

Introductory Physics II:

Waves, Acoustics, Electromagnetism, Optics, and Modern Physics

D.G. Simpson, Ph.D.

Department of Physical Sciences and Engineering
Prince George's Community College
Largo, Maryland

Spring 2018

Last updated: February 24, 2018

Contents

	Acknowledgments	10
	I Preliminaries	11
1	What is Physics?	12
2	Units	14
	2.1 Systems of Units.	14
	2.2 SI Units	15
	2.3 CGS Systems of Units	19
	2.4 British Engineering Units	19
	2.5 Units as an Error-Checking Technique.	19
	2.6 Unit Conversions	20
	2.7 Currency Units.	21
	2.8 Odds and Ends.	22
3	Problem-Solving Strategies	23
4	The Calculus	25
	4.1 Infinitesimal Numbers	25
	4.2 Differential Calculus — Finding Slopes	26
	4.3 Integral Calculus — Finding Areas	28
	4.4 The Fundamental Theorem of Calculus	32
	4.5 Approximations	32
	4.6 More Examples	33
	4.7 Main Ideas.	37
	4.8 Going Further	37
	II Waves	38
5	Simple Harmonic Motion	39
	5.1 Energy.	41
	5.2 The Vertical Spring	43
	5.3 Frequency and Period	43
	5.4 Mass on a Spring	43
	5.5 More on the Spring Constant.	44

6	Damped Oscillations	47
6.1	Underdamped	47
6.2	Overdamped	47
6.3	Critically Damped	48
7	Forced Oscillations	49
7.1	Resonance	49
8	The Pendulum	51
8.1	Equation of Motion	52
8.2	Period	52
8.3	The Spherical Pendulum.	52
8.4	The Conical Pendulum.	54
8.5	The Torsional Pendulum.	54
8.6	The Physical Pendulum	54
8.7	Other Pendulums	56
9	Waves	57
9.1	Types of Waves	57
9.2	Wave Speed	58
9.3	String Waves.	59
9.4	Reflection and Transmission.	59
9.5	Superposition	61
9.6	Interference	61
9.7	Wave Energy.	62
9.8	Wave Intensity.	64
9.9	Ocean Waves	64
9.10	Seismic Waves.	65
10	Standing Waves	66
10.1	Fixed or Free at Both Ends.	66
10.2	Fixed at One End and Free at the Other	68
10.3	Vibrations of Rods and Plates	68
	III Acoustics	70
11	Sound	71
11.1	Speed of Sound	71
11.2	Frequency of Sound	74
12	The Doppler Effect	75
12.1	Relativistic Doppler Effect.	76
13	Sound Intensity	79
13.1	Intensity	79
13.2	Decibels	79
13.3	Nepers	80

14	Music	81
14.1	Pitch	81
14.2	Musical Scales	82
14.3	Music Notation	84
14.4	Timing.	86
14.5	An Example	86
14.6	Musical Instruments	87
	IV Electricity and Magnetism	89
15	Electricity	90
15.1	Electric Charge	90
15.2	Coulomb's Law	91
15.3	Atomic View of Electricity.	91
15.4	Materials.	92
15.5	Coulomb's Law in Two or Three Dimensions	92
16	The Electric Field	94
16.1	Electric Field due to a Point Charge	94
16.2	Electric Field Lines	94
16.3	The Electric Dipole	95
16.4	Electric Flux.	95
16.5	Gauss's Law	96
16.6	Electric Fields of Conductors	96
16.7	Dielectric Breakdown	97
16.8	Lightning	97
17	Electric Potential	98
17.1	Potential Energy	98
17.2	Potential	98
17.3	Equipotential Surfaces	99
17.4	Comparison between Gravity and Electricity.	99
17.5	The Electron Volt	100
18	The Battery	101
19	Electric Current	103
20	Resistance	104
20.1	Resistivity	104
20.2	Resistors in Series and Parallel.	106
20.3	Conductance.	107
20.4	Wire	107
20.5	Battery Internal Resistance.	108
21	Ohm's Law	110
21.1	Electric Power	110

22	DC Electric Circuits	111
22.1	Schematic Diagrams	111
22.2	Kirchhoff Plots.	111
22.3	A Simple Circuit.	111
22.4	Circuit Analysis Principles.	115
23	Kirchhoff's Rules	117
23.1	Example Circuit	117
24	Electronic Instruments	120
24.1	Ammeter.	120
24.2	Voltmeter	120
24.3	Ohmmeter	120
24.4	Multimeter.	120
24.5	Oscilloscope	120
24.6	Logic Probe	121
25	Capacitance	122
25.1	Parallel-Plate Capacitor	122
25.2	Capacitors in Series and Parallel.	123
25.3	Dielectric Materials in Capacitors	124
25.4	Energy Stored in a Capacitor.	124
26	RC Circuits	126
26.1	Charging RC Circuit.	126
26.2	Discharging RC Circuit	127
27	Other Electronic Components	129
27.1	The Diode	129
27.2	The Transistor	129
27.3	Integrated Circuits	129
28	The Incandescent Light Bulb	130
29	Electronics as a Hobby	131
29.1	Analog Electronics.	131
29.2	Digital Electronics	131
29.3	Amateur Radio.	132
29.4	Robotics	133
29.5	Amateur Rocketry	133
29.6	Amateur Satellites	133
29.7	Sample Electronics Projects	134
30	Magnetism	135
30.1	Magnetic Poles	135
30.2	Atomic View of Magnetism	136

31	The Magnetic Field	137
31.1	Magnetic Field	137
31.2	Magnetic Field due to a Single Magnetic Pole	137
31.3	Magnetic Field Lines	137
31.4	The Magnetic Dipole	138
31.5	Magnetic Flux	138
31.6	Gauss's Law for Magnetism	139
31.7	Biot-Savart Law	139
31.8	Magnetic Field due to a Long Wire	139
31.9	Magnetic Field of a Solenoid	140
31.10	Magnetic Field of a Loop or Coil of Wire	140
31.11	Torque on a Magnetic Dipole in a Magnetic Field	140
31.12	Magnetic Pressure	141
32	The Lorentz Force	142
32.1	Plasmas	142
32.2	Force on a Wire in a Magnetic Field	143
32.3	Magnetic Force between Two Long Wires	143
32.4	The Hall Effect	144
33	Geomagnetism	146
33.1	Earth's Magnetic Dipole	146
33.2	Magnetic Declination	147
33.3	Magnetic Inclination.	147
33.4	Magnetic Reversals	149
33.5	The Magnetosphere	149
33.6	The Aurora	149
34	Magnetic Materials	156
34.1	Diamagnetism	156
34.2	Paramagnetism.	156
34.3	Ferromagnetism	156
34.4	Permanent Magnets	157
34.5	Curie Temperature	157
34.6	Eddy Currents	158
35	Ampère's Law	159
36	Faraday's Law	161
36.1	Lenz's Law	161
36.2	Motional EMF	162
37	Inductance	164
37.1	Solenoid Inductor	164
37.2	Inductors in Series and Parallel	165
37.3	Magnetic Materials in Inductors	165
37.4	Energy Stored in an Inductor.	166

38	LR Circuits	167
39	LC and LCR Circuits	169
39.1	LC Circuits	169
39.2	LCR Circuits.	172
40	AC Circuits	173
41	Memristance	175
42	Electromagnetism	176
42.1	Electromagnetic Waves	176
43	Radio	178
43.1	The Ionosphere	180
43.2	The Crystal Radio	181
43.3	The Radio Transmitter	184
	V Optics	187
44	Geometrical Optics	188
44.1	Law of Reflection	188
45	Mirrors	189
45.1	Ray Diagrams	191
45.2	Algebraic Method	192
45.3	Segmented Mirrors	192
46	Refraction	194
46.1	Snell's Law	194
46.2	Total Internal Reflection	194
47	Lenses	196
47.1	Ray Diagrams	197
47.2	Algebraic Method	198
47.3	The Fresnel Lens.	199
48	Optical Defects	200
48.1	Spherical Aberration.	200
48.2	Chromatic Aberration	200
48.3	Astigmatism	200
48.4	Coma	200
49	Optical Instruments	202
49.1	The Magnifying Glass	202
49.2	The Human Eye	202
49.3	The Trilobite Eye	204
49.4	The Camera	206
49.5	The Microscope	206

49.6	The Telescope	207
49.7	The Periscope	208
49.8	The Kaleidoscope	210
50	Photometry	211
50.1	Luminous Flux.	211
50.2	Luminous Intensity	212
50.3	Illuminance	212
50.4	Example: The Sun	213
50.5	Example: Incandescent Light Bulb	214
50.6	Astronomical Photometry	214
51	Young's Experiment	216
51.1	Quantum Effects	216
52	Diffraction	218
52.1	The Rayleigh Criterion.	218
52.2	Floater's in the Eye	219
52.3	The Diffraction Grating	219
53	Optics of the Hubble Space Telescope	220
53.1	The Hubble Space Telescope.	220
53.2	HST Optics Overview	221
53.3	Resolution	221
53.4	Spherical Aberration.	223
54	Dispersion	224
54.1	Cauchy Dispersion Formula	224
54.2	Sellmeier Dispersion Formula	224
55	Polarization	226
55.1	Selective Absorption.	226
55.2	Reflection; Brewster's Law	226
55.3	Scattering	227
55.4	Birefringence	227
56	Color	228
56.1	Lights	228
56.2	Pigments.	230
56.3	Spectral Colors	230
56.4	The Chromaticity Diagram.	231
57	The Rainbow	234
57.1	Colors	234
57.2	The Primary Rainbow	234
57.3	The Secondary Rainbow	237
57.4	Location of the Rainbow.	237
57.5	Alexander's Dark Band	238
57.6	Higher-Order Rainbows	239

VI	Modern Physics	243
58	Special Relativity	244
58.1	Introduction	244
58.2	Postulates	244
58.3	Time Dilation	244
58.4	Length Contraction	245
58.5	An Example	245
58.6	Momentum	245
58.7	Addition of Velocities	246
58.8	Energy	246
59	Superfluids	248
60	The Standard Model	251
60.1	Matter	251
60.2	Antimatter	252
60.3	Forces	252
60.4	The Higgs Boson	253
	Appendices	255
A	Further Reading	255
B	Greek Alphabet	259
C	Trigonometry	260
D	Useful Series	264
E	Table of Derivatives	265
F	Table of Integrals	267
G	Mathematical Subtleties	269
H	SI Units	271
I	Gaussian Units	274
J	Units of Physical Quantities	276
K	Physical Constants	279
L	Astronomical Data	280
M	Unit Conversion Tables	281

N	Angular Measure	285
	N.1 Plane Angle	285
	N.2 Solid Angle	285
O	Vector Arithmetic	287
P	Matrix Properties	290
Q	Moments of Inertia	292
R	The Simple Plane Pendulum: Exact Solution	294
	R.1 Equation of Motion	294
	R.2 Solution, $\theta(t)$	295
	R.3 Period	295
S	CIE Chromaticity Coordinates	299
T	Calculator Programs	301
U	Right-Hand Rules	302
V	The Earth's Magnetosphere	303
W	Round-Number Handbook of Physics	306
X	Short Glossary of Particle Physics	308
Y	Fundamental Physical Constants — Extensive Listing	309
Z	Periodic Table of the Elements	316
	References	316
	Index	319

Acknowledgments

The author wishes to express his thanks to David Benning, Jay Nelson, and Glenn Snyder for their help with the material on music in Chapter 14; and also to John McClure of Prince George's Community College for many valuable comments on the notes.

Part I

Preliminaries

Chapter 1

What is Physics?

Physics is the most fundamental of the sciences. Its goal is to learn how the Universe works at the most fundamental level—and to discover the basic laws by which it operates. *Theoretical physics* concentrates on developing the theory and mathematics of these laws, while *applied physics* focuses attention on the application of the principles of physics to practical problems. *Experimental physics* lies at the intersection of physics and engineering; experimental physicists have the theoretical knowledge of theoretical physicists, and they know how to build and work with scientific equipment.

Physics is divided into a number of sub-fields, and physicists are trained to have some expertise in all of them. This variety is what makes physics one of the most interesting of the sciences—and it makes people with physics training very versatile in their ability to do work in many different technical fields.

The major fields of physics are:

- *Classical mechanics* is the study the motion of bodies according to Newton's laws of motion.
- *Electricity and magnetism* are two closely related phenomena that are together considered a single field of physics. We'll study electricity and magnetism in this course.
- *Quantum mechanics* describes the peculiar motion of very small bodies (atomic sizes and smaller).
- *Optics* is the study of light, and we'll study it in this course.
- *Acoustics* is the study of sound; this is another subject we'll study in this course.
- *Thermodynamics* and *statistical mechanics* are closely related fields that study the nature of heat.
- *Solid-state physics* is the study of solids—most often crystalline metals.
- *Plasma physics* is the study of plasmas (ionized gases).
- *Atomic, nuclear, and particle physics* study of the atom, the atomic nucleus, and the particles that make up the atom.
- *Relativity* includes Albert Einstein's theories of special and general relativity. *Special relativity* describes the motion of bodies moving at very high speeds (near the speed of light), while *general relativity* is Einstein's theory of gravity.

The fields of *cross-disciplinary physics* combine physics with other sciences. These include *astrophysics* (physics of astronomy), *geophysics* (physics of geology), *biophysics* (physics of biology), *chemical physics* (physics of chemistry), and *mathematical physics* (mathematical theories related to physics).

