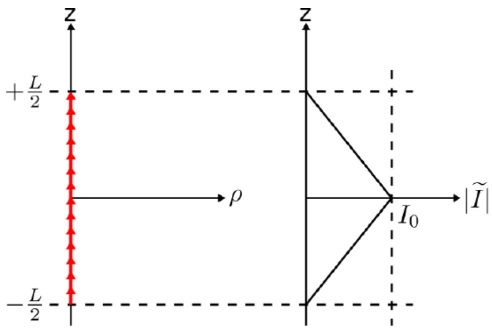


Figure $22.4.1$: Current distribution of the electrically-short dipole (ESD). (CC BY-SA 4.0; C. Wang)

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 $\widetilde{\mathbf{A}}$ by integration over the current distribution, calculate $\widetilde{\mathbf{H}} = (1/\mu)\nabla \times \widetilde{\mathbf{A}}$, and finally calculate $\widetilde{\mathbf{E}}$ from $\widetilde{\mathbf{H}}$ using Ampere's law. We shall employ a simpler approach, shown in Figure $22.4.2$.

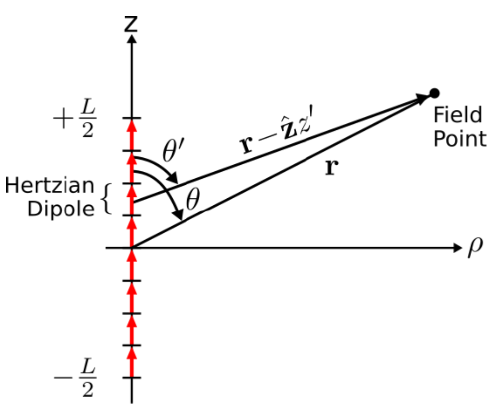


Figure $22.4.2$ Current distribution of the electricallyshort dipole (ESD) approximated as a large number of Hertzian dipoles. (CC BY-SA 4.0; C. Wang)

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

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\eta \frac{\tilde{I} \cdot \beta \Delta l}{4\pi} \; (\sin\theta) \; \frac{e^{-j\beta r}}{r}$$

where $\tilde{I}$ 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:

$$\widetilde{\mathbf{E}}(\mathbf{r}; z' = 0) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta \Delta l}{4\pi} \; (\sin\theta) \; \frac{e^{-j\beta r}}{r}$$

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:

$$d\widetilde{\mathbf{E}}(\mathbf{r}; z' = 0) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta dz'}{4\pi} \; (\sin\theta) \; \frac{e^{-j\beta r}}{r}$$

Using this approach, the electric field radiated by *any* segment can be written:

$$d\widetilde{\mathbf{E}}(\mathbf{r}; z') \approx \hat{\theta}' j\eta\beta \frac{\tilde{I}(z')}{4\pi} (\sin\theta') \; \frac{e^{-j\beta|\mathbf{r}-\hat{\mathbf{z}}z'|}}{|\mathbf{r}-\hat{\mathbf{z}}z'|} \; dz'$$

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:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \int_{-L/2}^{+L/2} d\widetilde{\mathbf{E}}(\hat{\mathbf{r}}; z')$$

yielding:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx j\frac{\eta\beta}{4\pi} \int_{-L/2}^{+L/2} \hat{\theta}' \tilde{I}(z') \; (\sin\theta') \; \frac{e^{-j\beta|\mathbf{r}-\hat{\mathbf{z}}z'|}}{|\mathbf{r}-\hat{\mathbf{z}}z'|} dz'$$

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:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\frac{\eta\beta}{4\pi} \; (\sin\theta) \int_{-L/2}^{+L/2} \tilde{I}(z') \frac{e^{-j\beta|\mathbf{r}-\hat{\mathbf{z}}z'|}}{|\mathbf{r}-\hat{\mathbf{z}}z'|} dz' \tag{22.4.1}$$

It is also possible to simplify the expression $|\mathbf{r}-\hat{\mathbf{z}}z'|$. Consider Figure 22.4.3

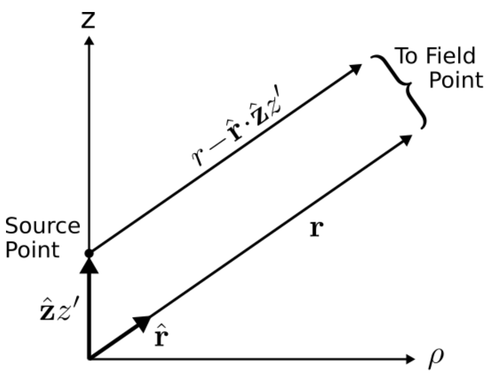


Figure 22.4.3 Parallel ray approximation for an ESD. (CC BY-SA 4.0; C. Wang)

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

$$|\mathbf{r}-\hat{\mathbf{z}}z'| \approx r - \hat{\mathbf{r}} \cdot \hat{\mathbf{z}} z' \tag{22.4.2}$$

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 $\widetilde{\mathbf{E}}(\mathbf{r})$, 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:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta\beta}{4\pi} (\sin\theta) \frac{e^{-j\beta r}}{r} \int_{-L/2}^{+L/2} \tilde{I}(z') dz'$$

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:

$$\boxed{\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta L}{8\pi} (\sin\theta) \frac{e^{-j\beta r}}{r}} \tag{22.4.3}$$

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 $\hat{\mathbf{z}}I_0 L/2$ (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:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta I_0}{4} \frac{L}{\lambda} (\sin\theta) \frac{e^{-j\beta r}}{r}$$

At field points $r \gg \lambda$, the wave appears to be locally planar. Therefore, we are justified using the plane wave relationship $\widetilde{\mathbf{H}} = \frac{1}{\eta}\hat{\mathbf{r}} \times \widetilde{\mathbf{E}}$ to calculate $\widetilde{\mathbf{H}}$. The result is:

$$\widetilde{\mathbf{H}}(\mathbf{r}) \approx \hat{\phi} j \frac{I_0}{4} \frac{L}{\lambda} (\sin\theta) \frac{e^{-j\beta r}}{r} \tag{22.4.4}$$

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 $\widetilde{\mathbf{E}} \times \widetilde{\mathbf{H}}$ 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.

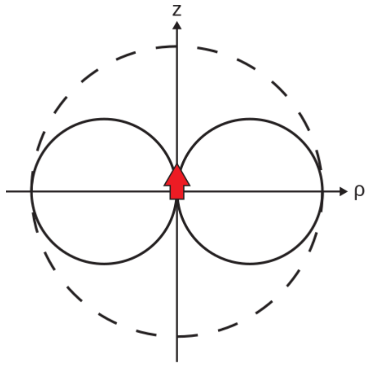


Figure 22.4.4 Magnitude of the radiated field in any plane of constant ϕ. (CC BY-SA 4.0; S. Lally)

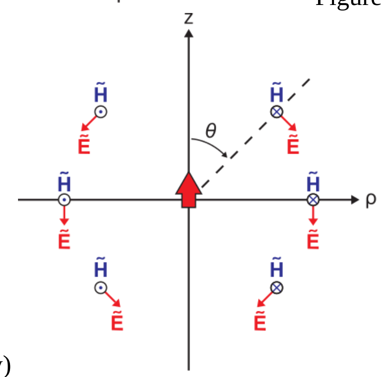


Figure 22.4.5 Orientation of the electric and magnetic fields in any plane of constant ϕ. (CC BY-SA 4.0; S. Lally)

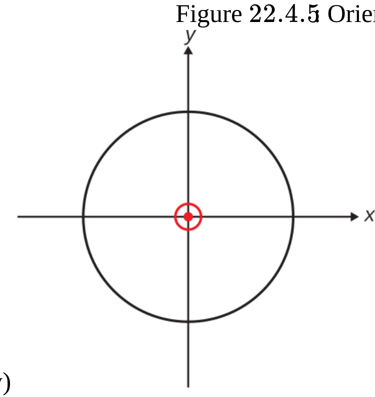


Figure 22.4.6 Magnitude of the radiated field in any plane of constant z. (CC BY-SA 4.0; S. Lally)

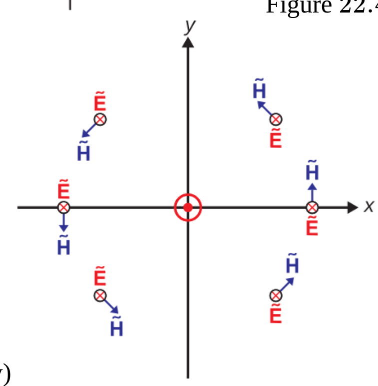


Figure 22.4.7 Orientation of the electric and magnetic fields in the $z=0$ plane. (CC BY-SA 4.0; S. Lally)

---

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.↩

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.

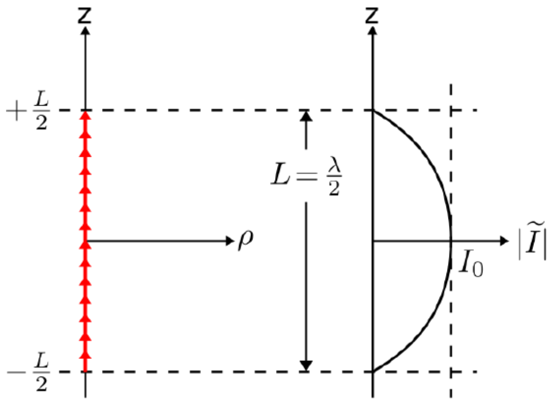


Figure 22.5.1: Current distribution of the half-wave dipole (HWD). (; C. Wang)

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:

$$\tilde{I}(z) \approx I_0 \cos\left(\pi \frac{z}{L}\right) \text{ for } |z| \leq \frac{L}{2} \tag{22.5.1}$$

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:

$$\tilde{I}(z) \approx I_0 \cos\left(2\pi \frac{z}{\lambda}\right) \tag{22.5.2}$$

The electromagnetic field radiated by this distribution of current may be calculated using the method described in Section 9.6, in particular:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta}{2} \frac{e^{-j\beta r}}{r}(\sin\theta) \cdot \left[\frac{1}{\lambda}\int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z' \cos\theta} dz'\right] \tag{22.5.3}$$

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

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j \frac{\eta I_0}{2\pi} \frac{\cos[(\pi/2)\cos\theta]}{\sin\theta} \frac{e^{-j\beta r}}{r}$$

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:

$$\widetilde{\mathbf{H}} = \frac{1}{\eta}\hat{\mathbf{r}} \times \widetilde{\mathbf{E}}$$

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 $22.5.2$

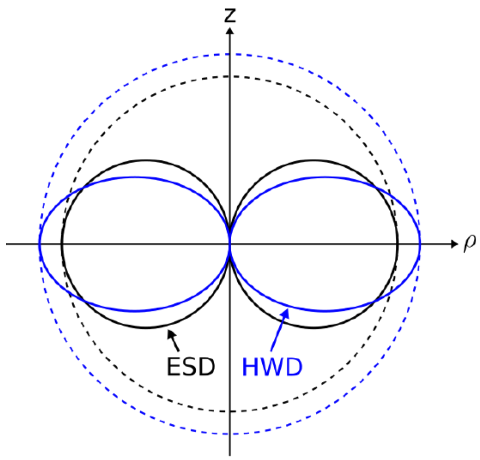


Figure $22.5.2$ Comparison of the magnitude of the radiated field of the HWD to that of an electrically-short dipole also oriented along the z-axis. This result is for any radial plane that includes the z-axis. (CC BY-SA 4.0; C. Wang)

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.

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 $22.6.1$.

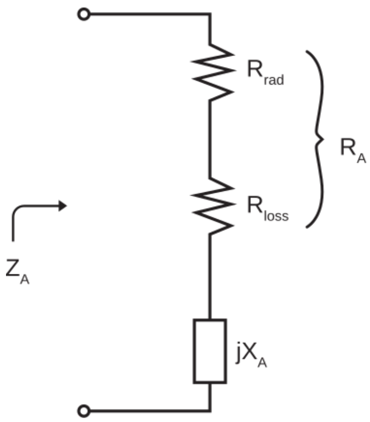


Figure $22.6.1$: Equivalent circuit for an antenna which is transmitting. (CC BY-SA 4.0; S. Lally)

Since the antenna is passive, it is reasonable to describe it as an impedance Z_A which is (by definition) the ratio of voltage $\widetilde{V}_A$ to current $\tilde{I}_A$ at the terminals; i.e.,

$$Z_A \triangleq \frac{\widetilde{V}_A}{\tilde{I}_A}$$

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

$$P_A = \frac{1}{2}\text{Re}\left\{\widetilde{V}_A \tilde{I}_A^*\right\}$$

where we have assumed peak (as opposed to root mean squared) units for voltage and current. Since $\widetilde{V}_A = Z_A \tilde{I}_A$, we have:

$$P_A = \frac{1}{2}\text{Re}\left\{(R_{rad} + R_{loss} + jX_A)\,\tilde{I}_A \tilde{I}_A^*\right\}$$

which reduces to:

$$P_A = \frac{1}{2}\left|\tilde{I}_A\right|^2 R_{rad} + \frac{1}{2}\left|\tilde{I}_A\right|^2 R_{loss} \tag{22.6.1}$$

As expected, the power transferred to the antenna is the sum of

$$P_{rad} \triangleq \frac{1}{2}\left|\tilde{I}_A\right|^2 R_{rad} \tag{22.6.2}$$

representing power transferred to the radiating electromagnetic field, and

$$P_{loss} \triangleq \frac{1}{2}\left|\tilde{I}_A\right|^2 R_{loss} \tag{22.6.3}$$

representing power dissipated within the antenna.