Besides acquiring a knowledge of physics for its own sake, the study of physics will give you a broad technical background and set of problem-solving skills that you can apply to wide variety of other fields. Some students of physics go on to study more advanced physics, while others find ways to apply their knowledge of physics to such diverse subjects as mathematics, engineering, biology, medicine, and finance.

Chapter 2

Units

The phenomena of Nature have been found to obey certain physical laws; one of the primary goals of physics research is to discover those laws. It has been known for several centuries that the laws of physics are appropriately expressed in the language of *mathematics*, so physics and mathematics have enjoyed a close connection for quite a long time.

In order to connect the physical world to the mathematical world, we need to make *measurements* of the real world. In making a measurement, we compare a physical quantity with some agreed-upon standard, and determine how many such standard units are present. For example, we have a precise definition of a unit of length called a *mile*, and have determined that there are about 92,000,000 such miles between the Earth and the Sun.

It is important that we have very precise definitions of physical units — not only for scientific use, but also for trade and commerce. In practice, we define a few *base units*, and derive other units from combinations of those base units. For example, if we define units for length and time, then we can define a unit for speed as the length divided by time (e.g. miles/hour).

How many base units do we need to define? There is no magic number; in fact it is possible to define a system of units using only *one* base unit (and this is in fact done for so-called *natural units*). For most systems of units, it is convenient to define base units for length, mass, and time; a base electrical unit may also be defined, along with a few lesser-used base units.

2.1 Systems of Units

Several different systems of units are in common use. For everyday civil use, most of the world uses *metric* units. The United Kingdom uses both metric units and an *imperial* system. Here in the United States, *U.S. customary units* are most common for everyday use.¹

There are actually several “metric” systems in use. They can be broadly grouped into two categories: those that use the meter, kilogram, and second as base units (MKS systems), and those that use the centimeter, gram, and second as base units (CGS systems). There is only one MKS system, called *SI units*. We will mostly use SI units in this course.

¹In the mid-1970s the U.S. government attempted to switch the United States to the metric system, but the idea was abandoned after strong public opposition. One remnant from that era is the two-liter bottle of soda pop.

2.2 SI Units

SI units (which stands for *Système International d'unités*) are based on the *meter* as the base unit of length, the *kilogram* as the base unit of mass, and the *second* as the base unit of time. SI units also define four other base units (the *ampere*, *kelvin*, *candela*, and *mole*, to be described later). Any physical quantity that can be measured can be expressed in terms of these base units or some combination of them. SI units are summarized in Appendix H.

SI units were originally based mostly on the properties of the Earth and of water. Under the *original* definitions:

- The *meter* was defined to be one ten-millionth the distance from the equator to the North Pole, along a line of longitude passing through Paris.
- The *kilogram* was defined as the mass of 0.001 m^3 of water.
- The *second* was defined as $1/86,400$ the length of a day.
- The definition of the *ampere* is related to electrical properties, ultimately relating to the meter, kilogram, and second.
- The *kelvin* was defined in terms of the thermodynamics properties of water, as well as absolute zero.
- The *candela* was defined by the luminous properties of molten tungsten.
- The *mole* is defined by the density of the carbon-12 nucleus.

Many of these original definitions have been replaced over time with more precise definitions, as the need for increased precision has arisen.

Length (Meter)

The SI base unit of length, the *meter* (m), has been re-defined more times than any other unit, due to the need for increasing accuracy. Originally (1793) the meter was defined to be $1/10,000,000$ the distance from the North Pole to the equator, along a line going through Paris.² Then, in 1889, the meter was re-defined to be the distance between two lines engraved on a prototype meter bar kept in Paris. Then in 1960 it was re-defined again: the meter was defined as the distance of 1,650,763.73 wavelengths of the orange-red emission line in the krypton-86 atomic spectrum. Still more stringent accuracy requirements led to the the current definition of the meter, which was implemented in 1983: the meter is now defined to be the distance light in vacuum travels in $1/299,792,458$ second. Because of this definition, the speed of light is now *exactly* 299,792,458 m/s.

U.S. Customary units are legally defined in terms of metric equivalents. For length, the *foot* (ft) is defined to be exactly 0.3048 meter.

Mass (Kilogram)

Originally the *kilogram* (kg) was defined to be the mass of 1 liter (0.001 m^3) of water. The need for more accuracy required the kilogram to be re-defined to be the mass of a standard mass called the *International Prototype Kilogram* (IPK, frequently designated by the Gothic letter $\mathfrak{K}$), which is kept in a vault at the Bureau International des Poids et Mesures (BIPM) in Paris. The kilogram is the only base unit still defined in terms of a prototype, rather than in terms of an experiment that can be duplicated in the laboratory.

²If you remember this original definition, then you can remember the circumference of the Earth: about 40,000,000 meters.

The International Prototype Kilogram is a small cylinder of platinum-iridium alloy (90% platinum), about the size of a golf ball. In 1884, a set of 40 duplicates of the IPK was made; each country that requested one got one of these duplicates. The United States received two of these: the duplicate called K20 arrived here in 1890, and has been the standard of mass for the U.S. ever since. The second copy, called K4, arrived later that same year, and is used as a constancy check on K20. Finally, in 1996 the U.S. got a third standard called K79; this is used for mass stability studies. These duplicates are kept at the National Institutes of Standards and Technology (NIST) in Gaithersburg, Maryland. They are kept under very controlled conditions under several layers of glass bell jars and are periodically cleaned. From time to time they are returned to the BIPM in Paris for re-calibration. For reasons not entirely understood, very careful calibration measurements show that the masses of the duplicates do not stay exactly constant. Because of this, physicists are considering re-defining the kilogram sometime in the next few years.

Another common metric (but non-SI) unit of mass is the *metric ton*, which is 1000 kg (a little over 1 short ton).

In U.S. customary units, the *pound-mass* (lbm) is defined to be exactly 0.45359237 kg.

Mass vs. Weight

Mass is not the same thing as *weight*, so it's important not to confuse the two. The *mass* of a body is a measure of the total amount of matter it contains; the *weight* of a body is the gravitational force on it due to the Earth's gravity. At the surface of the Earth, mass m and weight W are proportional to each other:

$$W = mg, \tag{2.1}$$

where g is the acceleration due to the Earth's gravity, equal to 9.80 m/s^2 . Remember: mass is mass, and is measured in kilograms; weight is a force, and is measured in force units of *newtons*.

Time (Second)

Originally the base SI unit of time, the *second* (s), was defined to be $1/60$ of $1/60$ of $1/24$ of the length of a day, so that 60 seconds = 1 minute, 60 minutes = 1 hour, and 24 hours = 1 day. High-precision time measurements have shown that the Earth's rotation rate has short-term irregularities, along with a long-term slowing due to tidal forces. So for a more accurate definition, in 1967 the second was re-defined to be based on a definition using atomic clocks. The second is now defined to be the time required for 9,192,631,770 oscillations of a certain type of radiation emitted from a cesium-133 atom.

Although officially the symbol for the second is "s", you will also often see people use "sec" to avoid confusing lowercase "s" with the number "5".

The Ampere, Kelvin, and Candela

For this course, most quantities will be defined entirely in terms of meters, kilograms, and seconds. There are four other SI base units, though: the *ampere* (A) (the base unit of electric current); the *kelvin* (K) (the base unit of temperature); the *candela* (cd) (the base unit of luminous intensity, or light brightness); and the *mole* (mol) (the base unit of amount of substance).

Amount of Substance (Mole)

Since we may have a use for the mole in this course, let's look at its definition in detail. The simplest way to think of it is as the name for a number. Just as "thousand" means 1,000, "million" means 1,000,000, and "billion" means 1,000,000,000, in the same way "mole" refers to the number 602,214,129,000,000,000,000,000.

or $6.02214129 \times 10^{23}$. You could have a mole of grains of sand or a mole of Volkswagens, but most often the mole is used to count atoms or molecules. There is a reason this number is particularly useful: since each nucleon (proton and neutron) in an atomic nucleus has an average mass of $1.660538921 \times 10^{-24}$ grams (called an *atomic mass unit*, or amu), then there are $1/(1.660538921 \times 10^{-24})$, or $6.02214129 \times 10^{23}$ nucleons per gram. In other words, one mole of nucleons has a mass of 1 gram. Therefore, if A is the atomic weight of an atom, then A moles of nucleons has a mass of A grams. But A moles of nucleons is the same as 1 mole of atoms, so *one mole of atoms has a mass (in grams) equal to the atomic weight*. In other words,

$$\text{moles of atoms} = \frac{\text{grams}}{\text{atomic weight}} \quad (2.2)$$

Similarly, when counting molecules,

$$\text{moles of molecules} = \frac{\text{grams}}{\text{molecular weight}} \quad (2.3)$$

In short, the mole is useful when you need to convert between the mass of a material and the number of atoms or molecules it contains.

It's important to be clear about what exactly you're counting (atoms or molecules) when using moles. It doesn't really make sense to talk about "a mole of oxygen", any more than it would be to talk about "100 of oxygen". It's either a "mole of oxygen atoms" or a "mole of oxygen molecules".³

Interesting fact: there is about $\frac{1}{2}$ mole of stars in the observable Universe.

SI Derived Units

In addition to the seven base units (m, kg, s, A, K, cd, mol), there are a number of so-called *SI derived units* with special names. We'll introduce these as needed, but a summary of all of them is shown in Appendix H (Table H-2). These are just combinations of base units that occur often enough that it's convenient to give them special names.

Plane Angle (Radian)

One derived SI unit that we will encounter frequently is the SI unit of plane angle. Plane angles are commonly measured in one of two units: *degrees* or *radians*.⁴ You're probably familiar with degrees already: one full circle is 360° , a semicircle is 180° , and a right angle is 90° .

The SI unit of plane angle is the *radian*, which is defined to be that plane angle whose arc length is equal to its radius. This means that a full circle is 2π radians, a semicircle is π radians, and a right angle is $\pi/2$ radians. To convert between degrees and radians, then, we have:

$$\text{degrees} = \text{radians} \times \frac{180}{\pi} \quad (2.4)$$

and

$$\text{radians} = \text{degrees} \times \frac{\pi}{180} \quad (2.5)$$

The easy way to remember these formulæ is to think in terms of units: 180 has units of degrees and π has units of radians, so in the first equation units of radians cancel on the right-hand side to leave degrees, and in the second equation units of degrees cancel on the right-hand side to leave radians.

³Sometimes chemists will refer to a "mole of oxygen" when it's understood whether the oxygen in question is in the atomic (O) or molecular (O₂) state.

⁴A third unit implemented in many calculators is the *grad*: a right angle is 100 grads and a full circle is 400 grads. You may encounter grads in some older literature, such as Laplace's *Mécanique Céleste*. Almost nobody uses grads today, though.

Occasionally you will see a formula that involves a “bare” angle that is not the argument of a trigonometric function like the sine, cosine, or tangent. In such cases it is understood that the angle must be *in radians*. For example, the radius of a circle r , angle θ , and arc length s are related by

$$s = r\theta, \quad (2.6)$$

where it is understood that θ is in radians.

See Appendix N for a further discussion of plane and solid angles.

SI Prefixes

It's often convenient to define both large and small units that measure the same thing. For example, in English units, it's convenient to measure small lengths in inches and large lengths in miles.

In SI units, larger and smaller units are defined in a systematic way by the use of *prefixes* to the SI base or derived units. For example, the base SI unit of length is the meter (m), but small lengths may also be measured in centimeters (cm, 0.01 m), and large lengths may be measured in kilometers (km, 1000 m). Table H-3 in Appendix H shows all the SI prefixes and the powers of 10 they represent. You should *memorize* the powers of 10 for all the SI prefixes in this table.

To use the SI prefixes, simply add the prefix to the front of the name of the SI base or derived unit. The symbol for the prefixed unit is the symbol for the prefix written in front of the symbol for the unit. For example, kilometer (km) = 10^3 meter, microsecond (μs) = 10^{-6} s. But put the prefix on the *gram* (g), *not* the kilogram: for example, 1 microgram (μg) = 10^{-6} g. For historical reasons, the kilogram is the only SI base or derived unit with a prefix.⁵

The Future of SI Units

There is currently a proposal to re-define the basis of SI units, probably starting in 2018. According to the proposal, instead of the seven base units, we would *define* the values of seven fundamental physical constants so that they have fixed, unchanging values—in much the same way that the meter is currently defined so that the speed of light in vacuum is defined to have the value 299,792,458 m/s. The proposed defined constants are shown in Table 2-1.

Table 2-1. Proposed new SI base quantities, defining constants, and definitions. (Here X indicates extra digits that have not yet been determined.) (Ref.: *Phys. Today* **67**, 7, 35 (July 2014).)

Base quantity	Defining constant	Definition
Frequency	$\Delta\nu(^{133}\text{Cs})_{\text{hfs}}$	The unperturbed ground-state hyperfine splitting frequency of the cesium-133 atom is exactly 9,192,631,770 Hz.
Velocity	c	The speed of light in vacuum c is exactly 299,792,458 m/s.
Action	h	The Planck constant h is exactly $6.626X \times 10^{-34}$ J s.
Electric charge	e	The elementary charge e is exactly $1.602X \times 10^{-19}$ C.
Heat capacity	k	The Boltzmann constant k is exactly $1.380X \times 10^{-23}$ J/K.
Amount of substance	N_A	The Avogadro constant N_A is exactly $6.022X \times 10^{23}$ mol ⁻¹ .
Luminous intensity	K_{cd}	The luminous efficacy K_{cd} of monochromatic radiation of frequency 540×10^{12} Hz is exactly 683 lm/W.

⁵Originally, the metric standard of mass was a unit called the *grave* (GRAH-veh), equal to 1000 grams. When the metric system was first established by Louis XVI following the French Revolution, the name *grave* was considered politically incorrect, since it resembled the German word *Graf*, or “Count” — a title of nobility, at a time when titles of nobility were shunned. The *grave* was retained as the unit of mass, but under the more acceptable name *kilogram*. The gram itself was too small to be practical as a mass standard.

2.3 CGS Systems of Units

In some fields of physics (e.g. solid-state physics, plasma physics, and astrophysics), it has been customary to use CGS units rather than SI units, so you may encounter them occasionally. There are several different CGS systems in use: *electrostatic*, *electromagnetic*, *Gaussian*, and *Heaviside-Lorentz* units. These systems differ in how they define their electric and magnetic units. Unlike SI units, none of these CGS systems defines a base electrical unit, so electric and magnetic units are all derived units. The most common of these CGS systems is Gaussian units, which are summarized in Appendix I.

SI prefixes are used with CGS units in the same way they're used with SI units.

2.4 British Engineering Units

Another system of units that is common in some fields of engineering is *British engineering units*. In this system, the base unit of length is the foot (ft), and the base unit of time is the second (s). There is no base unit of mass; instead, one uses a base unit of force called the *pound-force* (lbf). Mass in British engineering units is measured units of *slugs*, where 1 slug has a weight of 32.17404855 lbf.

A related unit of mass (not part of the British engineering system) is called the pound-mass (lbm). At the surface of the Earth, a mass of 1 lbm has a weight of 1 lbf, so sometimes the two are loosely used interchangeably and called the *pound* (lb), as we do every day when we speak of weights in pounds.

SI prefixes are not used in the British engineering system.

2.5 Units as an Error-Checking Technique

Checking units can be used as an important error-checking technique called *dimensional analysis*. If you derive an equation and find that the units don't work out properly, then you can be certain you made a mistake somewhere. If the units are correct, it doesn't necessarily mean your derivation is correct (since you could be off by a factor of 2, for example), but it does give you some confidence that you at least haven't made a units error. So checking units doesn't tell you for certain whether or not you've made a mistake, but it does help.

Here are some basic principles to keep in mind when working with units:

1. Units on both sides of an equation must match.
2. When adding or subtracting two quantities, they must have the same units.
3. Quantities that appear in exponents must be dimensionless.
4. The argument for functions like $\sin$, $\cos$, $\tan$, $\sin^{-1}$, $\cos^{-1}$, $\tan^{-1}$, $\log$, and $\exp$ must be dimensionless.
5. When checking units, radians and steradians can be considered dimensionless.
6. When checking complicated units, it may be useful to break down all derived units into base units (e.g. replace newtons with kg m s^{-2}).

Sometimes it's not clear whether or not the units match on both sides of the equation, for example when both sides involve derived SI units. In that case, it may be useful to break all the derived units down in terms of base SI units (m, kg, s, A, K, mol, cd). Table H-2 in Appendix H shows each of the derived SI units broken down in terms of base SI units.

2.6 Unit Conversions

It is very common to have to work with quantities that are given in units other than the units you'd like to work with. Converting from one set of units to another involves a straightforward, virtually foolproof technique that's very simple to double-check. We'll illustrate the method here with some examples.

Appendix M gives a number of important conversion factors. More conversion factors are available from sources such as the *CRC Handbook of Chemistry and Physics*.

1. Write down the unit conversion factor as a ratio, and fill in the units in the numerator and denominator so that the units cancel out as needed.
2. Now fill in the numbers so that the numerator and denominator contain the same length, time, etc. (This is because you want each factor to be a multiplication by 1, so that you don't change the quantity—only its units.)

Simple Conversions

A simple unit conversion involves only one conversion factor. The method for doing the conversion is best illustrated with an example.

Example. Convert 7 feet to inches.

Solution. First write down the unit conversion factor as a ratio, filling in the units as needed:

$$(7 \text{ ft}) \times \frac{\text{in}}{\text{ft}} \quad (2.7)$$

Notice that the units of feet cancel out, leaving units of inches. The next step is to fill in numbers so that the same length is in the numerator and denominator:

$$(7 \text{ ft}) \times \frac{12 \text{ in}}{1 \text{ ft}} \quad (2.8)$$

Now do the arithmetic:

$$(7 \text{ ft}) \times \frac{12 \text{ in}}{1 \text{ ft}} = 84 \text{ inches.} \quad (2.9)$$

More Complex Conversions

More complex conversions may involve more than one conversion factor. You'll need to think about what conversion factors you know, then put together a chain of them to get to the units you want.

Example. Convert 60 miles per hour to feet per second.

Solution. First, write down a chain of conversion factor ratios, filling in units so that they cancel out correctly:

$$60 \frac{\text{mile}}{\text{hr}} \times \frac{\text{ft}}{\text{mile}} \times \frac{\text{hr}}{\text{sec}} \quad (2.10)$$

Units cancel out to leave ft/sec. Now fill in the numbers, putting the same length in the numerator and denominator in the first factor, and the same time in the numerator and denominator in the second factor:

$$60 \frac{\text{mile}}{\text{hr}} \times \frac{5280 \text{ ft}}{1 \text{ mile}} \times \frac{1 \text{ hr}}{3600 \text{ sec}} \quad (2.11)$$

Finally, do the arithmetic:

$$60 \frac{\text{mile}}{\text{hr}} \times \frac{5280 \text{ ft}}{1 \text{ mile}} \times \frac{1 \text{ hr}}{3600 \text{ sec}} = 88 \frac{\text{ft}}{\text{sec}} \quad (2.12)$$

Example. Convert 250,000 furlongs per fortnight to meters per second.

Solution. We don't know how to convert furlongs per fortnight directly to meters per second, so we'll have to come up with a chain of conversion factors to do the conversion. We *do* know how to convert: furlongs to miles, miles to kilometers, kilometers to meters, fortnights to weeks, weeks to days, days to hours, hours to minutes, and minutes to seconds. So we start by writing conversion factor ratios, putting units where they need to be so that the result will have the desired target units (m/s):

$$250,000 \frac{\text{furlong}}{\text{fortnight}} \times \frac{\text{mile}}{\text{furlong}} \times \frac{\text{km}}{\text{mile}} \times \frac{\text{m}}{\text{km}} \times \frac{\text{fortnight}}{\text{week}} \times \frac{\text{week}}{\text{day}} \times \frac{\text{day}}{\text{hr}} \times \frac{\text{hr}}{\text{min}} \times \frac{\text{min}}{\text{sec}}$$

If you check the units here, you'll see that almost everything cancels out; the only units left are m/s, which is what we want to convert to. Now fill in the numbers: we want to put either the same length or the same time in both the numerator and denominator:

$$\begin{aligned} 250,000 \frac{\text{furlong}}{\text{fortnight}} &\times \frac{1 \text{ mile}}{8 \text{ furlongs}} \times \frac{1.609344 \text{ km}}{1 \text{ mile}} \times \frac{1000 \text{ m}}{1 \text{ km}} \times \frac{1 \text{ fortnight}}{2 \text{ weeks}} \times \frac{1 \text{ week}}{7 \text{ days}} \times \frac{1 \text{ day}}{24 \text{ hr}} \times \frac{1 \text{ hr}}{60 \text{ min}} \times \frac{1 \text{ min}}{60 \text{ sec}} \\ &= 41.58 \text{ m/s} \end{aligned}$$

Conversions Involving Powers

Occasionally we need to do something like convert an area or volume when we know only the length conversion factor.

Example. Convert 2000 cubic feet to gallons.

Solution. Let's think about what conversion factors we know. We know the conversion factor between gallons and cubic inches. We don't know the conversion factor between cubic feet and cubic inches, but we can convert between feet and inches. The conversion factors will look like this:

$$2000 \text{ ft}^3 \times \left(\frac{\text{in}}{\text{ft}} \right)^3 \times \frac{\text{gal}}{\text{in}^3} \quad (2.13)$$

With these units, the whole expression reduces to units of gallons. Now fill in the same length in the numerator and denominator of the first factor, and the same volume in the numerator and denominator of the second factor:

$$2000 \text{ ft}^3 \times \left(\frac{12 \text{ in}}{1 \text{ ft}} \right)^3 \times \frac{1 \text{ gal}}{231 \text{ in}^3} \quad (2.14)$$

Now do the arithmetic:

$$2000 \text{ ft}^3 \times \left(\frac{12 \text{ in}}{1 \text{ ft}} \right)^3 \times \frac{1 \text{ gal}}{231 \text{ in}^3} = 14,961 \text{ gallons} \quad (2.15)$$

2.7 Currency Units

Money has units that can be treated like any other units, using the same techniques we've just seen. Two things are unique about units of currency:

- Each country has its own currency units. Examples are United States dollars (\$), British pounds sterling (£), European euros (€), and Japanese yen (¥).
- The conversion factors from one country's currency to another's is a function of time, and even varies minute to minute during the day. These conversion factors are called *exchange rates*, and may be found, for example, on the Internet at <http://www.xe.com/currencyconverter/>.

Example. You're shopping in Reykjavík, Iceland, and see an Icelandic wool scarf you'd like to buy. The price tag says 6990 kr. What is the price in U.S. dollars?

Solution. The unit of currency in Iceland is the Icelandic króna (kr). Looking up the exchange rate on the Internet, you find it is currently \$1 = 119.050 kr. Then

$$6990 \text{ kr.} \times \frac{\$1.00}{119.050 \text{ kr.}} = \$58.71 \quad (2.16)$$

2.8 Odds and Ends

We'll end this chapter with a few miscellaneous notes about SI units:

- In a few special cases, we customarily drop the ending vowel of a prefix when combining with a unit that begins with a vowel: it's *megohm* (not "megaohm"); *kilohm* (not "kiloohm"); and *hectare* (not "hectoare"). In all other cases, keep both vowels (e.g. *microohm*, *kiloare*, etc.). There's no particular reason for this—it's just customary.
- In pharmacology (on bottles of vitamins or prescription medicine, for example), it is usual to indicate micrograms with "mcg" rather than " μg ". While this is technically incorrect, it is done to avoid misreading the units. Using "mc" for "micro" is not done outside pharmacology, and you should not use it in physics. Always use μ for "micro".
- Sometimes in electronics work the SI prefix symbol may be used in place of the decimal point. For example, 24.9 M Ω may be written "24M9". This saves space on electronic diagrams and when printing values on electronic components, and also avoids problems with the decimal point being nearly invisible when the print is tiny. This is unofficial use, and is only encountered in electronics.
- One sometimes encounters older metric units of length called the *micron* (μ , now properly called the *micrometer*, 10^{-6} meter) and the *millimicron* ($\text{m}\mu$, now properly called the *nanometer*, 10^{-9} meter). The micron and millimicron are now obsolete.
- At one time there was a metric prefix *myria-* (my) that meant 10^4 . This prefix is obsolete and is no longer used.
- In computer work, the SI prefixes are often used with units of bytes, but may refer to powers of 2 that are near the SI values. For example, the term "1 kB" may mean 1000 bytes, or it may mean $2^{10} = 1024$ bytes. Similarly, a 100 GB hard drive may have a capacity of 100,000,000,000 bytes, or it may mean $100 \times 2^{30} = 107,374,182,400$ bytes. To help resolve these ambiguities, a set of *binary prefixes* has been introduced (Table H-4 of Appendix H). These prefixes have not yet entirely caught on in the computing industry, though.

Chapter 3

Problem-Solving Strategies

Much of this course will focus on developing your ability to solve physics problems. If you enjoy solving puzzles, you'll find solving physics problems is similar in many ways. Here we'll look at a few general tips on how to approach solving problems.

- At the beginning of a problem stated in SI units, immediately convert the units of all the quantities you're given to base SI units. In other words, convert all lengths to meters, all masses to kilograms, all times to seconds, etc.: all quantities should be in un-prefixed SI units, except for masses in kilograms. When you do this, you're guaranteed that the final result will also be in base SI units, and this will minimize your problems with units. As you gain more experience in problem solving, you'll sometimes see shortcuts that let you get around this suggestion, but for now converting all units to base SI units is the safest approach.
- Similarly, if the problem is stated in CGS units immediately convert all given quantities to base CGS units (lengths in centimeters, masses in grams, and times in seconds). If the problem is stated in British engineering units, immediately convert all given quantities to base units (lengths in feet, masses in slugs, and times in seconds).
- Look at the information you're given, and what you're being asked to find. Then think about what equations you know that might let you get from what you're given to what you're trying to find.
- Be sure you understand under what conditions each equation is valid. For example, it would be inappropriate to use the equations for constant acceleration from kinematics (e.g. $x(t) = \frac{1}{2}at^2 + v_0t + x_0$) for a mass on a spring, since the acceleration of a mass under a spring force is *not* constant. For each equation you're using, you should be clear what each variable represents, and under what conditions the equation is valid.
- As a general rule, it's best to derive an algebraic expression for the solution to a problem first, then substitute numbers to compute a numerical answer as the very last step. This approach has a number of advantages: it allows you to check units in your algebraic expression, helps minimize roundoff error, and allows you to easily repeat the calculation for different numbers if needed.
- If you've derived an algebraic equation, *check the units* of your answer. Make sure your equation has the correct units, and doesn't do something like add quantities with different units.
- If you've derived an algebraic equation, you can check that it has the proper behavior for extreme values of the variables. For example, does the answer make sense if time $t \rightarrow \infty$? If the equation contains an angle, does it reduce to a sensible answer when the angle is 0° or 90° ?