The reactance X_A will play a role in determining $\tilde{I}_A$ given $\widetilde{V}_A$ (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 22.6.2tells us that

$$R_{rad} = 2P_{rad}\left|\tilde{I}_A\right|^{-2} \tag{22.6.4}$$

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 22.6.4to 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:

$$e_{rad} = \frac{R_{rad}}{R_{rad} + R_{loss}} \tag{22.6.5}$$

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, $P_{rad} = 60$ mW, $\left|\tilde{I}_A\right| = 20$ 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

$$R_{rad} \approx \frac{2 \cdot (60 \text{ mW})}{\left|\sqrt{2} \cdot 20 \text{ mA}\right|^2} = \underline{150\ \Omega}$$

Solving Equation 22.6.5 for the loss resistance, we find:

$$R_{loss} = \frac{1 - e_{rad}}{e_{rad}} R_{rad} \cong \underline{64.3\ \Omega}$$

Since $Z_A = R_{rad} + R_{loss} + jX_A$, we find $Z_A \cong \underline{214.3 + j0\ \Omega}$. 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.↩

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 $22.7.1$.

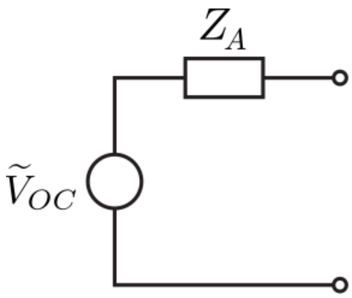


Figure $22.7.1$: Thévenin equivalent circuit for an antenna in the presence of an incident electromagnetic wave. (CC BY-SA 4.0 (modified); S. Lally)

The circuit consists of a voltage source $\widetilde{V}_{OC}$ and a series impedance Z_A. The source potential $\widetilde{V}_{OC}$ 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 $\widetilde{V}_{OC}$ 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 $\widetilde{V}_{OC}$ 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 $\widetilde{V}_{OC}$:

- $\widetilde{V}_{OC}$ must depend on the incident electric field intensity $\widetilde{\mathbf{E}}^i$. Presumably the relationship is linear, so $\widetilde{V}_{OC}$ is proportional to the magnitude of $\widetilde{\mathbf{E}}^i$.
- Since $\widetilde{\mathbf{E}}^i$ is a vector whereas $\widetilde{V}_{OC}$ is a scalar, there must be some vector $\mathbf{l}_e$ for which

$$\widetilde{V}_{OC} = \widetilde{\mathbf{E}}^i \cdot \mathbf{l}_e \tag{22.7.1}$$

- Since $\widetilde{\mathbf{E}}^i$ has SI base units of V/m and $\widetilde{V}_{OC}$ has SI base units of V, $\mathbf{l}_e$ must have SI base units of m; i.e., length.
- We expect that $\widetilde{V}_{OC}$ 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 $\widetilde{\mathbf{E}}^i$ to the observed scalar $\widetilde{V}_{OC}$ 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, $\widetilde{\mathbf{E}}^i \cdot \hat{\mathbf{l}}$ yields the scalar component of $\widetilde{\mathbf{E}}^i$ 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 $\widetilde{V}_{OC}$. Summarizing:

The *vector effective length* $\mathbf{l}_e = \hat{\mathbf{l}} l_e$ 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* l_e is

$$l_e \triangleq \frac{\widetilde{V}_{OC}}{\widetilde{\mathbf{E}}^i \cdot \hat{\mathbf{l}}}$$

where $\widetilde{V}_{OC}$ is the open-circuit potential induced at the antenna terminals in response to the incident electric field intensity $\widetilde{\mathbf{E}}^i$.

While this definition yields an unambiguous value for 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 $\widetilde{\mathbf{E}}^i$ be a $\hat{\theta}$-polarized plane wave arriving at the ESD. The ESD is open-circuited, so the potential induced in its terminals is $\widetilde{V}_{OC}$. One observes the following:

- When $\widetilde{\mathbf{E}}^i$ arrives from anywhere in the $\theta = \pi/2$ plane (i.e., broadside to the ESD), $\widetilde{\mathbf{E}}^i$ points in the $-\hat{\mathbf{z}}$ direction, and we find that $l_e \approx L/2$. It should not be surprising that l_e is proportional to L; this expectation was noted earlier in this section.
- When $\widetilde{\mathbf{E}}^i$ arrives from the directions $\theta = 0$ or $\theta = \pi$ – i.e., along the axis of the ESD – $\widetilde{\mathbf{E}}^i$ is perpendicular to the axis of the ESD. In this case, we find that l_e equals zero.

Taken together, these findings suggest that 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

$$\mathbf{l}_e \approx \hat{\theta}\frac{L}{2}\sin\theta \quad \text{(ESD)} \tag{22.7.2}$$

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 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 μV/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 $l_e = 3.54$ cm. When the electric field is $\hat{\theta}$-polarized, the magnitude of the induced open-circuit voltage is

$$\begin{aligned}\left|\widetilde{V}_{OC}\right| &= \left|\widetilde{\mathbf{E}}^i \cdot \mathbf{l}_e\right|\\ &\approx (10\ \mu\text{V/m})\,\hat{\theta}\cdot\hat{\theta}\,(3.54 \text{ cm})\\ &\approx \underline{354 \text{ nV rms}} \quad \text{(a)}\end{aligned}$$

When the electric field is $\hat{\phi}$-polarized:

$$\begin{aligned}\left|\widetilde{V}_{OC}\right| &\approx (10\ \mu\text{V/m})\,\hat{\phi}\cdot\hat{\theta}\,(3.54 \text{ cm})\\ &\approx \underline{0} \quad \text{(b)}\end{aligned}$$

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 $\widetilde{V}_L$ across the load is

$$\widetilde{V}_L = \widetilde{V}_{OC} \frac{Z_A^*}{Z_A + Z_A^*} = \widetilde{V}_{OC} \frac{Z_A^*}{2R_A}$$

The current $\tilde{I}_L$ through the load is

$$\tilde{I}_L = \frac{\widetilde{V}_{OC}}{Z_A + Z_A^*} = \frac{\widetilde{V}_{OC}}{2R_A}$$

Taking $\widetilde{V}_{OC}$ as an RMS quantity, the power P_L delivered to the load is

$$P_L = \text{Re}\left\{V_L I_L^*\right\} = \frac{\left|\widetilde{V}_{OC}\right|^2}{4R_A}$$

In part (a) of Example 22.7.1, $\left|\widetilde{V}_{OC}\right|$ is found to be ≈ 354 nV rms, so $P_L \approx \underline{6.33\ \text{pW}}$.

---

1. Formal derivations of these quantities are provided in subsequent sections. The starting point is the section on reciprocity.↩

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 $\hat{\mathbf{z}}$-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 $\widetilde{\mathbf{E}}^i$. The question is: What is $\widetilde{V}_{OC}$, the potential at the terminals when the terminals are open-circuited?

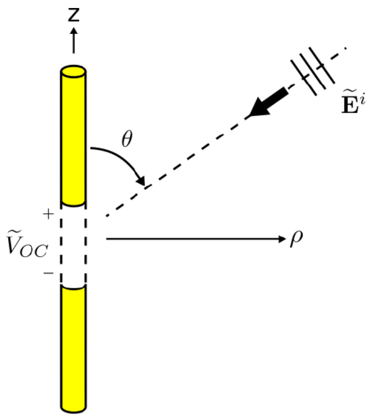


Figure $22.8.1$: A potential is induced at the terminals of a thin straight dipole in response to an incident plane wave. (CC BY-SA 4.0; C. Wang)

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:

$$\widetilde{V}_{OC} = -\int_{gap} \widetilde{\mathbf{E}}_{gap} \cdot d\mathbf{l} \tag{22.8.1}$$

where $\widetilde{\mathbf{E}}_{gap}$ is the electric field in the gap. The problem with this approach is that the value of $\widetilde{\mathbf{E}}_{gap}$ is not readily available. It is not simply $\widetilde{\mathbf{E}}^i$, 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 $22.8.2$.

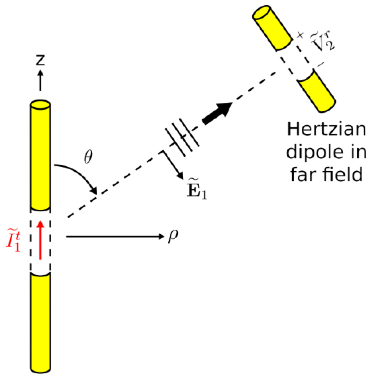


Figure 22.8.2 The dipole of interest driven by current $\tilde{I}_1^t$ radiates electric field $\widetilde{\mathbf{E}}_1$, resulting in open-circuit potential $\widetilde{V}_2^r$ at the terminals of a Hertzian dipole in the far field. (CC BY-SA 4.0; C. Wang)

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):

$$\begin{aligned}\widetilde{\mathbf{E}}_1(\mathbf{r}) \approx\; & \hat{\theta} j \frac{\eta}{2} \frac{e^{-j\beta r}}{r}(\sin\theta) \\ & \cdot \left[\frac{1}{\lambda}\int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z'\cos\theta} dz'\right]\end{aligned} \tag{22.8.2}$$

The second antenna is a $\hat{\theta}$-aligned Hertzian dipole in the far field, which receives $\widetilde{\mathbf{E}}_1$. (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 $\widetilde{V}_2^r$ 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:

$$\widetilde{V}_2^r = -\int_{gap} \widetilde{\mathbf{E}}_{gap} \cdot d\mathbf{l}$$

For the Hertzian dipole, $\widetilde{\mathbf{E}}_{gap}$ *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:

$$\widetilde{V}_2^r = -\int_{gap} \widetilde{\mathbf{E}}_1 \cdot d\mathbf{l} \tag{22.8.3}$$

Since the Hertzian dipole is very short and very far away from the transmitting dipole, $\widetilde{\mathbf{E}}_1$ is essentially constant over the gap. Also recall that we required the Hertzian dipole to be aligned with $\widetilde{\mathbf{E}}_1$. Choosing to integrate in a straight line across the gap, Equation 22.8.3 reduces to:

$$\widetilde{V}_2^r = -\widetilde{\mathbf{E}}_1(\mathbf{r}_2) \cdot \hat{\theta}\Delta l \tag{22.8.4}$$

where Δl is the length of the gap and $\mathbf{r}_2$ is the location of the Hertzian dipole. Substituting the expression for $\widetilde{\mathbf{E}}_1$ from Equation 22.8.2, we obtain:

$$\begin{aligned}\widetilde{V}_2^r \approx\; & -j\frac{\eta}{2}\frac{e^{-j\beta r_2}}{r_2}(\sin\theta) \\ & \cdot \left[\frac{1}{\lambda}\int_{-L/2}^{+L/2} \tilde{I}(z') e^{+j\beta z'\cos\theta} dz'\right]\Delta l\end{aligned} \tag{22.8.5}$$

where $r_2 = |\mathbf{r}_2|$.