- Check your answer for reasonableness—don't just write down whatever your calculator says. For example, suppose you're computing the speed of a pendulum bob in the laboratory, and find the answer is 14,000 miles per hour. That doesn't seem reasonable, so you should go back and check your work.
- You can avoid rounding errors by carrying as many significant digits as possible throughout your calculations; don't round off until you get to the final result.
- Write down a reasonable number of significant digits in the final answer—don't write down all the digits in your calculator's display. Nor should you round too much and use too few significant digits. There are rules for determining the correct number of significant digits, but for most problems in this course, 3 or 4 significant digits will be about right.
- Don't forget to put the correct units on the final answer! You will have points deducted for forgetting to do this.
- The best way to get good at problem solving (and to prepare for exams for this course) is *practice*—practice working as many problems as you have time for. Working physics problems is a skill much like learning to play a sport or musical instrument. You can't learn by watching someone else do it—you can only learn it by doing it yourself.

Chapter 4

The Calculus

Some ideas in physics are most naturally expressed in terms of a branch of mathematics called *the calculus of infinitesimals*, or simply *the calculus*. Here we will present a very brief overview of the ideas of the calculus so that the notation will be familiar when we encounter it. For a more complete, rigorous, and in-depth understanding of the calculus, the student is referred to courses on the subject.

4.1 Infinitesimal Numbers

Briefly stated, *the calculus is the mathematics of infinitesimal numbers*. Infinitesimal numbers are an extension to the set of real numbers. Following Leibniz, we will call an infinitesimal number on the number line (the x axis) by the notation dx . The symbol dx is to be thought of as one symbol; it does *not* mean $d \times x$.

Here's another way to think of the infinitesimal number dx . You've probably encountered the " Δ " notation before, meaning the difference between two real numbers. For example, if $x_1 = 3$ and $x_2 = 7$, then $\Delta x = x_2 - x_1 = 7 - 3 = 4$ is their difference. The notation dx is analogous to Δx , but refers to the difference between two numbers that are "infinitely close together."

Mathematically, we define the infinitesimal number dx by

$$\exists dx : 0 < dx < x, \forall x \in \mathbb{R} \quad (4.1)$$

In other words, *the (positive) infinitesimal number dx is greater than zero, but smaller than any real number*. You may wonder how this is possible. The answer is: it's just defined this way. Mathematicians have determined that infinitesimal numbers can be defined this way without mathematical contradiction.

Intuitively, you can think of the infinitesimal number dx as being "infinitely close" to zero, but *not* zero. Think of dx as a *very, very, very, very* small number — an "infinitely small" number.

Infinitesimal numbers obey many of the expected laws of arithmetic. Addition and subtraction work as you would expect:

$$dx + dx = 2dx \quad (4.2)$$

$$2dx + dx = 3dx \quad (4.3)$$

$$3dx - dx = 2dx \quad (4.4)$$

Multiplication is also defined:

$$dx \times dx = (dx)^2 \quad (4.5)$$

The number $(dx)^2$ is also an infinitesimal number, but is "infinitely smaller" than dx . This is as expected: if we approximate dx by a very small number like 10^{-6} , then its square (10^{-12}) is much smaller in comparison.

Division of infinitesimals leads to some interesting results. In general, dividing one infinitesimal number by another often leads to a *finite* result, as we'll see in the next section.

4.2 Differential Calculus — Finding Slopes

One important application of the calculus is that it allows us to determine the slope of a line that is not necessarily a straight line. You've learned in an algebra class how to find the slope of a straight line:

$$\text{slope} = \frac{\text{rise}}{\text{run}} \quad (4.6)$$

In other words, pick any two points along the line, and take the change in y (Δy , the “rise”) divided by the change in x (Δx , the “run”).

How can you calculate the slope of a line that is *not* straight — say, for example, the parabola $y = x^2$? For a curved line, the slope is different at different points along the curve; it is defined to be the slope of the straight line tangent to the curve at that point. We can calculate the slope of that tangent line by using the calculus.

As an example, let's take the parabola $f(x) = x^2$ and say we wish to find its slope at $x = 3$. We can approximate the slope of the tangent line at $x = 3$ by finding the slope of the straight line connecting the point on the parabola at $x = 3$ and a second point very close to $x = 3$. The closer the second point is to $x = 3$, the better the approximation to the actual slope at $x = 3$. For example, let the two points be $x = 3$ and $x = 3.01$. Then at $x = 3$, $y = f(x) = x^2 = 3^2 = 9$, and at $x = 3.01$, $y = f(x) = x^2 = 3.01^2 = 9.0601$. The slope of the line connecting these points is then

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{9.0601 - 9}{3.01 - 3} = 6.01 \quad (4.7)$$

Now let's try an even closer second point: $x = 3.001$. Then $y = x^2 = 3.001^2 = 9.006001$. Then

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{9.006001 - 9}{3.001 - 3} = 6.001 \quad (4.8)$$

And yet an even closer second point: $x = 3.0001$. Then $y = x^2 = 3.0001^2 = 9.00060001$. Then

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{9.00060001 - 9}{3.0001 - 3} = 6.0001 \quad (4.9)$$

The closer the second point is to 3, the closer the slope seems to be getting to 6. In other words, in the *limit* where Δx gets closer and closer to 0, the slope gets closer and closer to 6 — suggesting that the slope at $x = 3$ is *exactly* 6. We write this limit as:

$$\text{slope} = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{(x + \Delta x) - x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (4.10)$$

Since $f(x) = x^2$ in our example,

$$\text{slope} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (4.11)$$

$$= \lim_{\Delta x \rightarrow 0} \frac{(x + \Delta x)^2 - x^2}{\Delta x} \quad (4.12)$$

$$= \lim_{\Delta x \rightarrow 0} \frac{[x^2 + 2x\Delta x + (\Delta x)^2] - x^2}{\Delta x} \quad (4.13)$$

$$= \lim_{\Delta x \rightarrow 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \quad (4.14)$$

Canceling Δx in the numerator and denominator,

$$\text{slope} = \lim_{\Delta x \rightarrow 0} 2x + \Delta x \quad (4.15)$$

and as Δx approaches zero,

$$\text{slope} = 2x \quad (4.16)$$

So for at any point along the curve $f(x) = x^2$, its slope is given by $2x$. At $x = 3$, the slope is $2 \times 3 = 6$, in agreement with our earlier approximations.

The slope is called the *derivative* of $f(x)$ with respect to x . As we have just shown, the derivative of $f(x) = x^2$ with respect to x is $2x$. We indicate the derivative of $y = f(x)$ with respect to x by the notation

$$\frac{dy}{dx} \text{ or } \frac{d}{dx} f(x) \quad (4.17)$$

Thus the derivative can be thought of as the quotient of two infinitesimal numbers, and is defined as

$$\frac{dy}{dx} \equiv \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \quad (4.18)$$

For our example $y = f(x) = x^2$,

$$\frac{dy}{dx} = \frac{d}{dx} x^2 = 2x \quad (4.19)$$

More generally, it can be shown that for any n ,

$$\frac{d}{dx} x^n = nx^{n-1} \quad (4.20)$$

For example,

$$\frac{d}{dx} x^5 = 5x^4 \quad (4.21)$$

Here n need not necessarily be an integer. For example, since $\sqrt{x} = x^{1/2}$, we have

$$\frac{d}{dx} \sqrt{x} = \frac{d}{dx} x^{1/2} = \frac{1}{2} x^{-1/2} = \frac{1}{2\sqrt{x}} \quad (4.22)$$

Similar results can be worked out for many common functions. Appendix E gives a short table of derivatives. In conjunction with this table, we note the following properties (u and v are functions of x , and a is a constant):

$$\frac{d}{dx} (au) = a \frac{du}{dx} \quad (4.23)$$

$$\frac{d}{dx} (u + v) = \frac{du}{dx} + \frac{dv}{dx} \quad (4.24)$$

$$\frac{d}{dx} (u - v) = \frac{du}{dx} - \frac{dv}{dx} \quad (4.25)$$

$$\frac{d}{dx} (uv) = \frac{du}{dx} v + u \frac{dv}{dx} \quad (4.26)$$

$$\frac{d}{dx} \left(\frac{u}{v} \right) = \frac{v(du/dx) - u(dv/dx)}{v^2} \quad (4.27)$$

These results will be proved in a more rigorous calculus course.

Now we know how to find the slope of a line that is non necessarily straight: find a formula for the derivative of the curve, and the slope at any point is the derivative evaluated at that point. Why would we want to find the slope of a curved line? For one thing, a derivative with respect to time is how we describe the rate of change of something. For example, velocity is the rate of change of position, so the velocity of a body is written in terms of the derivative of its position with respect to time: $v = dx/dt$ — so that if you have a function $x(t)$ that gives the position x of a body at any time t , you can take the derivative with respect to t and get a formula that gives the velocity v of the body at any time t . Another use for the derivative is for optimization problems: the tangent at the peak of a curve is equal to zero, so to locate the peak of a curve, we calculate its derivative and set it equal to zero.

Here's an interesting calculus fact: there's one function that is equal to its own derivative. That function is e^x :

$$\frac{d}{dx} e^x = e^x \quad (4.28)$$

Example. Find the derivative of the function $f(x) = 4x^3 + 7x^2 - 5x + 6$ with respect to x , and find the slope of $f(x)$ at $x = 3$.

Solution. Using the above results,

$$\frac{d}{dx} f(x) = \frac{d}{dx} (4x^3 + 7x^2 - 5x + 6) \quad (4.29)$$

$$= \frac{d}{dx}(4x^3) + \frac{d}{dx}(7x^2) - \frac{d}{dx}(5x) + \frac{d}{dx}(6) \quad (4.30)$$

$$= 4 \frac{d}{dx}(x^3) + 7 \frac{d}{dx}(x^2) - 5 \frac{d}{dx}(x) + \frac{d}{dx}(6) \quad (4.31)$$

$$= 4(3x^2) + 7(2x) - 5 + 0 \quad (4.32)$$

$$= 12x^2 + 14x - 5 \quad (4.33)$$

The slope at $x = 3$ is then $12(3)^2 + 14(3) - 5 = 145$.

Example. Locate the peaks of the function $f(x) = 4x^3 + 7x^2 - 5x + 6$.

Solution. The peaks are where the derivative is equal to zero. We found the derivative in the previous example, so set this derivative equal to zero to find the peaks:

$$12x^2 + 14x - 5 = 0 \quad (4.34)$$

By the quadratic formula,

$$x = \frac{-14 \pm \sqrt{14^2 - 4 \times 12 \times (-5)}}{2 \times 12} = \frac{-7 \pm \sqrt{109}}{12} = \{-1.4534, 0.2867\} \quad (4.35)$$

This gives the two values of x at which the peaks are located.

4.3 Integral Calculus — Finding Areas

Besides finding slopes, another application of the calculus is to find the *area* under a curve (i.e. between the curve and the x axis). The area under a *straight* line is easy to find without the calculus: it's just the area of a trapezoid. But under a *curved* line, we use the calculus to compute the area.

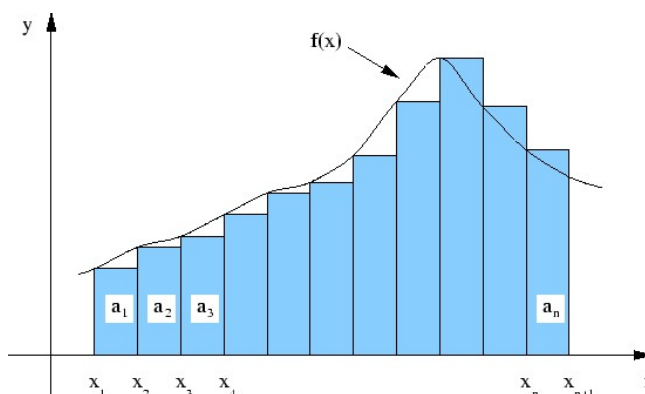


Figure 4.1: Finding the area under a curve using rectangles (*Credit: pleacher.com*)

To do this, imagine dividing the area under the curve into a number of very thin rectangles (Figure 4.1). The thinner the rectangles, the more rectangles we have, and the better the approximation to the actual area under the curve.

If we go to the limit where the rectangles are infinitesimally narrow, then we will have infinitely many of them, and the sum of the areas of all the rectangles exactly equals the area under the curve. Adding up an infinite number of infinitesimal numbers is called *integration*, and typically results in a finite result. If we have a curve $f(x)$, then a rectangle at x has infinitesimal width dx and finite height $f(x)$, so that that rectangle has area equal to its width times its height, or $f(x) dx$. We add together an infinite number of them by integration; the symbol for which is an elongated S (for “sum”), $\int$:

$$\int f(x) dx \quad (4.36)$$

This expression is called an *integral*, and the function $f(x)$ is called the *integrand* of the integral. The area under the curve clearly depends on where the left and right ends of the area are. The area under the curve $f(x)$ between $x = a$ and $x = b$ is indicated by

$$\int_a^b f(x) dx \quad (4.37)$$

Equation (4.36) is called an *indefinite integral*, and Equation (4.37) is called a *definite integral*. To compute a definite integral, we evaluate the *indefinite integral* at the upper bound b , and subtract the indefinite integral evaluated at the lower bound a :

$$\int_a^b f(x) dx = \int f(x) dx \text{ (at } x = b) - \int f(x) dx \text{ (at } x = a) \quad (4.38)$$

For example, suppose we want to find the area under the parabola $f(x) = x^2$ between $x = 1$ and $x = 3$. This would be

$$\text{area} = \int_1^3 x^2 dx = \left(\frac{x^3}{3} \right) \Big|_1^3 = \frac{3^3}{3} - \frac{1^3}{3} = \frac{26}{3} \text{ square units} \quad (4.39)$$

The vertical bar is used to indicate that we evaluate the expression at the top value (3), then subtract the expression evaluated at the bottom value (1).

It is important to note that the area under the curve counts as *negative* area if it lies below the x axis. For example, consider a sine curve, $f(x) = \sin x$. The function $\sin x$ has a positive “lobe” above the x axis from $x = 0$ to $x = \pi$, and a negative “lobe” beneath the x axis from $x = \pi$ to $x = 2\pi$. If we find the integral of $f(x) = \sin x$ from $x = 0$ to $x = 2\pi$, we’re finding the total area under the curve, but counting the part below the x axis as *negative*. We get (using Appendix F):

$$\int_0^{2\pi} \sin x \, dx = (-\cos x) \Big|_0^{2\pi} = -\cos 2\pi - (-\cos 0) = -1 - (-1) = 0. \quad (4.40)$$

so the positive area of the first lobe is exactly cancelled by the negative area of the second lobe, and the total area under the curve is zero. If we really wanted to find the total area under the sine curve from $x = 0$ to $x = 2\pi$, counting all area as positive, we could find the area under just one positive lobe and double it:

$$\text{area} = 2 \int_0^{\pi} \sin x \, dx = 2(-\cos x) \Big|_0^{\pi} = 2[(-\cos \pi) - (-\cos 0)] = 2[1 - (-1)] = 2 \times 2 = 4 \text{ sq. units} \quad (4.41)$$

The area under each lobe is 2 square units.

An unexpected result from the calculus is that the derivative (slope) and integration (area) are *inverse* operations of each other:

$$\frac{d}{dx} \int f(x) \, dx = f(x) \quad (4.42)$$

so the integral can be thought of as the “anti-derivative.” This result is called the *fundamental theorem of calculus*.

In a rigorous calculus course, you will learn how to work out formulas for a number of simple functions. For example,

$$\int x^2 \, dx = \frac{x^3}{3} + C \quad (4.43)$$

where C is an arbitrary constant. All *indefinite* integrals will include this arbitrary constant, because when we take the inverse (a derivative), the derivative of this constant is zero. In effect, some information about the original function is lost when computing its derivative, so that you can’t entirely recover the original function when computing the integral of the derivative. This lost information is expressed as an arbitrary constant C added to the indefinite integral. To find what C is, we would need some additional information, such as what value the integral is supposed to have at a specific point.

More generally,

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} + C \quad (4.44)$$

As with the similar formula for derivatives, n need not be an integer. For example, since $\sqrt{x} = x^{1/2}$, we have

$$\int \sqrt{x} \, dx = \int x^{1/2} \, dx = \frac{x^{3/2}}{3/2} + C = \frac{2}{3} \sqrt{x^3} + C \quad (4.45)$$

Similar results can be worked out for many common functions. Appendix F gives a short table of integrals. In conjunction with this table, we note the following properties (u and v are functions of x , and a is a

constant):

$$\int au \, dx = a \int u \, dx \quad (4.46)$$

$$\int (u + v) \, dx = \int u \, dx + \int v \, dx \quad (4.47)$$

$$\int (u - v) \, dx = \int u \, dx - \int v \, dx \quad (4.48)$$

These results will be proved in a more rigorous calculus course. There are no product or quotient rules for integrals as there are for derivatives.

Since the derivative and integration are inverses of each other, and the function e^x is equal to its own derivative, it is also equal to its own integral (to within an arbitrary constant of integration):

$$\int e^x \, dx = e^x + C \quad (4.49)$$

Example. Find the indefinite integral of the function $f(x) = 4x^3 + 7x^2 - 5x + 6$ with respect to x , and find the area under $f(x)$ between $x = 3$ and $x = 4$.

Solution. Using the above results,

$$\int f(x) \, dx = \int (4x^3 + 7x^2 - 5x + 6) \, dx \quad (4.50)$$

$$= \int 4x^3 \, dx + \int 7x^2 \, dx - \int 5x \, dx + \int 6 \, dx \quad (4.51)$$

$$= 4 \int x^3 \, dx + 7 \int x^2 \, dx - 5 \int x \, dx + 6 \int dx \quad (4.52)$$

$$= 4 \left(\frac{x^4}{4} \right) + C_1 + 7 \left(\frac{x^3}{3} \right) + C_2 - 5 \left(\frac{x^2}{2} \right) + C_3 + 6(x) + C_4 \quad (4.53)$$

$$= x^4 + \frac{7}{3}x^3 - \frac{5}{2}x^2 + 6x + C \quad (4.54)$$

where we have combined all the individual constants of integration C_1, C_2, C_3, C_4 into a single constant C .

To find the area under the curve between $x = 3$ and $x = 4$, we compute the definite integral

$$\int_3^4 f(x) \, dx \quad (4.55)$$

We've already found the indefinite integral; all we need to do is evaluate the indefinite integral at $x = 4$, and subtract the indefinite integral evaluated at $x = 3$:

$$\text{area} = \int_3^4 f(x) \, dx = \left(x^4 + \frac{7}{3}x^3 - \frac{5}{2}x^2 + 6x + C \right) \Big|_3^4 \quad (4.56)$$

$$= \left[(4)^4 + \frac{7}{3}(4)^3 - \frac{5}{2}(4)^2 + 6(4) + C \right] - \left[(3)^4 + \frac{7}{3}(3)^3 - \frac{5}{2}(3)^2 + 6(3) + C \right] \quad (4.57)$$

$$= \frac{1499}{6} \quad (4.58)$$

Notice that the constant of integration C always cancels out in a definite integral.

4.4 The Fundamental Theorem of Calculus

The *fundamental theorem of calculus* states an unexpected result: the derivative (slope-finding) and integral (area-finding) are inverses of each other. Thus

$$\frac{d}{dx} \int f(x) dx = f(x) \quad (4.59)$$

4.5 Approximations

It may sometimes happen that we have *data points* for which we need to calculate a derivative or integral. For example, suppose we have the following data for a moving body:

Time t (s)	Position x (m)
0.0	0.0
1.0	0.34
2.0	1.36
3.0	3.06
4.0	5.44
5.0	8.50
6.0	12.24

What is the velocity v of the body at time $t = 2.5$ seconds? By definition, the velocity v is found by a derivative: $v = dx/dt$. One way to *approximate* this derivative is by finding $\Delta x / \Delta t$, for the interval from 2.0 to 3.0 seconds,

$$\frac{dx}{dt} \approx \frac{\Delta x}{\Delta t} = \frac{3.06 \text{ m} - 1.36 \text{ m}}{3.0 \text{ s} - 2.0 \text{ s}} = 1.70 \text{ m/s} \quad (4.60)$$

We could do the same for every time interval in the table, and use the midpoint of the time intervals as the time. We get the following table:

Time t (s)	Velocity v (m/s)
0.5	0.34
1.5	1.02
2.5	1.70
3.5	2.38
4.5	3.06
5.5	3.74

If the data in the table is “noisy” (has lots of measurement errors), then this kind of computing derivatives numerically can lead to very noisy results: small measurement errors can lead to a large change in slope from one point to the next.

Integrals can be computed numerically as well. There are a number of methods for doing this; the simplest is called the *rectangular rule*, in which we imagine drawing a rectangle at each data point, and approximate

the integral as the sum of the rectangle areas. For example, for the body we've been using for our example, how far does the body travel from time $t = 0$ to time $t = 6$ seconds? That can be found as an integral:

$$x = \int_0^6 v(t) dt \approx \sum_{t=0}^6 v(t) \Delta t \quad (4.61)$$

Using the data from the above table of velocities,

$$x \approx \sum_{t=0}^6 v(t) \Delta t \quad (4.62)$$

$$= (0.34 \text{ m/s})(1.5 - 0.5 \text{ s}) + (1.02 \text{ m/s})(2.5 - 1.5 \text{ s}) + (1.70 \text{ m/s})(3.5 - 2.5 \text{ s}) \quad (4.63)$$

$$+ (2.38 \text{ m/s})(4.5 - 3.5 \text{ s}) + (3.06 \text{ m/s})(5.5 - 4.5 \text{ s}) + (3.74 \text{ m/s})(6.5 - 5.5 \text{ s}) \quad (4.64)$$

$$= 12.24 \text{ m} \quad (4.65)$$

Numerical integration has a tendency to smooth out noise, so in general it is not as subject to the “noise” problem as numerical derivatives are. When using the rectangular rule, one may evaluate the function at the left edge of the horizontal (e.g. time) interval, at the right, edge, or at the center. There are other, more sophisticated, numerical integration methods that may give better results, such as the trapezoidal rule and Simpson's rule. You'll study these in a more comprehensive calculus course.

4.6 More Examples

Area of a Circle

You learned the formula for the area of a circle in elementary school: $A = \pi R^2$, where R is the radius of the circle. We can use integral calculus to derive this formula. The simplest way to approach this using rectangular coordinates is to find the area of a quarter circle and multiply by 4. Let's say the circle has radius R and center at the origin. Then the equation for the circle is

$$x^2 + y^2 = R^2 \quad (4.66)$$

or

$$y = \pm \sqrt{R^2 - x^2} \quad (4.67)$$

For the quarter circle in the first quadrant, we use only the $+$ sign, which corresponds to the upper semicircle:

$$y = \sqrt{R^2 - x^2} \quad (4.68)$$

as let x go from 0 to R to get the quarter-circle in the first quadrant. The area under this quarter-circle curve is then

$$\int_0^R \sqrt{R^2 - x^2} dx \quad (4.69)$$

This is a fairly complicated integral to work out. Often in cases like this, we consult a published table of integrals¹ to find the result already worked out for us. From a published table of integrals, we find the integral

¹Some well-known tables of integrals are found in the *CRC Standard Mathematical Tables and Formulae*; *Tables of Integrals and Other Mathematical Data* by Dwight; and the massive *Table of Integrals, Series, and Products* by Gradshteyn and Ryzhik.

to be

$$\int_0^R \sqrt{R^2 - x^2} dx = \frac{1}{2} \left[x \sqrt{R^2 - x^2} + R^2 \tan^{-1} \left(\frac{x}{\sqrt{R^2 - x^2}} \right) \right] \Big|_0^R \quad (4.70)$$

$$= \frac{1}{2} \left(\frac{\pi}{2} R^2 - 0 \right) = \frac{\pi}{4} R^2 \quad (4.71)$$

The area of a circle is then 4 times this:

$$A = 4 \times \frac{\pi}{4} R^2 = \pi R^2 \quad (4.72)$$

and we have derived the famous formula $A = \pi R^2$.

It's actually simpler to work this problem in polar coordinates, although it leads to a *double integral*. Imagine a circle of radius R , whose center is at the origin. Now imagine a series of straight lines radiating away from the origin, and concentric circles around the origin, just as you have with polar graph paper. These lines divide the interior of the circle up into a series of little "boxes" with curved edges. If you make lots of lines, these boxes will be very small, and if they're infinitesimally small, you can treat them as rectangles. A general infinitesimal "rectangle" will have one side of length dr , and another of (arc) length $r d\theta$. The infinitesimal area of the little box is then the product of the lengths of the sides, $dA = r dr d\theta$. To get the area of a circle, we just add together the infinitesimal areas of all the little boxes inside the circle by integrating r from 0 to R , and integrating θ from 0 to 2π :

$$\text{area} = \int_0^{2\pi} \int_0^R dA = \int_0^{2\pi} \int_0^R r dr d\theta \quad (4.73)$$

This is called a *double integral*. The way to evaluate it is to evaluate the "inner" integral first, then make the result the integrand for the "outer" integral:

$$\int_0^{2\pi} \int_0^R r dr d\theta = \int_0^{2\pi} \left[\int_0^R r dr \right] d\theta \quad (4.74)$$

$$= \int_0^{2\pi} \left[\frac{r^2}{2} \Big|_0^R \right] d\theta \quad (4.75)$$

$$= \int_0^{2\pi} \left[\frac{R^2}{2} - \frac{0^2}{2} \right] d\theta \quad (4.76)$$

$$= \int_0^{2\pi} \left[\frac{R^2}{2} \right] d\theta \quad (4.77)$$

$$= \frac{R^2}{2} \int_0^{2\pi} d\theta \quad (4.78)$$

where in the last step we moved $R^2/2$ outside the integral because it's a constant. Now evaluate the θ integral:

$$\int_0^{2\pi} \int_0^R r \, dr \, d\theta = \frac{R^2}{2} \int_0^{2\pi} d\theta \quad (4.79)$$

$$= \frac{R^2}{2} \theta \Big|_0^{2\pi} \quad (4.80)$$

$$= \frac{R^2}{2} (2\pi - 0) \quad (4.81)$$

$$= \pi R^2 \quad (4.82)$$

And again we have derived the classical formula for the area of a circle.