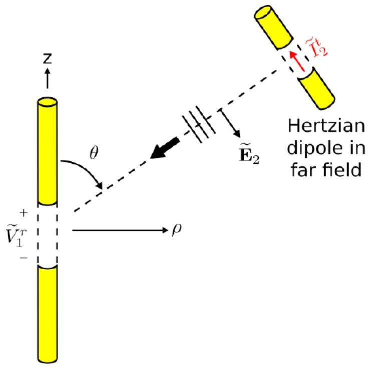


Figure 22.8.3 The Hertzian dipole driven by current Iet $\tilde{I}_2^t$ radiates electric field $\widetilde{\mathbf{E}}_2$, resulting in open-circuit potential $\widetilde{V}_1^r$ at the terminals of the dipole of interest. (CC BY-SA 4.0; C. Wang)

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):

$$\widetilde{\mathbf{E}}_2(\mathbf{r}=0) \approx \hat{\theta}\, j\eta \frac{\tilde{I}_2^t \cdot \beta \Delta l}{4\pi} (1) \frac{e^{-j\beta r_2}}{r_2} \tag{22.8.6}$$

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 $\widetilde{V}_1^r$ 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 $\widetilde{V}_1^r$ should be proportional to $\widetilde{\mathbf{E}}_2(\mathbf{r}=0)$, 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

$$\widetilde{V}_1^r = \widetilde{\mathbf{E}}_2(\mathbf{r}=0) \cdot \mathbf{l}_e \tag{22.8.7}$$

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:

$$\widetilde{V}_1^r = -\widetilde{\mathbf{E}}_2(\mathbf{r}=0) \cdot \hat{\theta}\, l_e$$

We may go a bit further and substitute the expression for $\widetilde{\mathbf{E}}_2(\mathbf{r}=0)$ from Equation 22.8.6:

$$\widetilde{V}_1^r \approx -j\eta \frac{\tilde{I}_2^t \cdot \beta \Delta l}{4\pi} \frac{e^{-j\beta r_2}}{r_2} l_e \tag{22.8.8}$$

Now we invoke reciprocity. As a two-port linear time-invariant system, it must be true that:

$$\tilde{I}_1^t \widetilde{V}_1^r = \tilde{I}_2^t \widetilde{V}_2^r$$

Thus:

$$\widetilde{V}_1^r = \frac{\tilde{I}_2^t}{\tilde{I}_1^t} \widetilde{V}_2^r$$

Substituting the expression for $\widetilde{V}_2^r$ from Equation 22.8.5:

$$\widetilde{V}_1^r \approx -\frac{\tilde{I}_2^t}{\tilde{I}_1^t}\cdot j\frac{\eta}{2}\frac{e^{-j\beta r_2}}{r_2}\ (\sin\theta) \\ \cdot\left[\frac{1}{\lambda}\int_{-L/2}^{+L/2}\tilde{I}\,(z')e^{+j\beta z'\cos\theta}dz'\right]\Delta l$$

Thus, reciprocity has provided a second expression for $\widetilde{V}_1^r$. We may solve for 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:

$$l_e \approx \left[\frac{1}{\tilde{I}_1^t}\int_{-L/2}^{+L/2}\tilde{I}\,(z')e^{+j\beta z'\cos\theta}dz'\right]\sin\theta \tag{22.8.9}$$

Thus, you can calculate 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 $\widetilde{V}_{OC}$ in response to an incident electric field $\widetilde{\mathbf{E}}^i$ is

$$\boxed{\widetilde{V}_{OC} = \widetilde{\mathbf{E}}^i\cdot\mathbf{l}_e} \tag{22.8.10}$$

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)

A 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:

$$l_e \approx \left[\frac{1}{I_0}\int_{-L/2}^{+L/2} I_0\left(1-\frac{2}{L}|z'|\right)e^{+j\beta z'\cos\theta}dz'\right] \\ \cdot\sin\theta$$

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!↩

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

$$\boxed{G \triangleq 10 \log_{10} \frac{P_1}{P_0} \quad \text{(dB)}} \tag{22.9.1}$$

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:

G &= 10\log_{10}\left(5 \times 10^{-6}\right) \cong -53.0~\mbox{dB} \nonumber \\ L &= -G \cong \underline{+53.0~\mbox{dB}} \nonumber

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:

$$\begin{aligned} 10 \log_{10} P_1 &= 10 \log_{10} \frac{P_0}{L} \\ &= 10 \log_{10} P_0 - 10 \log_{10} L \end{aligned} \tag{22.9.4}$$

Division has been transformed into subtraction; i.e.,

$$P_1 = P_0 - L \quad \text{(dB)} \tag{22.9.5}$$

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 22.9.5 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.

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:

$$D(\theta, \phi) \triangleq \frac{S(\mathbf{r})}{S_{ave}(r)} \tag{22.10.1}$$

In this expression, $S(\mathbf{r})$ is the power density at (r, θ, ϕ); i.e., at a distance r in the direction (θ, ϕ). $S_{ave}(r)$ is the *average* power density at that distance; that is, $S(\mathbf{r})$ averaged over all possible directions at distance 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 r. Even though the numerator and denominator of Equation 22.10.1 both vary with r, one finds that the distance dependence always cancels because power density and average power density are both proportional to 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:

$$\widetilde{\mathbf{E}}(\mathbf{r}) \approx \hat{\theta} j\eta \frac{I_0 \cdot \beta L}{8\pi} (\sin\theta) \frac{e^{-j\beta r}}{r} \tag{22.10.2}$$

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:

$$S(\mathbf{r}) \approx \eta \frac{|I_0|^2 (\beta L)^2}{128\pi^2} (\sin\theta)^2 \frac{1}{r^2} \tag{22.10.3}$$

and we subsequently find that the total power radiated is:

$$P_{rad} \approx \eta \frac{|I_0|^2 (\beta L)^2}{48\pi} \tag{22.10.4}$$

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 22.10.2and 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:

$$\begin{aligned} D(\theta,\phi) &\triangleq \frac{S(\mathbf{r})}{S_{ave}(r)} \\ &\approx 1.5(\sin\theta)^2 \end{aligned}$$

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.

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 $22.11.1$:

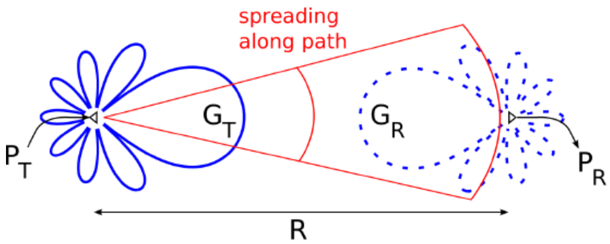


Figure $22.11.1$: Copy and Paste Caption here

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 22.11.1may be written in the following form:

$$\boxed{P_{R,max} = P_T G_T \left(\frac{\lambda}{4\pi R}\right)^2 G_R} \tag{22.11.2}$$

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:

$$\boxed{L_p \triangleq \left(\frac{\lambda}{4\pi R}\right)^{-2}}$$

which is known as *free space path loss*. Thus, Equation 22.11.2may be expressed as follows:

$$P_{R,max} = P_T G_T L_p^{-1} G_R \tag{22.11.3}$$

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 22.11.2applies. We find:

$$\begin{aligned} P_T &\geq \frac{P_{R,max}}{G_T(\lambda/4\pi R)^2 G_R} \\ &\cong 2.26 \times 10^{-7} \text{ W} \\ &\cong 2.26 \times 10^{-4} \text{ mW} \\ &\cong \underline{-36.5 \text{ dBm}} \end{aligned}$$

Additional Reading:

- “Free-space path loss” on Wikipedia.
- “Friis transmission equation” on Wikipedia.

CHAPTER OVERVIEW

23: Signal Modulation

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-

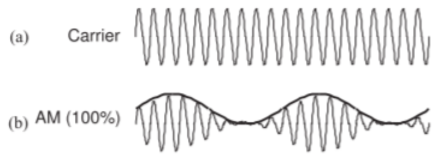


Figure $23.1.1$: AM showing the relationship between the carrier and modulation envelope: (a) carrier; and (b) 100% amplitude modulated carrier.

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 where 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.

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

Band	EIA Waveguide Band	Operating Frequency (GHz)	Internal Dimensions ($a \times b$, inches)
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**, E, with units of volts per meter (V/m), and the magnetic field, H, with units of amperes per meter (A/m). E and 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 D, the **electric flux** density, with units of coulombs per square meter $(\mathrm{C}/\mathrm{m}^2)$, and the other is B, the **magnetic flux** density, with units of teslas (T). B and H, and D and E, are related to each other by the properties of the medium, which are embodied in the quantities μ and ε (with the caligraphic letter, e.g. $\mathcal{B}$, denoting a time-domain quantity):

$$\overline{\mathcal{B}} = \mu\overline{\mathcal{H}} \tag{23.2.1}$$

$$\overline{\mathcal{D}} = \varepsilon\overline{\mathcal{E}} \tag{23.2.2}$$

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 $\mu_0 = 4\pi \times 10^{-7}\ \mathrm{H}/\mathrm{m}$ and the magnetic flux and magnetic field are related as

$$\overline{\mathcal{B}} = \mu_0\overline{\mathcal{H}} \tag{23.2.3}$$

The other material quantity is the **permittivity**, ε, which describes the ability to store energy in a volume and in a vacuum

$$\overline{\mathcal{D}} = \varepsilon_0\overline{\mathcal{E}} \tag{23.2.4}$$

where $\varepsilon_0 = 8.854 \times 10^{-12}\ \mathrm{F}/\mathrm{m}$ is the permittivity of a vacuum. The **relative permittivity**, ε_r, is the ratio the permittivity of a material to that of vacuum:

$$\varepsilon_r = \varepsilon/\varepsilon_0 \tag{23.2.5}$$

Similarly, the **relative permeability**, μ_r, refers to the ratio of permeability of a material to its value in a vacuum:

$$\mu_r = \mu/\mu_0 \tag{23.2.6}$$

1.3.2 Biot-Savart Law

The Biot–Savart law relates current to magnetic field as, see Figure $23.2.1$,

$$d\overline{H} = \frac{I d\ell \times \hat{a}_R}{4\pi R^2} \tag{23.2.7}$$

which has the units of amperes per meter in the SI system. In Equation (23.2.7) $d\overline{H}$ is the incremental static H field, I is current, $d\ell$ is the vector of the length of a filament of current I, $\hat{a}_R$ 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 $d\overline{H}$ field is directed at right angles to $\hat{a}_R$ 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,

$$V = \oint_\ell \overline{\mathcal{E}} \cdot d\ell = -\oint_s \frac{\partial \overline{\mathcal{B}}}{\partial t} \cdot d\mathrm{s} \tag{23.2.8}$$

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

$$\oint_\ell \overline{H} \cdot d\ell = I_{\text{enclosed}} \tag{23.2.9}$$

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

$$H = |I|/(2\pi r) \tag{23.2.10}$$

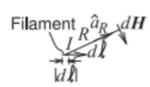


Figure $23.2.1$: Diagram illustrating the Biot-Savart law. The law relates a static filament of current to the incremental H field at a distance.

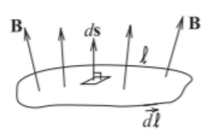


Figure $23.2.2$: Diagram illustrating Faraday's law. The contour ℓ encloses the surface.

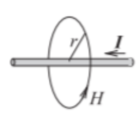


Figure $23.2.3$: Diagram illustrating Ampere's law. Ampere's law relates the current, I, on a wire to the magnetic field around it, H.

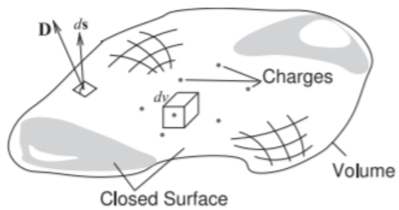


Figure $23.2.4$: Diagram illustrating Gauss's law. Charges are distributed in the volume enclosed by the closed surface. An incremental area is described by the vector $d\mathbf{S}$, which is normal to the surface and whose magnitude is the area of the incremental area.