Area of a Trapezoid

Suppose we have a trapezoid consisting of a side along the x axis, two parallel vertical sides at $x = 0$ and $x = h$, and a slanted top side that is a straight line. Let the vertical side at $x = 0$ have length a , and the vertical side at $x = h$ have length b . Then the classical formula for the area of a trapezoid is the mean of the lengths of the parallel sides times the distance between the parallel sides:

$$A = \frac{a+b}{2} h \quad (4.83)$$

Let's see if we can derive this formula from integral calculus. The slanted top side of the trapezoid passes through the points $(0, a)$ and (h, b) . It therefore has equation

$$(y - a) = \frac{b-a}{h-0} (x - 0) \quad (4.84)$$

or

$$y = \frac{b-a}{h} x + a \quad (4.85)$$

Using integral calculus, the area of the trapezoid is then the area under this line:

$$\int_0^h \left(\frac{b-a}{h} x + a \right) dx = \frac{b-a}{h} \int_0^h x \, dx + a \int_0^h dx \quad (4.86)$$

$$= \left(\frac{b-a}{h} \frac{x^2}{2} + ax \right) \Big|_0^h \quad (4.87)$$

$$= \left(\frac{b-a}{h} \frac{h^2}{2} + ah \right) - \left(\frac{b-a}{h} \frac{0^2}{2} + a(0) \right) \quad (4.88)$$

$$= h \left(\frac{b-a}{2} + a \right) \quad (4.89)$$

$$= h \left(\frac{b-a}{2} + \frac{2a}{2} \right) \quad (4.90)$$

$$= \frac{a+b}{2} h \quad (4.91)$$

and we have derived the classical formula.

Fence Enclosing Maximum Area

Let's look at an optimization problem. Say you have a pet dog, and want to make a rectangular fenced-in area in the back of your house for him to run around in. You get some fencing material, and plan to use the side of the house for one side of the play area, and the fencing material for the other three sides. Let's say you bought a total length L of fencing material, and let x be the length of the side of the play area that's along the side of the house. Now if $x = 0$, you'll have folded the fencing in half and set it perpendicular to the side of the house — you'll have a rectangle of size zero on one side, and therefore zero area. On the other hand, if $x = L$, then you'll have just set the fencing up against the house, and the play area will be a rectangle whose *other* side is size zero, and therefore encloses zero area again. Clearly there's some value of x in between 0 and L that must *maximize* the enclosed area. The question is: how do you maximize the total area of the play area? In other words, what must be the dimensions of the play area that maximizes the enclosed area for a given length of fencing L ?

To solve this, we'll need to find a formula that gives the enclosed area as a function of x . Since x is the length of the side of the rectangle that's against the house, then the opposite side must also have length x ; therefore the amount of fencing you have left over is $L - x$. This fencing will be used to make the other two sides, so each of the other sides of the rectangle will have length $(L - x)/2$. The rectangular play area will therefore be a rectangle whose sides parallel to the side of the house is x , and whose other sides have length $(L - x)/2$. The area of the rectangular play area is then

$$A(x) = x \frac{L - x}{2} = \frac{1}{2}(-x^2 + Lx) \quad (4.92)$$

This is the equation of a parabola opening downward, so it will have a peak that gives the maximum area. We can find the value of x at the peak (the maximum) because the slope of this curve is zero at the peak. All we need to do is compute the derivative (i.e. slope) of $A(x)$ with respect to x , then set that to zero.

$$\frac{d}{dx} A(x) = 0 \quad (4.93)$$

$$\frac{d}{dx} \left[\frac{1}{2}(-x^2 + Lx) \right] = 0 \quad (4.94)$$

$$\frac{1}{2} \frac{d}{dx} (-x^2 + Lx) = 0 \quad (4.95)$$

$$\frac{1}{2} \left[\frac{d}{dx} (-x^2) + \frac{d}{dx} (Lx) \right] = 0 \quad (4.96)$$

$$\frac{1}{2} \left[-\frac{d}{dx} x^2 + L \frac{d}{dx} x \right] = 0 \quad (4.97)$$

$$\frac{1}{2} [-2x + L] = 0 \quad (4.98)$$

$$-x + \frac{L}{2} = 0 \quad (4.99)$$

$$x = \frac{L}{2} \quad (4.100)$$

Therefore, to maximize the play area for your dog, you should make one side (the side parallel to the side of the house) equal to half the total amount of fencing ($L/2$); the remaining fencing will be divided equally among the other two sides, so the other sides will have length $L/4$. The total area enclosed — the maximum possible area for a length L of fencing — will be $(L/2)(L/4) = L^2/8$.

4.7 Main Ideas

We won't be doing anything very complicated with the calculus in this course; we'll leave mathematical rigor and more complicated problems to a dedicated calculus course. The the purposes of this course, here are the main ideas:

- The number dx is an *infinitesimal* number—a number on the x axis that is ‘infinitely small,’ but not zero.
- The notation $\frac{d}{dx} f(x)$ (the *derivative*) gives the *slope* of the curve $f(x)$ at any x .
- As a special case, the notation $\frac{d}{dt} f(t)$ gives the *rate of change* of $f(t)$ with respect to time t .
- The notation $\int_a^b f(x) dx$ (the *integral*) gives the *area* under the curve $f(x)$ between $x = a$ and $x = b$.
- The derivative and integral are inverses of each other: $\frac{d}{dx} \int f(x) dx = f(x)$

4.8 Going Further

In this chapter we've only just touched on a few of the basic ideas behind the calculus. In a multi-semester course, you'll learn, among other things, how to derive the results presented here; about infinite series and sequences; how to take derivatives and integrals of more complex functions; advanced techniques; how to work in polar coordinates; how to work with functions of several variables; finding areas and volumes of solids of revolution; and how to solve differential equations.

An excellent and brief introduction to the calculus, at about the level of these notes, is *How to Enjoy Calculus* by Eli S. Pine. A typical college-level calculus textbook is *Calculus with Analytic Geometry* by Earl W. Swokowski.

Part II

Waves

Chapter 5

Simple Harmonic Motion

We begin our study of waves with the study of *simple harmonic motion*. Simple harmonic motion is the motion that a particle exhibits when under the influence of a force of the form given by *Hooke's law* (named for the 17th century English scientist Robert Hooke):

$$F = -kx. \quad (5.1)$$

A force of this form describes, for example, the force on a mass attached to a horizontal spring with spring constant k , where k is a measure of the stiffness of the spring. In this case F is the force exerted by the spring, and x is the distance of the mass from its *equilibrium position*—that is, the “resting” position at which the mass can be left where it will not oscillate.

It can be shown using the calculus that when the particle is displaced from the equilibrium position and released, then this force results in an oscillating motion of the particle about the equilibrium position that varies sinusoidally with time t :

$$x(t) = A \cos(\omega t + \delta). \quad (5.2)$$

Here ω is called the *angular frequency* of the motion, and measures how fast the particle oscillates back and forth. The constant A is called the *amplitude* of the motion, and is the maximum distance the particle travels from its equilibrium position, $x = 0$. The constant δ called the *phase constant*, and determines where in its cycle the particle is at time $t = 0$. A plot of $x(t)$ is shown in Fig. 5.1.

Since the sine and cosine function differ only by a phase shift ($\sin \theta \equiv \cos(\theta - \pi/2)$; $\cos \theta \equiv \sin(\theta + \pi/2)$), we could replace the cosine function in Eq. (5.2) with a sine by simply adding an extra $\pi/2$ to the phase constant δ . So either the sine or the cosine can be used equally well to describe simple harmonic motion (Fig. 5.2); here we will choose to use the cosine function.

The calculus may also be used to find the velocity of the particle at any time t ; the result is

$$v(t) = -A\omega \sin(\omega t + \delta). \quad (5.3)$$

Further, it can be shown that the acceleration at any time t is

$$a(t) = -A\omega^2 \cos(\omega t + \delta) \quad (5.4)$$

$$= -\omega^2 x(t). \quad (5.5)$$

Multiplying Eq. (5.5) by the particle mass m , we find

$$ma(t) = F(t) = -m\omega^2 x(t). \quad (5.6)$$

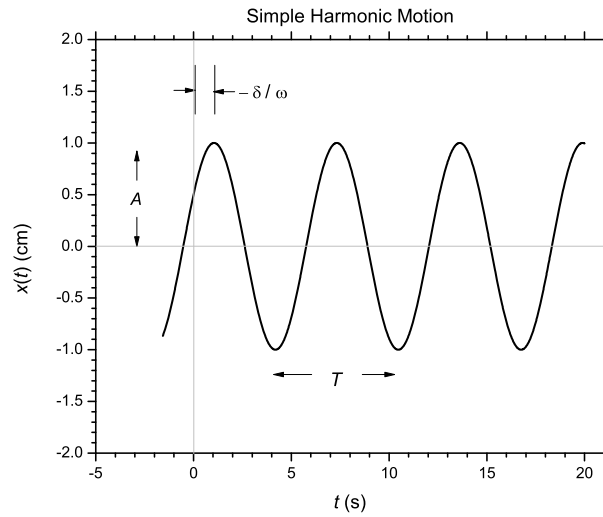


Figure 5.1: Simple harmonic motion. Shown are the amplitude A , period T , and phase constant δ . The horizontal line $x(t) = 0$ is the equilibrium position.

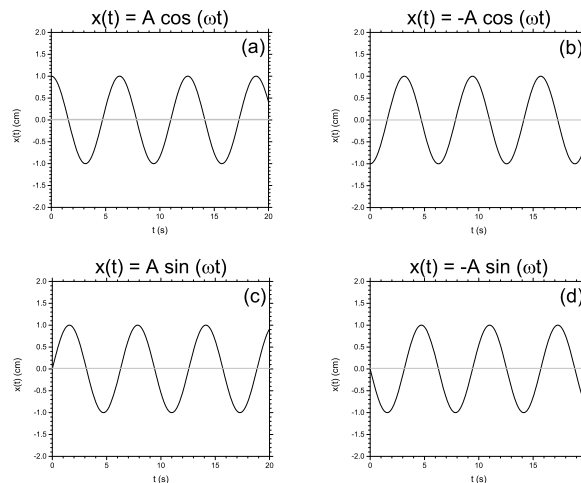


Figure 5.2: Four common special cases of simple harmonic motion phase constant. These are physically identical, and differ only by where the oscillator is in its motion at $t = 0$. (a) $A \cos(\omega t)$; (b) $-A \cos(\omega t)$; (c) $A \sin(\omega t)$; (d) $-A \sin(\omega t)$. In Eq. (5.2), these correspond to: (a) $\delta = 0$, (b) $\delta = \pi$, (c) $\delta = -\pi/2$, (d) $\delta = \pi/2$.

Comparing this with Eq. (5.1) we see that

$$k = m\omega^2, \quad (5.7)$$

or

$$\omega = \sqrt{\frac{k}{m}}. \quad (5.8)$$

In Eq. (5.2), the amplitude A depends on how far the particle was displaced from equilibrium before being released; the phase constant δ just depends on when we choose time $t = 0$; but the angular frequency ω depends on the physical parameters of the system: the stiffness of the spring k and the mass of the particle m .

5.1 Energy

The kinetic energy K of a particle of mass m moving with speed v is defined to be the work required to accelerate the particle from rest to speed v ; this is found to be

$$K = \frac{1}{2}mv^2. \quad (5.9)$$

From Hooke's law, the potential energy U of a simple harmonic oscillator particle at position x can be shown to be

$$U = \frac{1}{2}kx^2. \quad (5.10)$$

The *total* mechanical energy $E = K + U$ of a simple harmonic oscillator can be found by observing that when $x = \pm A$, we have $v = 0$, and therefore the kinetic energy $K = 0$ and the total energy is all potential. Since the potential energy at $x = \pm A$ is $U = kA^2/2$ (by Eq. (5.10)), the total energy must be

$$E = \frac{1}{2}kA^2. \quad (5.11)$$

Since total energy is conserved, the energy E is constant and does not change throughout the motion, although the kinetic energy K and potential energy U do change.

In a simple harmonic oscillator, the energy sloshes back and forth between kinetic and potential energy, as shown in Fig. 5.3. At the endpoints of its motion ($x = \pm A$), the oscillator is momentarily at rest, and the energy is entirely potential; when passing through the equilibrium position ($x = 0$), the energy is entirely kinetic. In between, kinetic energy is being converted to potential energy or vice versa.

We can find the velocity v of a simple harmonic oscillator as a function of position x (rather than time t) by writing an expression for the conservation of energy:

$$E = K + U \quad (5.12)$$

$$\frac{1}{2}kA^2 = \frac{1}{2}mv^2 + \frac{1}{2}kx^2 \quad (5.13)$$

Solving for v , we find

$$v(x) = \pm A \sqrt{\frac{k}{m}} \sqrt{1 - \frac{x^2}{A^2}}. \quad (5.14)$$

This can be simplified somewhat by using Eq. (5.8) to give

$$v(x) = \pm A\omega \sqrt{1 - \frac{x^2}{A^2}}, \quad (5.15)$$

where $A\omega$ is, by inspection of Eq. (5.3), the maximum speed of the oscillator (the speed it has while passing through the equilibrium position).