1.3.5 Gauss's Law

The final static EM law is Gauss's law, which relates the static electric flux density vector, $\overline{\mathcal{D}}$, to charge. With reference to Figure $23.2.4$, Gauss's law in integral form is

$$\oint_s \overline{D} \cdot ds = \int_v \rho_v \cdot dv = Q_{\text{enclosed}} \tag{23.2.11}$$

This states that the integral of the electric flux vector, $\overline{D}$, 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

$$\oint_s \overline{B} \cdot ds = 0 \tag{23.2.12}$$

This states that the integral of the magnetic flux vector, $\overline{D}$, 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	. -

Symbol	Code
B	-...
C	-.-.
D	-..
E	.
F	..-.
G	--.
H	....
I	. .
J	.---
K	-.-
L	.-..
M	- -
N	- .
O	---
P	.--.
Q	--.-
R	.-.
S	...
T	-
U	..-
V	...-
W	.--
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

$$\nabla\times\overline{\mathcal{E}}=-\frac{\partial\overline{\mathcal{B}}}{\partial t}-\overline{\mathcal{M}} \tag{23.2.13}$$

$$\nabla\cdot\overline{\mathcal{D}}=\rho_V \tag{23.2.14}$$

$$\nabla\times\overline{\mathcal{H}}=\frac{\partial\overline{\mathcal{D}}}{\partial t}+\overline{\mathcal{J}} \tag{23.2.15}$$

$$\nabla\cdot\overline{\mathcal{B}}=\rho_{mV} \tag{23.2.16}$$

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

- $\overline{\mathcal{J}}$, the **electric current** density, with units of amperes per square meter (A/m^2);
- ρ_V, the **electric charge** density, with units of coulombs per cubic meter (C/m^3);
- ρ_{mV}, the magnetic charge density, with units of webers per cubic meter (Wb/m^3); and
- $\overline{\mathcal{M}}$, 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, ρ_{mV}, and the **magnetic current** density, $\overline{\mathcal{M}}$, 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 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, $\partial\overline{\mathcal{B}}/\partial t$ and $\partial\overline{\mathcal{D}}/\partial t$, depends on frequency. The more interesting property is how fast a field can change spatially, $\nabla\times\overline{\mathcal{E}}$ and $\nabla\times\overline{\mathcal{H}}$—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 E and 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

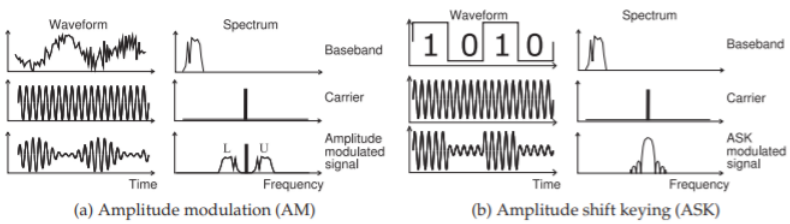


Figure $23.2.5$: Waveform and spectra of simple modulation schemes. The modulating signal, at the top in (a) and (b), is also called the baseband signal.

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.

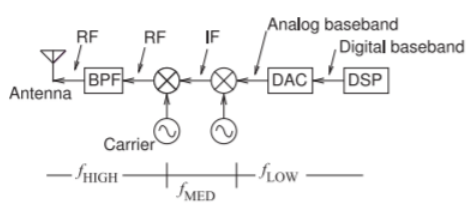


Figure $23.2.6$: A simple transmitter with low, f_{LOW}, medium, f_{MED}, and high frequency, f_{HIGH}, sections. The mixers can be idealized as multipliers, shown as circles with crosses, that boost the frequency of the input baseband or IF signal by the frequency of the carrier.

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.

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.

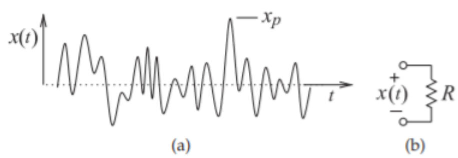


Figure $23.3.1$: Definition of crest factor: (a) arbitrary waveform; and (b) voltage across a resistor.

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 leads 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]

$$\text{CF} = x_p / x_{\text{rems}} \tag{23.3.1}$$

More formally,

$$\text{CF} = \frac{\|x\|_\infty}{\|x\|_2} \tag{23.3.2}$$

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 $x(t)$:

$$x_{\text{rms}} = \|x\|_2 = \lim_{T\to\infty} \sqrt{\frac{1}{T}\int_0^T x(t)\cdot dt} \tag{23.3.3}$$

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

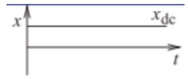
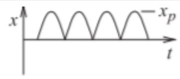
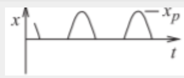
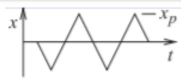
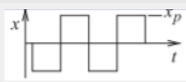
Waveform	$x(t)$	Max. value	rms (x_{rms})	CF	PAPR
DC		x_{dc}	x_{dc}	1	0 dB
Sinewave		x_p	$\frac{x_p}{\sqrt{2}}$	1.414	3.01 dB
Full-wave rectified sinewave		x_p	$\frac{x_p}{\sqrt{2}} = 0.717x_p$	1.414	3.01 dB
Half-wave rectified sinewave		x_p	$\frac{x_p}{2}$	2	6.02 dB
Triangle wave		x_p	$\frac{x_p}{\sqrt{3}} = 0.577x_p$	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

$$P_p = |x_p|^2/R \tag{23.3.4}$$

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

$$P_{\text{avg}} = |x_{\text{rms}}|^2/R \tag{23.3.5}$$

Then

$$\text{PAPR} = \frac{P_p}{P_{\text{avg}}} = \text{CF}^2 = (x_p/x_{\text{rms}})^2 \tag{23.3.6}$$

In decibels,

$$\begin{aligned} \text{PAPR}|_{\text{dB}} &= 10\log(\text{PAPR}) \\ &= 20\log(\text{CF}) = 20\log(x_p/x_{\text{rms}}) \end{aligned} \tag{23.3.7}$$

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),

$$\begin{aligned} x_{\text{rms}} &= \sqrt{\lim_{T\to\infty} \frac{1}{T}\int_0^T [x(t)]^2 dt} = \sqrt{\lim_{T\to\infty} \frac{1}{T}\int_0^T [a + b\sin(\omega t)]^2 dt} \\ &= \sqrt{\lim_{T\to\infty} \frac{1}{T}\int_0^T \left[a^2 + ab\sin(\omega t) + b^2\sin^2(\omega t)\right] dt} \\ &= \sqrt{\lim_{T\to\infty} \frac{1}{T}\left\{\int_0^T a^2 dt + \int_0^T ab\sin(\omega t)dt + \int_0^T b^2 \frac{1}{2}[1 + \cos(2\omega t)]\, dt\right\}} \\ &= \sqrt{\lim_{T\to\infty} \frac{1}{T}\left\{a^2 T dt + 0 + \frac{1}{2}b^2 T\right\}} \end{aligned} \tag{23.3.8}$$

since the integral of sin and cos over a period is zero. Thus

$$x_{\text{rms}} = \sqrt{a^2 + b^2/2} = \sqrt{0.1^2 + \frac{1}{2}0.5^2} = 0.3674 \tag{23.3.9}$$

the crest factor is

$$\text{CF} = \frac{x_p}{x_{\text{rms}}} = \frac{0.6}{0.3674} = 1.6311 \tag{23.3.10}$$

and PAPR is

$$\text{PAPR} = 20\log(1.6311) = 4.260\text{ dB} \tag{23.3.11}$$

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

$$P_{\text{avg}} = \frac{1}{R}\left[0.1^2 + \frac{1}{2}0.5^2\right]\frac{1}{R} = \frac{0.1350}{R} \tag{23.3.12}$$

Thus, in decibels,

$$\text{PAPR}|_{\text{dB}} = 10\log\left(\frac{P_p}{P_{\text{avg}}}\right) = \frac{x_p^2}{x_{\text{rms}}^2} = 10\log\left(\frac{0.36}{0.135}\right) = 10\log(2.667) = 4.260\text{ dB} \tag{23.3.13}$$

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}

$$\text{PMEPR} = \frac{\text{PEP}}{P_{\text{avg}}} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} \tag{23.3.14}$$

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

$$x(t)=A_c\left[1+m\cos(2\pi f_m t)\right]\cos(2\pi f_c t) \tag{23.3.15}$$

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:

$$P_{\text{avg}}=\lim_{\tau\to\infty}\int_{-\tau}^{\tau}\frac{1}{2\tau}Gv^2(t)dt \tag{23.3.16}$$

Now, if $v(t)$ is a sinusoidal, $v(t)=A\cos\omega t$, then

$$\begin{aligned}P_{\text{avg}} &=\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}A_c^2G\cos^2(\omega t)dt\\ &=\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}A_c^2G\frac{1}{2}\left[1+\cos(2\omega t)\right]dt\\ &=\frac{1}{2}A_c^2G\left\{\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}1dt+\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos(2\omega t)dt\right\}=\frac{1}{2}A_c^2G\end{aligned} \tag{23.3.17}$$

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, $T=2\pi/\omega$:

$$\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos^n(\omega t)dt=\frac{1}{T}\int_{-T/2}^{T/2}\cos^n(\omega t)dt \tag{23.3.18}$$

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:

$$\begin{aligned}\cos A\cos B &=\frac{1}{2}[\cos(A-B)+\cos(A+B)]\\ \cos^2 A &=\frac{1}{2}[1+\cos(2A)]\end{aligned} \tag{23.3.19}$$

$$\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos\omega t dt=\frac{1}{T}\int_{-T/2}^{T/2}\cos(\omega t)dt=0 \tag{23.3.20}$$

$$\frac{1}{T}\int_{-T/2}^{T/2}\cos^2(\omega t)dt=\frac{1}{T}\int_{-T/2}^{T/2}\frac{1}{2}[\cos(2\omega t)+\cos(0)] \tag{23.3.21}$$

$$=\frac{1}{2T}\left[\int_{-T/2}^{T/2}\cos(2\omega t)dt+\int_{-T/2}^{T/2}1dt\right] \tag{23.3.22}$$

$$=\frac{1}{2T}(0+T)=\frac{1}{2} \tag{23.3.23}$$

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,

$$\int_{-\tau}^{\tau}\cos\omega_A t\cos\omega_B t dt=\int_{-\tau}^{\tau}\frac{1}{2}\left[\cos(\omega_A+\omega_B)t+\cos(\omega_A-\omega_B)t\right]dt=0$$

and if $\omega_A\neq\omega_B\neq 0$

$$\int_{-\infty}^{\infty}\cos\omega_A t\cos^n\omega_B t dt=0 \tag{23.3.24}$$

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),

$$x_p(t) = A_c[1+m]\cos(2\pi f_c t) \tag{23.3.25}$$

Then the power (P_{PEP}) of the peak pseudo carrier is obtained by integrating over one period of the pseudo carrier:

$$\begin{aligned} P_{\text{PEP}} &= \frac{1}{T}\int_{-T/2}^{T/2} Gx^2(t)dt = \frac{1}{T}\int_{-T/2}^{T/2} A_c^2 G(1+m)^2\cos^2(\omega_c t)dt \\ &= A_c^2 G(1+m)^2 \frac{1}{T}\int_{-T/2}^{T/2}\cos^2(\omega_c t)dt = \frac{1}{2}A_c^2 G(1+m)^2 \end{aligned} \tag{23.3.26}$$

The **average power** (P_{avg}) of the modulated signal is obtained by integrating over all time, so

$$\begin{aligned} P_{\text{avg}} &= \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} Gx^2(t)dt \\ &= A_c^2 G\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\{[1+m\cos(\omega_m t)]\cos(\omega_c t\}^2 dt \\ &= A_c^2 G\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\{[1+2m\cos(\omega_m t)+m^2\cos^2(\omega_m t)]\cos^2(\omega_c t)\}dt \\ &= A_c^2 G\left[\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos^2(\omega_c t)dt + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} 2m\cos(\omega_m t)\cos^2(\omega_c t)dt\right. \\ &\quad\left. + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau} m^2\cos^2(\omega_m t)\cos^2(\omega_c t)dt\right] \\ &= A_c^2 G\left\{\frac{1}{2}+0+m^2\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\frac{1}{4}[1+\cos(2\omega_m t)]\,[1+\cos(2\omega_c t)]\,dt\right\} \\ &= A_c^2 G\left\{\frac{1}{2}+\frac{m^2}{4}\left[\lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}1dt + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos(2\omega_m t)dt\right.\right. \\ &\quad\left.\left. + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos(2\omega_c t)dt + \lim_{\tau\to\infty}\frac{1}{2\tau}\int_{-\tau}^{\tau}\cos(2\omega_m t)\cos(2\omega_c t)dt\right]\right\} \\ &= A_c^2 G\left[\frac{1}{2}+m^2\left(\frac{1}{4}+0+0+0\right)\right] = \frac{1}{2}A_c^2 G(1+m^2/2) \end{aligned} \tag{23.3.27}$$

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

$$\text{PMEPR}_{100\%\text{AM}} = \frac{(1+1)^2}{1+1^2/2} = \frac{4}{1.5} = 2.667 = 4.26\text{ dB} \tag{23.3.28}$$

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

$$\text{PMEPR}_{50\%\text{AM}} = \frac{(1+0.5)^2}{1+0.5^2/2} = \frac{2.25}{1.125} = 2 = 3\text{ dB} \tag{23.3.29}$$

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:

$$y(t) = X_A \cos(\omega_A t) + X_B \cos(\omega_B t) \tag{23.3.30}$$

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 $X_A + X_B$. The peak pseudo carrier is the single RF sinusoid where the sinusoid 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.

$$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{2X^2}{X^2} = 2 = 3 \text{ dB} \tag{23.3.31}$$

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

$$\text{PMEPR} = \frac{P_{\text{PEP}}}{P_{\text{avg}}} = \frac{0.01125}{0.00625} = 1.8 = 2.55 \text{ dB} \tag{23.3.32}$$

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

$$P_{\text{peak, PAPR}} = (2A_c)^2 G \tag{23.3.33}$$

P_{avg} is the same for PAPR and PMEPR for the AM signal, see Equation (23.3.27), so that

$$\text{PAPR} = \frac{P_{\text{peak, PAPR}}}{P_{\text{avg}}} = \frac{(2A_c)^2 G}{\frac{1}{2}A_c^2(1+\frac{1}{2})} = \frac{4}{3/4} = \frac{16}{3} = 5.333 = 7.27 \text{ dB} \tag{23.3.34}$$

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.

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.

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

$$x(t)=A_c\left[1+my(t)\right]\cos(2\pi f_c t) \tag{23.5.1}$$

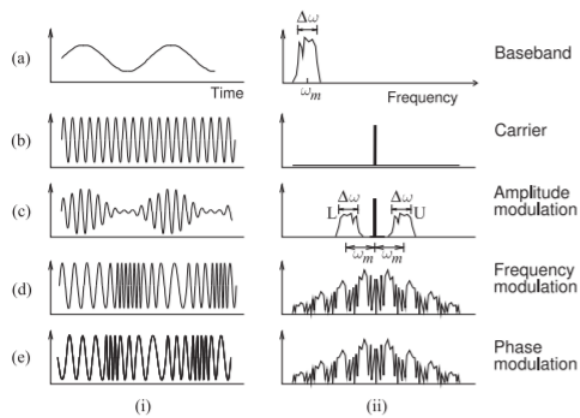


Figure $23.5.1$: Basic analog modulation showing the (i) waveform and (ii) spectrum for (a) baseband signal; (b) carrier; (c) carrier modulated using amplitude modulation; (d) carrier modulated using frequency modulation; and (e) carrier modulated using phase modulation.

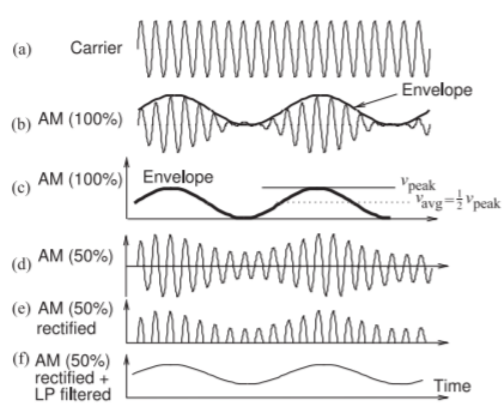


Figure $23.5.2$: AM showing the relationship between the carrier and modulation envelope: (a) carrier; (b) 100% amplitude modulated carrier; (c) modulating or baseband signal; (d) 50% amplitude-modulated carrier; (e) rectified 50% AM modulated signal; and (f) rectified and lowpass (LP) filtered 50% modulated 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 $1\ \mathrm{MHz}$ (which is in the middle of the AM broadcast band). The highest frequency component of the modulating signal corresponding to voice is about $4\ \mathrm{kHz}$. 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

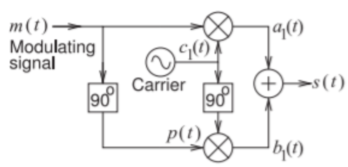


Figure $23.5.3$: Hartley modulator implementing singlesideband suppressed-carrier (SSB-SC) modulation. The "90°" blocks shift the phase of the signal by $+90^{\circ}$. The mixer indicated by the circle with a cross is an ideal multiplier, e.g. $a_1(t) = m(t) \cdot c_1(t)$.

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

$$\begin{aligned} x(t) &= [1 + m\cos(\omega_m t + \phi)]\cos(\omega_c t) \\ &= \cos(\omega_c t) + \frac{1}{2} m\left[\cos((\omega_c - \omega_m)t - \phi) + \cos((\omega_c + \omega_m)t + \phi)\right] \end{aligned} \tag{23.5.2}$$

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 ωm 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 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 $c_1(t)=\cos(\omega_c t)$:

$$\begin{aligned} a_1(t) &=\cos(\omega_m t+\phi)\cos(\omega_c t)=\frac{1}{2}[\cos((\omega_c-\omega_m)t-\phi)+\cos((\omega_c+\omega_m)t+\phi)] \\ b_1(t) &=\sin(\omega_m t+\phi)\sin(\omega_c t)=\frac{1}{2}[\cos((\omega_c-\omega_m)t-\phi)-\cos((\omega_c+\omega_m)t+\phi)] \\ s(t) &=a_1(t)+b_1(t)=\cos((\omega_c-\omega_m)t-\phi) \end{aligned} \tag{23.5.3}$$

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

$$\begin{aligned} s(t) &=\cos(\omega_c t+\beta\cos(\omega_m t)]) \\ &=\cos(\omega_c t)\cos(\cos(\beta\omega_m t))-\sin(\omega_c t)\sin(\cos(\beta\omega_m t)) \end{aligned} \tag{23.5.4}$$

which has the Bessel function-based expansion

$$\begin{aligned} & s(t)=J_0(\beta)\cos(\omega_c t) \\ & \quad +J_1(\beta)\cos(\omega_c+\omega_m)t+\pi/2)+J_1(\beta)\cos(\omega_c-\omega_m)t+\pi/2) \\ & \quad +J_2(\beta)\cos(\omega_c+2\omega_m)t+\pi)+J_2(\beta)\cos(\omega_c-2\omega_m)t+\pi] \\ & \quad +J_3(\beta)\cos(\omega_c+3\omega_m)t+3\pi/2)+J_3(\beta)\cos(\omega_c-3\omega_m)t+3\pi/2)+\dots \end{aligned} \tag{23.5.5}$$

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 fm as well. This is DSB

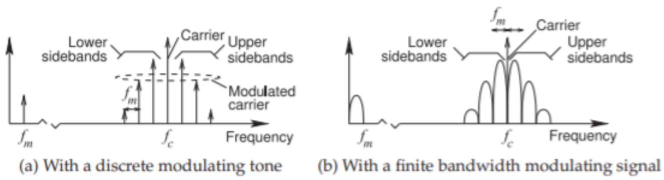


Figure $23.5.4$: Spectrum of a phase-modulated carrier which includes the carrier at f_c and upper and lower sidebands with the spectrum of the discrete modulating signal at f_m.

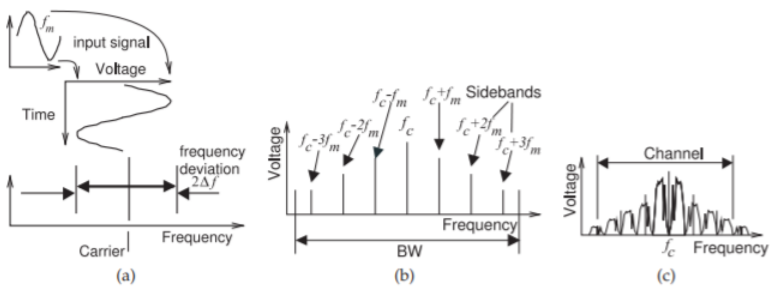


Figure $23.5.5$: Frequency modulation: (a) sinusoidal baseband signal shown varying the frequency of the carrier and so FM modulating the carrier; (b) the spectrum of the resulting FM-modulated waveform; and (c) spectrum of the modulated carrier when it is modulated by a broadband baseband signal such 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 is 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

$$\begin{aligned} s(t) &= \cos([\omega_c + \beta\cos(\omega_m t)]t) \\ &= \cos(\omega_c t)\cos(\cos(\omega_m t)\beta t) - \sin(\omega_c t)\sin(\cos(\omega_m t)\beta t) \end{aligned} \tag{23.5.6}$$

which has the Bessel function-based expansion

$$\begin{aligned} s(t) = & J_0(\beta t)\cos(\omega_c t) \\ & - J_1(\beta)\sin(\omega_c+\omega_m)t+\pi/2) - J_1(\beta t)\sin(\omega_c-\omega_m)t+\pi/2) \\ & - J_2(\beta t)\cos(\omega_c-2\omega_m)t+\pi) + K_2(\beta t)\cos(\omega_c-2\omega_m)t+\pi] \\ & + J_3(\beta t)\sin(\omega_c+3\omega_m)t+3\pi/2) + J_3(\beta t)\sin(\omega_c-3\omega_m)t+3\pi/2)+\ldots \end{aligned} \tag{23.5.7}$$

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 to 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.

a. What are PAPR and PMEPR of just one FM signal?
b. 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

a. An FM signal has a constant envelope just like a single sinusoid, and so $\text{PAPR} = 1.414 = 3.01\text{ dB}$ and $\text{PMEPR} = 1 = 0\text{ dB}$.
b. 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,

$$P_P = k(2X)^2 \quad \text{and} \quad \text{PMEPR} = \frac{P_P}{P_{\text{avg}}} = \frac{k(2X)^2}{2k\frac{1}{2}X^2} = 4 = 6.0 \text{ dB} \tag{23.5.9}$$

For PMEPR, $P_{\text{PEP}} = \text{ power of the pseudo-carrier } = k\frac{1}{2}(2X)^2$
and

$$\text{PMEPR} = \frac{P_P''}{P_{\text{avg}}} = \frac{k\frac{1}{2}(2X)^2}{2k\frac{1}{2}X^2} = 2 = 3.0 \text{ dB} \tag{23.5.10}$$

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.

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

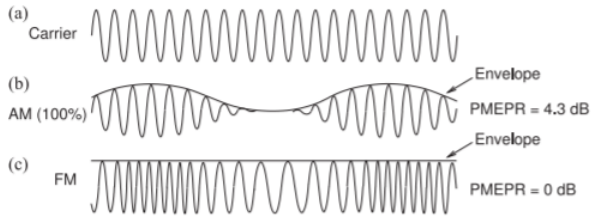


Figure $23.6.1$: Comparison of 100% AM and FM highlighting the envelopes of both: (a) carrier; (b) AM signal; and (c) FM signal with constant envelope.

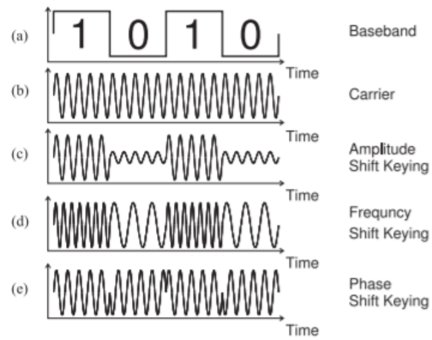


Figure $23.6.2$: Modes of digital modulation: (a) modulating bitstream; (b) **carrier**; (c) carrier modulated using amplitude shift keying (ASK); (d) carrier modulated using frequency shift keying (FSK); and (e) carrier modulated using binary phase shift keying (BPSK).

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 more 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 that 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

$$\text{modulation efficiency } = \eta_c = \frac{\text{gross bit rate}}{\text{bandwidth}} \tag{23.6.1}$$

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

$$\text{link spectrum efficiency } = \frac{\text{information bit rate}}{\text{bandwidth}} \leq \eta_c \tag{23.6.2}$$

Example 23.6.1: Modulation Efficiency

A radio transmits a bit stream of 2 Mbits/s using a bandwidth of 1 MHz.

a. What is the modulation efficiency?
b. If 25% of the bits are used for error correction, what is the modulation efficiency?
c. With error correction coding, what is the information rate?
d. 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.

d.
$$\text{link spectrum efficiency } = \frac{\text{information bit rate}}{\text{bandwidth}} = \frac{1.5 \text{ Mbits/s}}{1 \text{ MHz}} = 1.5 \text{ bits/s/Hz}$$

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

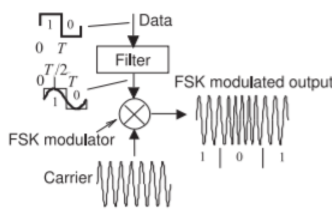


Figure $23.7.1$: The frequency shift keying (FSK) modulation system. In the GSM four-state cellular system-adjacent constellation points differ in frequency by 33.25 kHz.

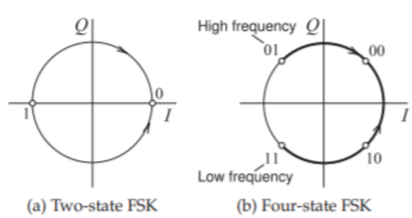


Figure $23.7.2$: Constellation diagrams of FSK modulation. In two-state FSK a symbol indicates whether a bit is a '0' or a '1'. In four-state FSK there are four symbols and each symbol has a different frequency and indicates the values of two bits.