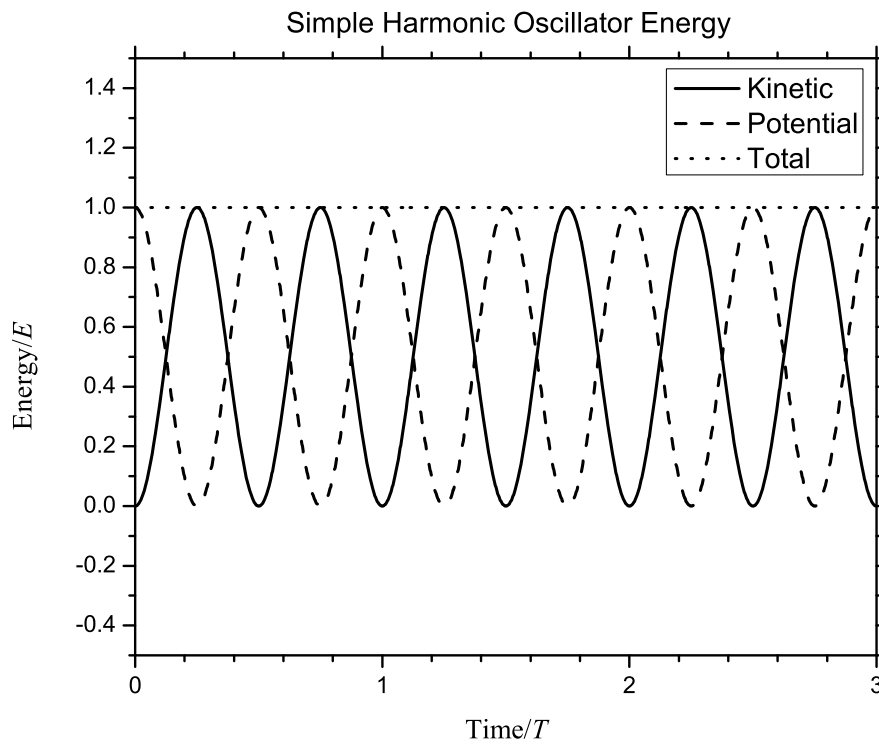


Figure 5.3: Kinetic, potential, and total energy of the simple harmonic oscillator as a function of time. The oscillator continuously converts potential energy to kinetic energy and back again, but the total energy E remains constant.

5.2 The Vertical Spring

If a horizontal mass on a spring is turned to a vertical position, then the spring is stretched by an amount $x_0 = mg/k$, giving it a new equilibrium position. For the vertical spring, the potential energy is still given by $U = \frac{1}{2}kx^2$, but x in this case refers to the distance from the *original* (horizontal) equilibrium position.

5.3 Frequency and Period

The angular frequency ω described earlier is a measure of how fast the oscillator oscillates; specifically, it measures how many radians of its motion the oscillator moves through each second, where one complete cycle of motion is 2π radians. A related quantity is the *frequency* f , which describes how many complete cycles of motion the oscillator moves through per second. The two frequencies are related by

$$\omega = 2\pi f. \quad (5.16)$$

You can think of ω and f as really being the same thing, but measured in different units. The angular frequency ω is measured in units of radians per second (rad/s); the frequency f is measured in units of hertz (Hz), where $1 \text{ Hz} = 1/\text{sec}$.

The reciprocal of the frequency is the *period* T , and is the time required to complete one cycle of the motion:

$$T = \frac{1}{f} = \frac{2\pi}{\omega}. \quad (5.17)$$

The period is measured in units of seconds. As shown in the plot of $x(t)$ (Fig. 5.1), the period T is the time between peaks in the motion.

5.4 Mass on a Spring

The discussion so far has applied to simple harmonic motion in general; there are many specific examples of physical systems that act as simple harmonic oscillators. The most commonly cited example is a mass m on a spring with spring constant k . The spring constant k is a measure of how stiff the spring is, and is measured in units of newtons per meter (N/m). Specifically, k describes how much force the spring exerts per unit distance it is extended or compressed.

A mass on a spring oscillates with angular frequency

$$\omega = \sqrt{\frac{k}{m}}, \quad (5.18)$$

and therefore has period $T = 2\pi/\omega$, or

$$T = 2\pi \sqrt{\frac{m}{k}}. \quad (5.19)$$

It really doesn't matter whether a mass on a spring moves horizontally on a frictionless surface, or bobs up and down vertically. The motion is the same—the only difference is that if you take a horizontal spring and hang it vertically, the equilibrium position will change because of gravity. The period and frequency of motion will be the same.

The importance of the spring example is not that there are government laboratories filled with researchers studying springs; rather the spring example serves as an important model and approximation for other problems. Often even a complicated force can be *approximated* as a linear force (Eq. (5.1)) over some limited

range. In this case one may approximately model the force as a spring force with an “effective spring constant” k , and allow at least an approximate answer to what might otherwise be a difficult problem.

There are several other examples of systems that form simple harmonic oscillators: the torsional pendulum, the simple plane pendulum, a ball rolling back and forth inside a bowl, etc. The simple plane pendulum will be discussed in more detail in Chapter 8.

5.5 More on the Spring Constant

It is often not appreciated that the spring constant k depends not only on the *rigidity* of the spring, but also on the diameter of the spring and the total number of turns of wire in the spring. Consider a vertical spring with spring constant k , and a mass m hanging on one end. Assume the system is in its equilibrium position, and in this position it has length L_0 and consists of N turns of wire. Now if you apply an additional downward force F to the mass, the string will stretch by an additional amount x given by Hooke's law: $x = F/k$. This stretching will manifest itself as an additional spacing of x/N between adjacent turns of the spring. It is this additional spacing per turn that is the true measure of the inherent “stiffness” of the spring.

Now suppose this spring is cut in half and put in its equilibrium position. Its new length will be $L_0/2$, and will consist of $N/2$ turns of wire. When the same additional force F is applied to the mass m , the additional spacing between adjacent turns of the spring will be the same as before, x/N , because the spring still has the same stiffness. Since the number of turns is now $N/2$, this means that the additional total stretching of the spring is $x/2$, so it will stretch by only half as much as before. By Hooke's law, the spring constant is now $k' = F/(x/2) = 2F/x = 2k$, so the spring constant is now twice what it was before. In other words, *cutting the spring in half will double the spring constant*. Likewise, doubling the length (number of turns) of the spring will halve its spring constant.

Another way to think of this is to consider two springs connected in series or in parallel (Fig. 5.4). If several springs are connected end-to-end (i.e. *in series*), then the equivalent spring constant k_s of the system will be given by

$$\frac{1}{k_s} = \sum_i \frac{1}{k_i} \quad (5.20)$$

$$= \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \dots \quad (5.21)$$

If the springs are connected *in parallel*, then the equivalent spring constant k_p of the system will be

$$k_p = \sum_i k_i \quad (5.22)$$

$$= k_1 + k_2 + k_3 + \dots \quad (5.23)$$

For example, if two identical springs, each of spring constant k , are connected in series, then the combination will have an equivalent spring constant of $k/2$. If the two identical springs were instead connected in parallel, then the combination would have an equivalent spring constant of $2k$, as shown in Figure (5.4).

Now imagine you have a long spring of spring constant k . You can imagine it as being two identical springs connected in series, each having spring constant $2k$, so that the combination has a total equivalent spring constant of $[(1/2k) + (1/2k)]^{-1} = k$. If the long spring is cut in half, then you are left with only one of those smaller springs of spring constant $2k$, so again we reach the conclusion that cutting the spring in half will double the spring constant.

It's possible to calculate the spring constant from the geometry of the spring. The formula is ¹

$$k = \frac{Gd^4}{8ND^3} \quad (5.24)$$

¹See e.g. http://www.engineersedge.com/spring_comp_calc_k.htm

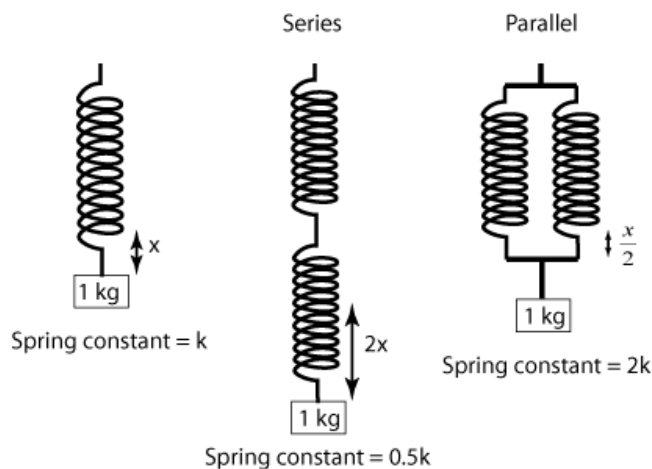


Figure 5.4: Springs in series and parallel (Credit: <http://spmphysics.onlinetuition.com.my>).

where d is the wire diameter, N is the number of active turns in the spring, D is the coil diameter (measured from the *center* of the wire), and G is called the *modulus of rigidity* of the spring material; G is given by

$$G = \frac{Y}{2(1 + \nu)} \quad (5.25)$$

where Y is the *Young's modulus* of the material (a measure of how much it stretches when pulled or compressed), and ν is the material's *Poisson ratio* (a measure of how much it squeezes sideways when compressed). These are properties that are characteristic of the material, and can be looked up in a handbook of material properties. Values for a few materials are shown in the table below.

Table 5-1. Young's Moduli and Poisson Ratios.

Material	Young's Modulus Y (N/m ²)	Poisson Ratio ν
Aluminum	69×10^9	0.334
Bronze	100×10^9	0.34
Copper	117×10^9	0.355
Lead	14×10^9	0.431
Magnesium	45×10^9	0.35
Stainless steel	180×10^9	0.305
Titanium	110×10^9	0.32
Wrought iron	200×10^9	0.278

Notice from Eq. (5.24) that if the spring is cut in half, N will be half its original value, and so the spring constant k will be doubled, in agreement with what we've found earlier.

Example. Suppose we make a spring of 1 mm diameter copper wire, the diameter of the spring is 1 cm, and there are 50 turns of wire in the spring. What is the spring constant?

Solution. From the above table, for copper, $Y = 117 \times 10^9 \text{ N/m}^2$ and $\nu = 0.355$. From Eq. (5.25), we have

$$G = \frac{Y}{2(1 + \nu)} = \frac{117 \times 10^9 \text{ N/m}^2}{2(1 + 0.355)} = 43.2 \times 10^9 \text{ N/m}^2$$

And the spring constant is found from Eq. (5.24)

$$k = \frac{Gd^4}{8ND^3} = \frac{(43.2 \times 10^9 \text{ N/m}^2)(10^{-3} \text{ m})^4}{8(50)(10^{-2} \text{ m})^3} = 108 \text{ N/m}$$

Chapter 6

Damped Oscillations

If you build a real simple harmonic oscillator by attaching a mass to a spring and letting it oscillate back and forth, you'll find that it doesn't oscillate forever, as would be predicted by Eq. (5.2). Instead, the motion will damp out due to frictional forces, and the oscillator will eventually stop oscillating.

We can model the damping force F_d as being proportional to the speed v of the oscillator:

$$F_d = -bv, \quad (6.1)$$

where b is a damping constant (in units of kg/s). There are three different cases of damped motion: *underdamped*, *overdamped*, and *critically damped*. In the following discussion, the natural oscillation frequency of the undamped oscillator is¹ $\omega_0 = \sqrt{k/m}$.

6.1 Underdamped

In the underdamped case, the damping constant b is small ($b < 2m\omega_0$), and the oscillations gradually decrease in amplitude. In this case, the motion will be described by

$$x(t) = Ae^{-(b/2m)t} \cos(\omega' t + \delta), \quad (6.2)$$

where A is the initial amplitude and δ is the phase constant. The underdamped oscillator oscillates at a slower frequency ω' than if it were undamped, where ω' is given by

$$\omega' = \omega_0 \sqrt{1 - \left(\frac{b}{2m\omega_0}\right)^2}. \quad (6.3)$$

Fig. 6.1 shows what the motion looks like: it is a cosine curve modulated by an overall exponentially decaying “envelope”.

6.2 Overdamped

Now imagine that a simple harmonic oscillator is immersed in a thick liquid like honey. In this case the damping constant b is large (specifically, $b > 2m\omega_0$), and the motion is said to be *overdamped*. If the mass is

¹The quantity ω_0 is customarily pronounced “omega-nought”, *nought* being an old-fashioned term for *zero*.