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 $(1\text{ GHz} - 33.25\text{ kHz})$, $(1\text{ GHz} - 16.62\text{ kHz})$, $(1\text{ GHz} + 16.62\text{ kHz})$, and $(1\text{ GHz} + 33.25\text{ kHz})$. 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 $\text{GMSK, PMEPR } = 3.01\text{ dB},\ 6.02\text{ dB},\ 9.01\text{ dB},\ 11.40\text{ dB},\ 14.26\text{ dB},$ 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

$$\Delta f = f v_s / c \tag{23.7.1}$$

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.

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 course 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.

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.

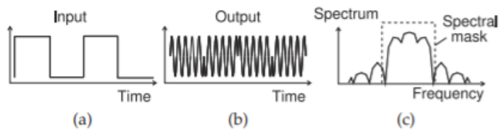


Figure $23.9.1$: Binary PSK modulation: (a) modulating bitstream; (b) the modulated waveform; and (c) its spectrum after smoothing the transitions from one phase state to another.

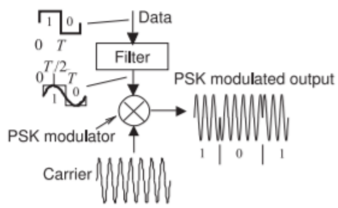


Figure $23.9.2$: A binary phase shift keying (PSK) modulation system.

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,

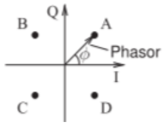


Figure $23.9.3$: Phasor diagram of QPSK modulation. Here there are four discrete phase states of the phasor indicated by the points A, B, C, and D. The PSK modulator moves the phasor from one phase state to another. The task at the receiver is determining the phase of the phasor.

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

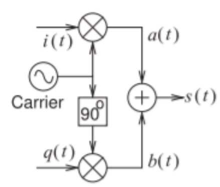


Figure $23.9.4$: Quadrature modulator block diagram. In PSK modulation $i(t)$ and $q(t)$ have the same amplitude and indicate a phase ϕ of the modulated carrier so that $i(t) = \cos[\phi(t)]$ and $q(t) = \sin[\phi(t)]$. The particular example shows two possible values of I_k and Q_k and this indicates 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]).

$$a(t) = \cos(\phi(t))\cos(\omega_c t) = \frac{1}{2}[\cos(\omega_c t - \phi(t)) + \cos(\omega_c t + \phi(t))] \tag{23.9.1}$$

$$b(t) = \sin(\phi(t))\,[-\sin(\omega_c t)] = -\frac{1}{2}[\cos(\omega_c - \phi(t)) - \cos(\omega_c t + \phi(t))] \tag{23.9.2}$$

$$s(t) = a(t) + b(t) = \cos(\omega_c t + \phi(t)) \tag{23.9.3}$$

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 $1\ \text{bits/s/Hz}$. 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

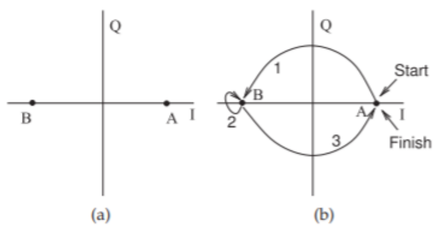


Figure $23.9.5$: BPSK modulation with constellation points A and B: (a) constellation diagram; and (b) constellation diagram with possible transitions from one phase state to the other, or possibly no change in the phase state. In practical systems the transition should not go through the origin, as then the RF signal would drop below the noise level.

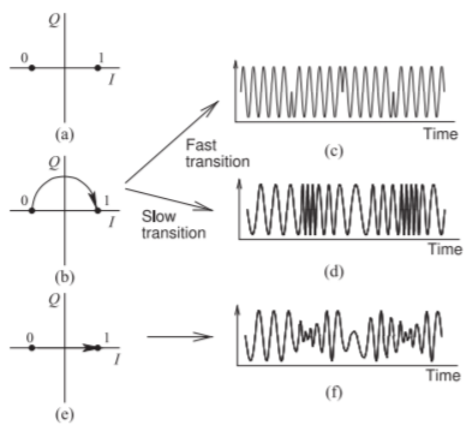


Figure $23.9.6$: BPSK modulation: (a) constellation diagram; (b) constellation diagram with a constant amplitude transition; (c) time-domain waveform if the transition is fast; (d) time-domain waveform if the transition is slow; (e) constellation diagram with transition through the origin; and (f) time-domain 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

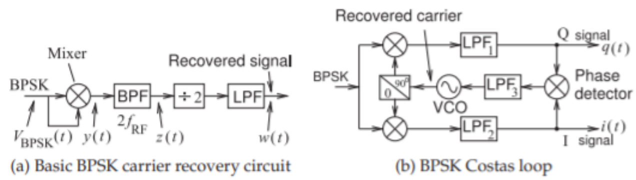


Figure $23.9.7$: Block diagram of carrier recovery circuits for BPSK signals.

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

$$v_{\mathrm{BPSK}}(t) = A(t)\cos(\omega_{\mathrm{RF}}t + n\pi) \tag{23.9.4}$$

where the carrier frequency $f_{\mathrm{RF}} = \omega_{\mathrm{RF}}/(2\pi)$ and n can have a value of 0 or 1. Squaring this produces a signal

$$\begin{aligned} y(t) = v_{\mathrm{BPSK}}^2(t) &= A^2(t)\cos^2(\omega_{\mathrm{RF}}t + n\pi) = \frac{1}{2}A^2(t)[1 + \cos(2\omega_{\mathrm{RF}}t + n2\pi)] \\ &= \frac{1}{2}A^2(t)[1 + \cos(2\omega_{\mathrm{RF}}t)] \end{aligned} \tag{23.9.5}$$

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 $v_{\mathrm{BPSK}}(t)$ 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.

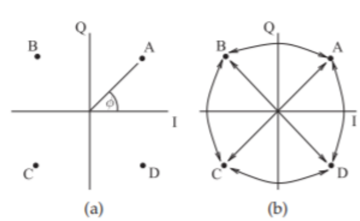


Figure $23.9.8$: QPSK modulation: (a) constellation diagram; and (b) constellation diagram with possible transitions. Each constellation point indicates the phase, ϕ, of the modulated carrier, i.e. $\cos(\omega_c t + \phi)$ where ω_c is the radian frequency of the carrier.

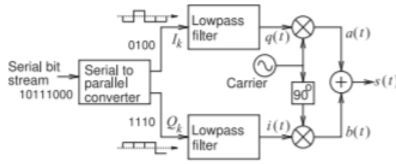


Figure 23.9.9: QPSK modulator block diagram. I_k and Q_k are similar to a stream of one-bit binary signals but are analog with a either a postitive value or a negative voltage so that after lowpass filtering $i(t)$ and $q(t)$ each have either a positive or a negative voltage at each clock tick.

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 Kth 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

$$v_{\text{QPSK}}(t) = A(t)\cos(\omega_{\text{RF}}t + n\pi/2); \quad n = 0, 1, 2, 3, \tag{23.9.6}$$

where the carrier frequency $f_{\text{RF}} = \omega_{\text{RF}}/(2\pi)$. The fourth power of this produces

$$\begin{aligned} v^4_{\text{QPSK}}(t) &= A^4(t)\cos^4(\omega_{\text{RF}}t + n\pi/2) \\ &= \frac{1}{8}A^4(t)[3 + 4\cos(2\omega_{\text{RF}}t + n\pi) + \cos(4\omega_{\text{RF}}t + n2\pi)] \end{aligned} \tag{23.9.7}$$

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

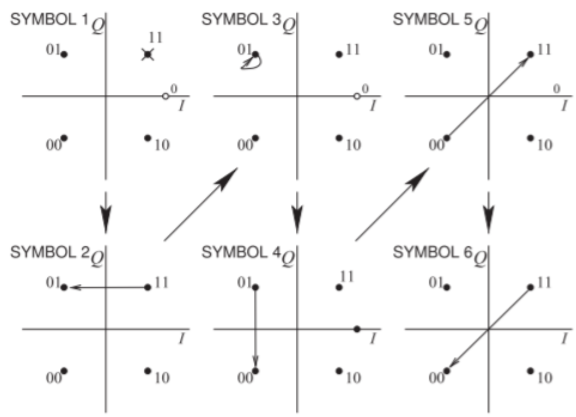


Figure $23.9.10$ Constellation diagram and transitions for the bit sequence 110101001100 sent as the set of symbols 110101001100 using QPSK. Note that symbols 2 and 3 are identical, so there is no transition. The SYMBOL numbers indicated reference the symbol at the end of the transition (end of the arrow). 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.

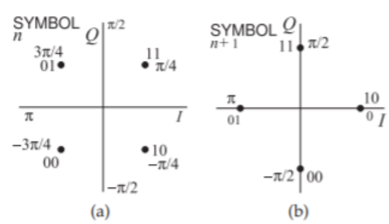


Figure $23.9.11$: Constellation diagram of $\pi/4$- QPSK modulation: (a) initial constellation diagram at one symbol; and (b) the constellation diagram at the time of the next symbol.

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

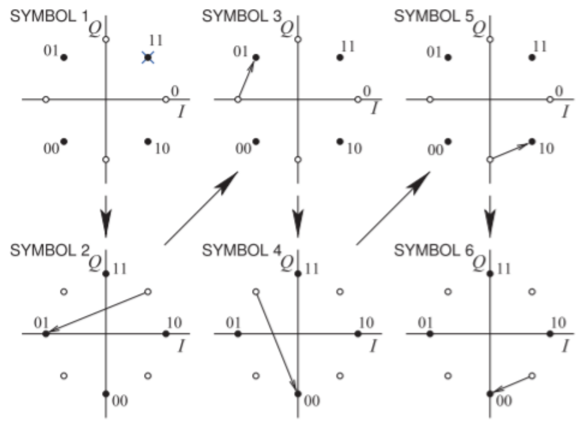


Figure $23.9.12$ Constellation diagram states and transitions for the bit sequence 110101001000 sent as the set of symbols 110101001000 using $\pi/4$QPSK modulation.

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

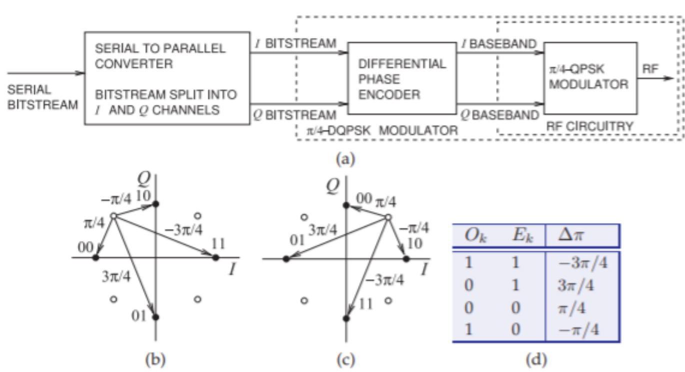


Figure $23.9.13$: A $\pi/4$-DQPSK modulator: (a) differential phase encoder with a $\pi/4$-QPSK modulator; (b) constellation diagram of $\pi/4$-DQPSK; (c) a second constellation diagram; and (d) phase changes in a $\pi/4$-DQPSK modulation scheme. Note that the information is in the phase change rather than the phase state.

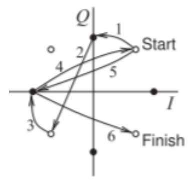


Figure $23.9.14$: Constellation diagram of $\pi/4$- DQPSK modulation showing six symbol intervals coding the bit sequence $\mathsf{000110110101}$

$\pi/4$-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 $\mathsf{000110110101}$ 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

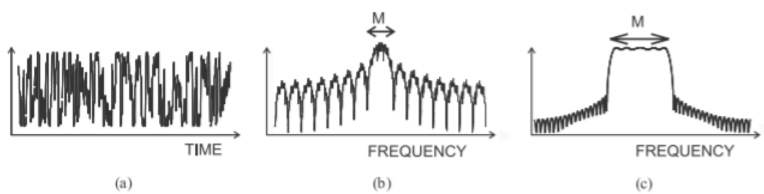


Figure $23.9.15$: Details of digital modulation obtained using differential phase shift keying ($\pi/4$- DQPSK): (a) modulating waveform; (b) spectrum of the modulated carrier, with M denoting the main channel; and (c) details of the spectrum of the modulated carrier focusing on the main channel.

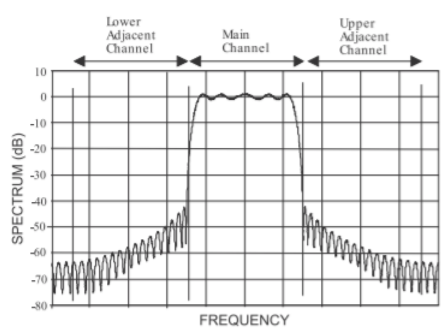


Figure $23.9.16$: Detailed spectrum of a $\pi/4$-DQPSK signal showing the main channel and lower and upper adjacent channels.

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

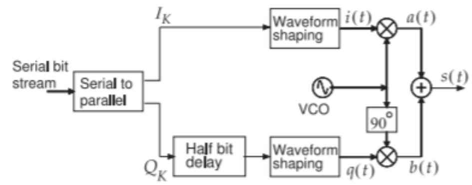


Figure $23.9.17$: Block diagram of an OQPSK modulator.

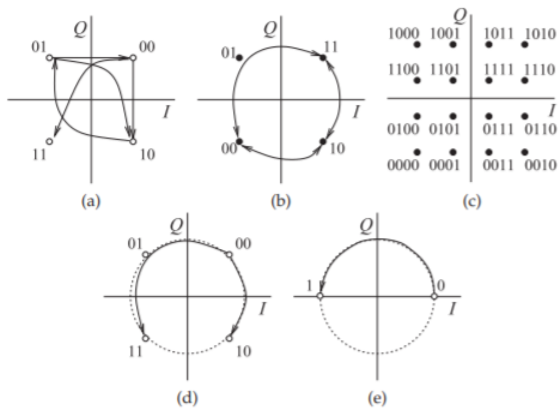


Figure $23.9.18$: Constellation diagrams for various modulation formats: (a) OQPSK; (b) GMSK; (c) 16-QAM; (d) SOQPSK (also FOQPSK); and (e) SBPSK.

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

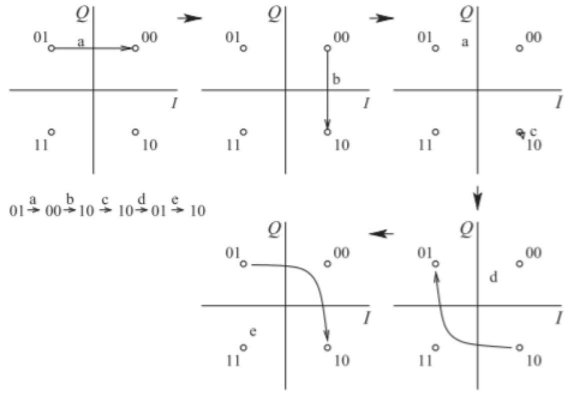


Figure $23.9.19$: Constellation diagram of OQPSK modulation for the bit sequence 010010100110

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 kbits/s, faster than the 48 kbits/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

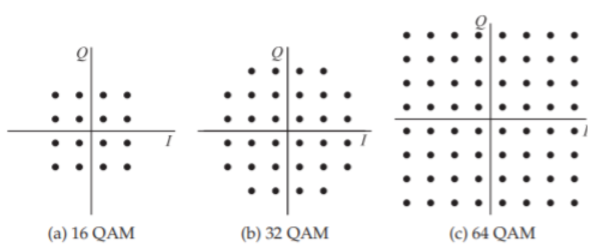


Figure $23.9.20$: QAM constellation diagrams.

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.”

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

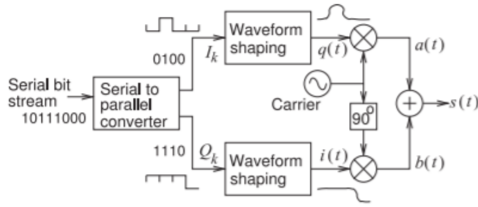


Figure $23.10.1$: QAM modulator block diagram. In QAM modulation $i(t)$ and $q(t)$ address the real and imaginary components of a phasor. The wave-shaping block ensures that the symbol has the correct amplitude and phase at each clock tick.

Modulation	$\mathrm{bits/s/Hz}$
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 $\mathrm{bits/s/Hz}$ (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 $3pi/8$-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.

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.

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.

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 Ω 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 $1\ \mathrm{GHz}$ with a maximum modulated frequency deviation of $\pm 10\ \mathrm{kHz}$?
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 $20\ \mathrm{kHz}$ and a modulating signal between $300\ \mathrm{Hz}$ and $5\ \mathrm{kHz}$. What is the bandwidth required to transmit the modulated RF signal when the carrier is $200\ \mathrm{MHz}$? Is this considered to be narrowband FM or wideband FM?
17. A high-fidelity stereo audio signal has a frequency content ranging from $50\ \mathrm{Hz}$ to $20\ \mathrm{kHz}$. If the signal is to be modulated on an FM carrier at $100\ \mathrm{MHz}$, what is the bandwidth required for the modulated RF signal? The maximum frequency deviation is $5\ \mathrm{kHz}$ 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%.

a. What is the PMEPR in dB of just one FM modulated signal?
b. 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.
    a. What is the PAPR of the AM signal $x(t)$?
    b. What is the PAPR of the FM signal $z(t)$?
    c. What is the PAPR of $x(t)$?
    d. What is the PMEPR of the AM signal $x(t)$?
    e. What is the PMEPR of the FM signal $z(t)$?
    f. 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 $25\ \mathrm{kHz}$ and a modulation scheme with a gross bit rate of $100\ \mathrm{kbits/s}$ that is made of an information bit rate of $60\ \mathrm{kbits/s}$ and a code bit rate of $40\ \mathrm{kbits/s}$.
    a. What is the modulation efficiency in $\mathrm{bits/s/Hz}$?
    b. What is the spectral efficiency in $\mathrm{bits/s/Hz}$?
23. A cellular communication system uses $\pi/4$- DQPSK modulation with a modulation efficiency of $1.63\ \mathrm{bits/s/Hz}$ to transmit data at the rate of $30\ \mathrm{kbits/s}$. 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.
    a. What is the gross bit rate?
    b. What is the information bit rate?
    c. What is the bandwidth required to transmit the information and code bits?
    d. What is the spectral efficiency in $\mathrm{bits/s/Hz}$?
24. A radio uses a channel with a $5\ \mathrm{MHz}$ bandwidth and uses 256-QAM modulation with a modulation efficiency of $6.33\ \mathrm{bits/s/Hz}$. 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).
    a. What is gross bit rate in $\mathrm{Mbits/s}$?
    b. What is information rate in $\mathrm{Mbits/s}$?
    c. What is the spectral efficiency in $\mathrm{bits/s/Hz}$?
25. The following sequence of bits 0100110111is 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 0100110111is 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 0100110111is 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 $5\ \mathrm{V}$. If the noise on the signal has an rms value of $0.2\ \mathrm{V}$, 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

$\S 12.21, 2, 3, 4, 5, 6, 7, 8, 910, 11, 12$

$\S 12.413, 14, 15, 16, 17, 18, 1920, 21$

$\S 12.522, 23, 24$

$\S 12.825, 26, 27, 28, 29, 30, 31$

$\S 12.932, 33, 34, 35, 36, 37, 38$

$\S 12.1139, 40, 41, 42, 43$

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

Index

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: Electricity and Magnetism with Applications to Amateur Radio and Wireless Technology

Webpages: 271

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 63.8% (173 pages)
- CC BY-SA 4.0: 14% (38 pages)
- Undeclared: 13.7% (37 pages)
- CC BY-NC 4.0: 4.8% (13 pages)
- GNU Free Documentation License: 2.6% (7 pages)
- CC BY 1.0: 1.1% (3 pages)

By Page

- Electricity and Magnetism with Applications to Amateur Radio and Wireless Technology - *CC BY-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *CC BY 4.0*
    - Licensing - *Undeclared*
    - Preface - *Undeclared*
  - 1: Preliminary Concepts - *CC BY 4.0*
    - 1.1: What is Electricity and Magnetism? - *CC BY-SA 4.0*
    - 1.2: Wireless Technology and Amateur Radio - What and Why? - *CC BY-SA 4.0*
    - 1.3: Units - *CC BY-SA 4.0*
    - 1.4: Electromagnetic Spectrum - *CC BY-SA 4.0*
    - 1.5: Amateur Radio Equipment Basics - *CC BY-SA 4.0*
    - 1.6: Notation - *CC BY-SA 4.0*
    - 1.7: Coordinate Systems - *CC BY-SA 4.0*
    - 1.8: Where Do We Go from Here? - *CC BY-SA 4.0*
  - 2: The Electric Field - *CC BY 4.0*
    - 2.1: Introduction - *CC BY-SA 4.0*
    - 2.2: Electric Charge Model - *CC BY 4.0*
    - 2.3: Conduction and Charging - *CC BY 4.0*
    - 2.4: Electric Fields and Forces - *CC BY-SA 4.0*
    - 2.5: Electric Fields and Forces with Multiple Charges - *CC BY 4.0*
    - 2.6: Electric Field Diagrams - *CC BY 4.0*
    - 2.7: Common Models of Electric Field - *CC BY 4.0*
    - 2.8: Motion of a Charged Particle in an Electric Field - *CC BY 4.0*
    - 2.9: Conclusion - *CC BY-SA 4.0*
    - 2.10: The Electric Field (Summary) - *CC BY 4.0*
    - 2.11: The Electric Field (Exercises) - *CC BY 4.0*
    - 2.12: The Electric Field (Answers) - *CC BY 4.0*
  - 3: The Electric Potential - *Undeclared*
    - 3.1: Introduction - *CC BY 4.0*
    - 3.2: Work and Energy - *CC BY 4.0*
    - 3.3: Electric Potential Energy - *CC BY 4.0*
    - 3.4: Electric Potential Energy of Point Charges - *CC BY 4.0*
    - 3.5: Electric Potential - *CC BY 4.0*
    - 3.6: Electric Potential of a Point Charge - *CC BY 4.0*
    - 3.7: Common Models of Electric Potential - *CC BY 4.0*
    - 3.8: Electric Potential (Summary) - *CC BY 4.0*
    - 3.9: The Electric Potential (Exercises) - *CC BY 4.0*
    - 3.10: The Electric Potential (Answers) - *CC BY 4.0*
  - 4: Potential and Field Relationships - *Undeclared*
    - 4.1: Electric Potential from Electric Field - *CC BY 4.0*
    - 4.2: Electric Field from Electric Potential - *CC BY 4.0*
    - 4.3: Equipotential Curves and Surfaces - *CC BY 4.0*
    - 4.4: Conductors in Electrostatic Equilibrium - *CC BY 4.0*
    - 4.5: Applications of Electric Potential and Conductors in Electrostatic Equilibrium - *CC BY 4.0*
    - 4.6: Potential and Field Relationships (Summary) - *CC BY 4.0*
    - 4.7: Potential and Field Relationships (Exercises) - *CC BY 4.0*
    - 4.8: Potential and Field Relationships (Answers) - *CC BY 4.0*
  - 5: Electric Current and Resistance - *Undeclared*
    - 5.1: Introduction - *CC BY 4.0*
    - 5.2: Electric Current - *CC BY 4.0*

- 12: Antenna Systems - *Undeclared*
  - 12.1: Introduction - *CC BY 4.0*
  - 12.2: Production of Electromagnetic Waves - *CC BY 4.0*
  - 12.3: Transmission Lines and Characteristic Impedance - *GNU Free Documentation License*
  - 12.4: Finite-length Transmission Lines - *GNU Free Documentation License*
  - 12.5: "Long" and "Short" Transmission Lines - *GNU Free Documentation License*
  - 12.6: Standing Waves and Resonance - *GNU Free Documentation License*
  - 12.7: Antenna Systems (Summary) - *CC BY 4.0*
- 13: Propagation of Electromagnetic Waves - *Undeclared*
  - 13.1: Introduction - *CC BY 4.0*
  - 13.2: Ray and Wave Models of Propagation - *CC BY 4.0*
  - 13.3: Reflection of Rays - *CC BY 4.0*
  - 13.4: Refraction of Rays - *CC BY 4.0*
  - 13.5: Application- Line-of-Sight Transmission - *CC BY 1.0*
  - 13.6: Diffraction of Waves - *CC BY 4.0*
  - 13.7: Interference of Waves - *CC BY 4.0*
  - 13.8: Double-Slit Interference - *CC BY 4.0*
  - 13.9: Propagation of Electromagnetic Waves (Summary) - *CC BY 4.0*
  - 13.10: Propagation of Electromagnetic Waves (Exercises) - *CC BY 4.0*
  - 13.11: Propagation of Electromagnetic Waves (Answers) - *CC BY 4.0*
- 14: Introduction to Semiconductor Devices - *Undeclared*
  - 14.1: Introduction - *CC BY 4.0*
  - 14.2: Band Theory of Solids - *CC BY 4.0*
  - 14.3: Semiconductors and Doping - *CC BY 4.0*
  - 14.4: Introduction to Semiconductor Devices - *CC BY 4.0*
  - 14.5: Junction Diodes - *GNU Free Documentation License*
  - 14.6: Light Emitting Diode - *CC BY 1.0*
  - 14.7: Solar Cells - *CC BY 1.0*
  - 14.8: Bipolar Junction Transistors - *GNU Free Documentation License*
  - 14.9: Junction Field-effect Transistors - *GNU Free Documentation License*
- 15: Part 2 - Detailed and/or Advanced Content - *Undeclared*
- 16: Direct Calculation of Electrical Quantities from Charge Distributions - *Undeclared*
  - 16.1: Introduction - *CC BY 4.0*
  - 16.2: Electric Dipoles - *CC BY 4.0*
  - 16.3: Calculating Electric Fields of Charge Distributions - *CC BY 4.0*
  - 16.4: Calculating Electric Potential of Charge Distributions - *CC BY 4.0*
  - 16.5: Direct Calculation of Electrical Quantities from Charge Distributions (Summary) - *CC BY 4.0*
  - 16.6: Direct Calculation of Electrical Quantities from Charge Distributions (Exercises) - *Undeclared*
  - 16.7: Direct Calculation of Electrical Quantities from Charge Distributions (Answers) - *Undeclared*
- 17: Gauss's Law for Calculation of Electrical Field from Charge Distributions - *Undeclared*
  - 17.1: Introduction to Gauss's Law - *CC BY 4.0*
  - 17.2: Electric Flux - *CC BY 4.0*
  - 17.3: Gauss's Law - *CC BY 4.0*
  - 17.4: Calculating Electric Field Using Gauss's Law - *CC BY 4.0*
  - 17.5: Conductors in Electrostatic Equilibrium via Gauss's Law - *CC BY 4.0*
  - 17.6: Gauss's Law (Summary) - *CC BY 4.0*
  - 17.7: Gauss's Law (Exercises) - *CC BY 4.0*
  - 17.8: Gauss's Law (Answers) - *CC BY 4.0*
- 18: Calculation of Magnetic Quantities from Currents - *Undeclared*
  - 18.1: Introduction - *CC BY 4.0*
  - 18.2: Magnetic Field due to a Thin Straight Wire - *CC BY 4.0*
  - 18.3: Magnetic Field of a Current Loop - *CC BY 4.0*
  - 18.4: Magnetic Field using Ampère's Law - *CC BY 4.0*
  - 18.5: Magnetic Field of Solenoids and Toroids - *CC BY 4.0*
  - 18.6: Magnetic Force between Two Parallel Currents - *CC BY 4.0*
  - 18.7: (edit) Magnetic Force and Torque on a Current Loop - Motors and Meters - *CC BY 4.0*
  - 18.8: Magnetic Forces in a Conductor - The Hall Effect - *CC BY 4.0*
  - 18.9: More Applications of Magnetism - *CC BY 4.0*
  - 18.10: Superconductors - *CC BY 4.0*
  - 18.11: Conclusion - *Undeclared*
  - 18.12: Magnetic Forces and Fields (Summary) - *CC BY 4.0*
  - 18.13: Sources of Magnetic Fields (Summary) - *CC BY 4.0*
  - 18.14: Current and Resistance (Summary) - *CC BY 4.0*
  - 18.15: Magnetic Forces and Fields (Exercise) - *CC BY 4.0*
  - 18.16: Sources of Magnetic Fields (Exercise) - *CC BY 4.0*

Detailed Licensing

Overview

Title: Electricity and Magnetism with Applications to Amateur Radio and Wireless Technology

Webpages: 271

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY 4.0: 63.8% (173 pages)
- CC BY-SA 4.0: 14% (38 pages)
- Undeclared: 13.7% (37 pages)
- CC BY-NC 4.0: 4.8% (13 pages)
- GNU Free Documentation License: 2.6% (7 pages)
- CC BY 1.0: 1.1% (3 pages)

By Page

- Electricity and Magnetism with Applications to Amateur Radio and Wireless Technology - *CC BY-SA 4.0*
    - Front Matter - *Undeclared*
        - TitlePage - *Undeclared*
        - InfoPage - *Undeclared*
        - Table of Contents - *CC BY 4.0*
        - Licensing - *Undeclared*
        - Preface - *Undeclared*
    - 1: Preliminary Concepts - *CC BY 4.0*
        - 1.1: What is Electricity and Magnetism? - *CC BY-SA 4.0*
        - 1.2: Wireless Technology and Amateur Radio - What and Why? - *CC BY-SA 4.0*
        - 1.3: Units - *CC BY-SA 4.0*
        - 1.4: Electromagnetic Spectrum - *CC BY-SA 4.0*
        - 1.5: Amateur Radio Equipment Basics - *CC BY-SA 4.0*
        - 1.6: Notation - *CC BY-SA 4.0*
        - 1.7: Coordinate Systems - *CC BY-SA 4.0*
        - 1.8: Where Do We Go from Here? - *CC BY-SA 4.0*
    - 2: The Electric Field - *CC BY 4.0*
        - 2.1: Introduction - *CC BY-SA 4.0*
        - 2.2: Electric Charge Model - *CC BY 4.0*
        - 2.3: Conduction and Charging - *CC BY 4.0*
        - 2.4: Electric Fields and Forces - *CC BY-SA 4.0*
        - 2.5: Electric Fields and Forces with Multiple Charges - *CC BY 4.0*
        - 2.6: Electric Field Diagrams - *CC BY 4.0*
        - 2.7: Common Models of Electric Field - *CC BY 4.0*
        - 2.8: Motion of a Charged Particle in an Electric Field - *CC BY 4.0*
        - 2.9: Conclusion - *CC BY-SA 4.0*
        - 2.10: The Electric Field (Summary) - *CC BY 4.0*
        - 2.11: The Electric Field (Exercises) - *CC BY 4.0*
        - 2.12: The Electric Field (Answers) - *CC BY 4.0*
    - 3: The Electric Potential - *Undeclared*
        - 3.1: Introduction - *CC BY 4.0*
        - 3.2: Work and Energy - *CC BY 4.0*
        - 3.3: Electric Potential Energy - *CC BY 4.0*
        - 3.4: Electric Potential Energy of Point Charges - *CC BY 4.0*
        - 3.5: Electric Potential - *CC BY 4.0*
        - 3.6: Electric Potential of a Point Charge - *CC BY 4.0*
        - 3.7: Common Models of Electric Potential - *CC BY 4.0*
        - 3.8: Electric Potential (Summary) - *CC BY 4.0*
        - 3.9: The Electric Potential (Exercises) - *CC BY 4.0*
        - 3.10: The Electric Potential (Answers) - *CC BY 4.0*
    - 4: Potential and Field Relationships - *Undeclared*
        - 4.1: Electric Potential from Electric Field - *CC BY 4.0*
        - 4.2: Electric Field from Electric Potential - *CC BY 4.0*
        - 4.3: Equipotential Curves and Surfaces - *CC BY 4.0*
        - 4.4: Conductors in Electrostatic Equilibrium - *CC BY 4.0*
        - 4.5: Applications of Electric Potential and Conductors in Electrostatic Equilibrium - *CC BY 4.0*
        - 4.6: Potential and Field Relationships (Summary) - *CC BY 4.0*
        - 4.7: Potential and Field Relationships (Exercises) - *CC BY 4.0*
        - 4.8: Potential and Field Relationships (Answers) - *CC BY 4.0*
    - 5: Electric Current and Resistance - *Undeclared*
        - 5.1: Introduction - *CC BY 4.0*
        - 5.2: Electric Current - *CC BY 4.0*

- 18.17: Magnetic Forces and Fields (Answers) - *CC BY 4.0*
- 18.18: Sources of Magnetic Fields (Answers) - *CC BY 4.0*
- 19: Alternating-Current (AC) Circuits - *Undeclared*
  - 19.1: Introduction - *CC BY 4.0*
  - 19.2: AC Sources - *CC BY 4.0*
  - 19.3: Simple AC Circuits - *CC BY 4.0*
  - 19.4: RLC Series Circuits with AC - *CC BY 4.0*
  - 19.5: Power in an AC Circuit - *CC BY 4.0*
  - 19.6: Resonance in an AC Circuit - *CC BY 4.0*
  - 19.7: AC Safety - Grounding and Bonding - *Undeclared*
  - 19.8: Alternating-Current Circuits (Summary) - *CC BY 4.0*
  - 19.9: Alternating-Current Circuits (Exercise) - *CC BY 4.0*
  - 19.10: Alternating-Current Circuits (Answers) - *CC BY 4.0*
- 20: Maxwell's Equations - *Undeclared*
  - 20.1: Introduction - *CC BY 4.0*
  - 20.2: Electric Flux - *CC BY 4.0*
  - 20.3: Gauss's Law - *CC BY 4.0*
  - 20.4: Ampère's Law - *CC BY 4.0*
  - 20.5: Maxwell's Equations and Electromagnetic Waves - *CC BY 4.0*
  - 20.6: Plane Electromagnetic Waves - *CC BY 4.0*
  - 20.7: Momentum and Radiation Pressure - *CC BY 4.0*
- 21: Electrical Transmission Lines - *Undeclared*
  - 21.1: Introduction - *CC BY 4.0*
  - 21.2: Phasors - *CC BY-SA 4.0*
  - 21.3: Introduction to Transmission Lines - *CC BY-SA 4.0*
  - 21.4: Types of Transmission Lines - *CC BY-SA 4.0*
  - 21.5: Transmission Lines as Two-Port Devices - *CC BY-SA 4.0*
  - 21.6: Lumped-Element Model - *CC BY-SA 4.0*
  - 21.7: Telegrapher's Equations - *CC BY-SA 4.0*
  - 21.8: Wave Equation for a Transmission Line - *CC BY-SA 4.0*
  - 21.9: Characteristic Impedance of a Transmission Line - *CC BY-SA 4.0*
  - 21.10: Wave Propagation on a Transmission Line - *CC BY-SA 4.0*
  - 21.11: Lossless and Low-Loss Transmission Lines - *CC BY-SA 4.0*
  - 21.12: Voltage Reflection Coefficient - *CC BY-SA 4.0*
  - 21.13: Standing Waves - *CC BY-SA 4.0*
  - 21.14: Standing Wave Ratio - *CC BY-SA 4.0*
  - 21.15: Parallel Wire Transmission Line - *CC BY-SA 4.0*
  - 21.16: Attenuation in Coaxial Cable - *CC BY-SA 4.0*
  - 21.17: Power Handling Capability of Coaxial Cable - *CC BY-SA 4.0*
  - 21.18: Why 50 Ohms? - *CC BY-SA 4.0*
  - 21.19: Conclusion - *Undeclared*
- 22: Generation and Detection of Electromagnetic Waves - *Undeclared*
  - 22.1: Introduction - *CC BY 4.0*
  - 22.2: Production of Electromagnetic Waves - The Antenna - *CC BY 4.0*
  - 22.3: Radiation from a Current Moment - *CC BY-SA 4.0*
  - 22.4: Radiation from an Electrically-Short Dipole - *CC BY-SA 4.0*
  - 22.5: Far-Field Radiation from a Half-Wave Dipole - *CC BY-SA 4.0*
  - 22.6: Equivalent Circuit Model for Transmission; Radiation Efficiency - *CC BY-SA 4.0*
  - 22.7: Equivalent Circuit Model for Reception - *CC BY-SA 4.0*
  - 22.8: Potential Induced in a Dipole - *CC BY-SA 4.0*
  - 22.9: Decibel Scale for Power Ratio - *CC BY-SA 4.0*
  - 22.10: Antenna Radiation Patterns, Directivity, and Gain - *CC BY-SA 4.0*
  - 22.11: Friis Transmission Equation - *CC BY-SA 4.0*
- 23: Signal Modulation - *Undeclared*
  - 23.1: Introduction - *CC BY-NC 4.0*
  - 23.2: Historical Context - The Origins of Radio Communication - *CC BY-NC 4.0*
  - 23.3: Radio Signal Metrics - *CC BY-NC 4.0*
  - 23.4: Modulation Overview - *CC BY-NC 4.0*
  - 23.5: Analog Modulation - *CC BY-NC 4.0*
  - 23.6: Digital Modulation - *CC BY-NC 4.0*
  - 23.7: Frequency Shift Keying, FSK - *CC BY-NC 4.0*
  - 23.8: Carrier Recovery - *CC BY-NC 4.0*
  - 23.9: Phase Shift Keying Modulation - *CC BY-NC 4.0*
  - 23.10: Quadrature Amplitude Modulation - *CC BY-NC 4.0*
  - 23.11: Digital Modulation Summary - *CC BY-NC 4.0*
  - 23.12: References - *CC BY-NC 4.0*
  - 23.13: Exercises - *CC BY-NC 4.0*
- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*