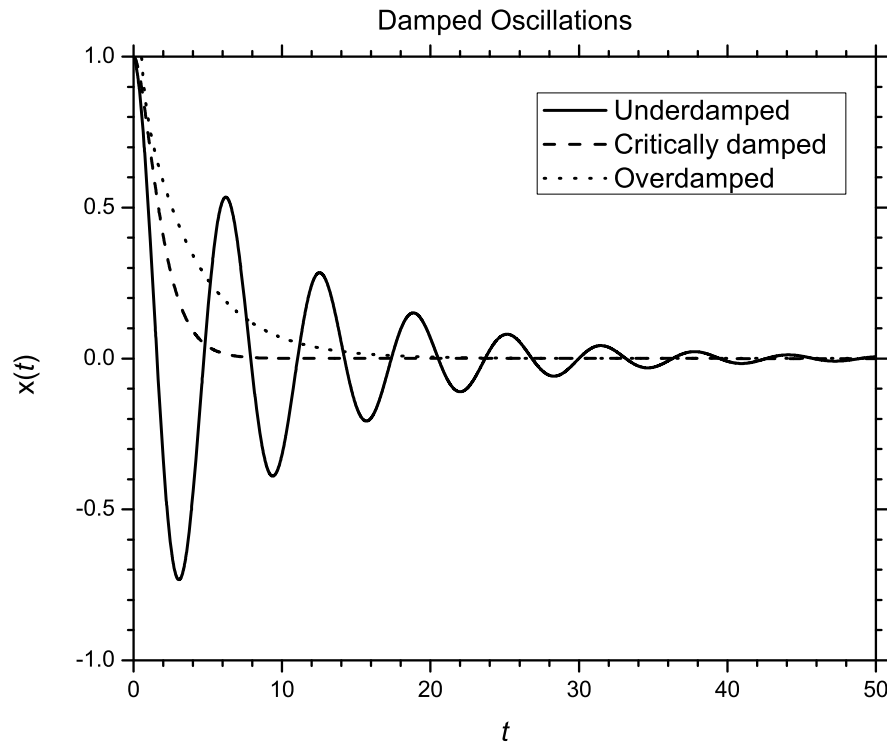


Figure 6.1: Damped oscillations.

displaced from its equilibrium position, then it will slowly move toward equilibrium, but will not overshoot it, so no oscillations will occur. In this case the motion is described by

$$x(t) = e^{-(b/2m)t} (Ae^{Ct} + Be^{-Ct}), \quad (6.4)$$

where $C = \sqrt{(b/2m)^2 - \omega_0^2}$, and the constants A and B depend on the initial conditions. This case is also illustrated in Fig. 6.1.

6.3 Critically Damped

In between the underdamped and overdamped case is the case of *critical* damping, where the damping constant $b = 2m\omega_0$. In this case, the mass returns to its equilibrium position as quickly as possible, without overshooting. The motion in this case is

$$x(t) = e^{-(b/2m)t} (At + B), \quad (6.5)$$

where again the constants A and B depend on the initial conditions. Fig. 6.1 shows critical damping compared to the similar-looking overdamped case.

Chapter 7

Forced Oscillations

Now suppose that we have a harmonic oscillator that is being driven, or *forced*. For example, imagine a spring that has a mass m attached to one end, and the other end is connected to a motor-driven piston that moves back and forth. What happens in this case is that the motion of the oscillator is fairly complicated at first, then settles down to a “steady-state” motion, where the oscillator oscillates at the same frequency as the driving force.

Suppose we have a damped oscillator whose natural oscillation frequency is $\omega_0 = \sqrt{k/m}$, and the oscillator is being driven by a force of the form $F(t) = F_0 \sin \Omega t$, so the driving force has amplitude F_0 and angular frequency Ω . Then after the initial complicated motion has died out, the steady-state motion will be an oscillatory motion with the same frequency as the driving force,

$$x(t) = A \cos(\Omega t + \delta). \quad (7.1)$$

Here A is the amplitude of the motion, which will depend on how far the driving frequency Ω is from the natural frequency ω_0 :

$$A = \frac{F_0/m}{\sqrt{(\Omega^2 - \omega_0^2)^2 + (b\Omega/m)^2}}. \quad (7.2)$$

7.1 Resonance

Notice that in Eq. (7.2), the denominator will be smallest when $\Omega = \omega_0$, so that the oscillator is being driven at its natural frequency of oscillation. This situation is called *resonance*, and can result in very large oscillations. (Note that in Eq. (7.2) if the damping constant $b = 0$ and $\Omega = \omega_0$, the denominator is zero and amplitude becomes infinite!) We’re familiar with examples of resonance in everyday life: for example, an opera singer who sings a loud, high note and is able to shatter a crystal goblet. Engineers have to be careful in designing things like buildings, bridges, aircraft, spacecraft, etc. that the objects won’t be subjected to being driven at one of the natural frequencies of oscillation of the object. Marching soldiers break step when crossing a bridge, just in case the cadence of the march is at one of the natural frequencies of oscillation of the bridge, which could cause the bridge to collapse.

Fig. 7.1 shows a plot of amplitude vs. forcing frequency for a typical forced oscillator. Resonance is shown by the large increase in the amplitude of the forced oscillations when $\Omega = \omega_0$. The smaller the damping force, the larger the amplitude at resonance.

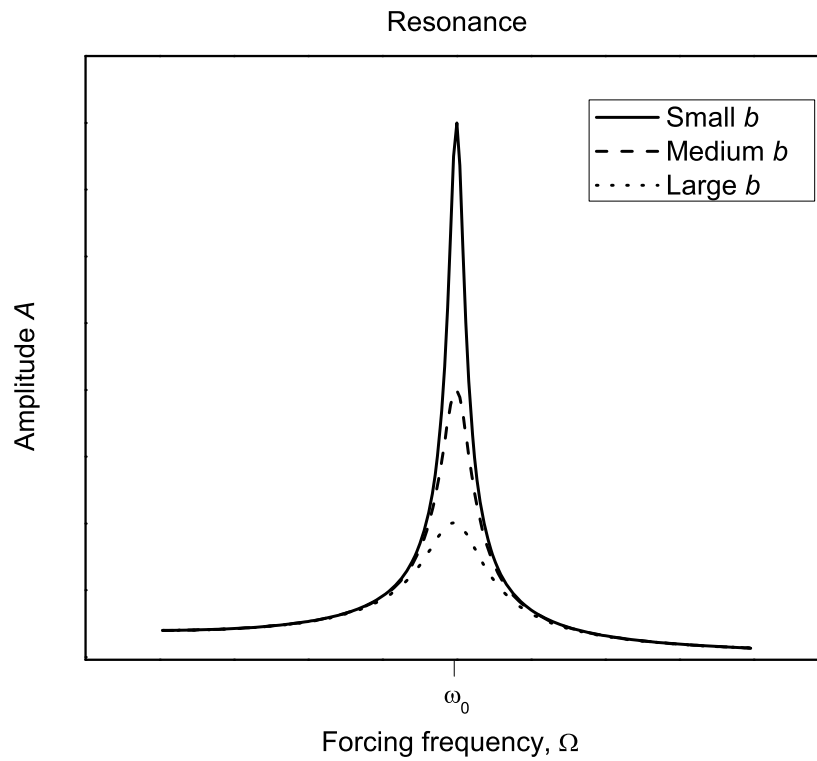


Figure 7.1: Amplitude vs. forcing frequency for forced oscillations, for various damping coefficients. The maximum amplitude occurs when the forcing frequency Ω is equal to the natural frequency ω_0 , a phenomenon known as *resonance*.

Chapter 8

The Pendulum

A *simple plane pendulum* (Fig. 8.1) consists of a mass m attached to one end a light rod of length L ; the other end of the rod is attached to a frictionless pivot. The pendulum is initially displaced from the vertical by an angle θ_0 and released, causing it to swing back and forth. Is the pendulum a simple harmonic oscillator?

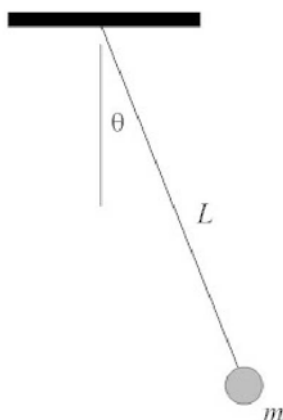


Figure 8.1: A simple plane pendulum.

Analyzing the geometry of the pendulum shows that the restoring force—the force acting on the pendulum directing it back to its equilibrium position (vertical)—is $-mg \sin \theta$, where θ is the angle from the vertical, g is the acceleration due to gravity, and the minus sign indicates that the restoring force acts opposite the direction of angular displacement. We can write the restoring force as

$$F = -mg \sin \theta. \quad (8.1)$$

But for a simple harmonic oscillator, the restoring force must be in the form $F = -kx$, so the pendulum is *not* a simple harmonic oscillator.

Suppose, however, that we restrict the pendulum to *small* oscillations. For small angles, we can make the approximation $\sin \theta \approx \theta$, where θ is in radians. Under this approximation, Eq. (8.1) becomes

$$F \approx -mg\theta, \quad (8.2)$$

which is the form of equation of a simple harmonic oscillator. So while the pendulum is not strictly a simple harmonic oscillator, it is *approximately* a simple harmonic oscillator when the oscillations are small.

8.1 Equation of Motion

Since the pendulum (in the small-angle approximation) is a simple harmonic oscillator, its motion is given by (cf. Eq. (5.2))

$$\theta(t) = \theta_0 \cos(\omega t + \delta), \quad (8.3)$$

where θ_0 is the (angular) amplitude in radians and δ is the phase constant. To find the angular frequency ω , note from geometry that the horizontal displacement distance of the pendulum is $x = L \sin \theta$. Writing Eq. (8.2) as

$$F \approx -\left(\frac{mg}{L}\right) (L \sin \theta), \quad (8.4)$$

and comparing with Eq. (5.1), we can see that the effective spring constant for the pendulum is

$$k_{\text{eff}} = \frac{mg}{L}. \quad (8.5)$$

Now for the harmonic oscillator we know $\omega = \sqrt{k/m}$, and so

$$\omega = \sqrt{\frac{k_{\text{eff}}}{m}} = \sqrt{\frac{mg}{mL}} \quad (8.6)$$

or

$$\omega = \sqrt{\frac{g}{L}}. \quad (8.7)$$

So the small-amplitude motion of the simple plane pendulum is the same as the mass on a spring; but the angular frequency of the spring system is given by $\omega = \sqrt{k/m}$, and for the pendulum it is $\omega = \sqrt{g/L}$. Other simple harmonic oscillators with have other expressions for their angular frequency ω , each depending on the physical parameters of the system.

8.2 Period

Since the period of a simple harmonic oscillator is given by $T = 2\pi/\omega$, we find, using Eq. (8.7), that the period of the pendulum is

$$T = 2\pi \sqrt{\frac{L}{g}}. \quad (8.8)$$

Remember that this is just an *approximate* expression for the period of a pendulum, with the approximation being better the smaller the amplitude θ_0 . An exact treatment requires the period T to be expressed as an infinite series. The details require some advanced mathematics that is beyond the scope of this course, but if you're interested, an exact treatment of the simple plane pendulum is given in Appendix R.

8.3 The Spherical Pendulum

A *spherical pendulum* is similar to a simple plane pendulum, except that the pendulum is not constrained to move in a plane; the mass m is free to move in two dimensions along the surface of a sphere. Figure 8.2 shows a photograph of the movement of a spherical pendulum.

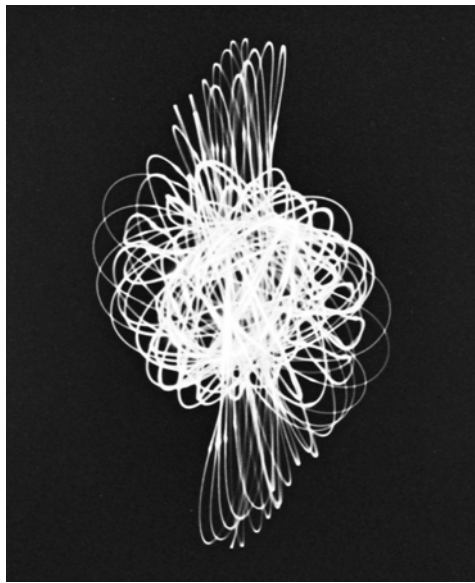


Figure 8.2: Trace of the motion of a spherical pendulum, made by the author. A flashlight lens was covered with a piece of cardboard in which a small hole was punched. The flashlight was then suspended by a string from the ceiling (lens downward) to create a pendulum. The room was then darkened, the flashlight turned on, and the flashlight pendulum allowed to swing back and forth for several minutes above a camera which was on the floor pointing up toward the ceiling. The camera shutter was kept open, allowing this time-exposure image to be made on the film. (*Image Copyright* © 2011 D.G. Simpson.)

8.4 The Conical Pendulum

A *conical pendulum* is also similar to a simple plane pendulum, except that the pendulum is constrained to move along the surface of a cone, so that the mass m moves in a horizontal circle of radius r , maintaining a constant angle θ from the vertical.

For a conical pendulum, we might ask: what speed v must the pendulum bob have in order to maintain an angle θ from the vertical? To solve this problem, let the pendulum have length L , and let the bob have mass m . A general approach to solving problems involving circular motion like this is to identify the force responsible for keeping the mass moving in a circle, then set that equal to the centripetal force mv^2/r . In this case, the force keeping the mass moving in a circle is the horizontal component of the tension T , which is $T \sin \theta$. Setting that equal to the centripetal force, we have

$$T \sin \theta = \frac{mv^2}{r}. \quad (8.9)$$

The vertical component of the tension is

$$T \cos \theta = mg \quad (8.10)$$

Dividing Eq. (8.9) by Eq. (8.10),

$$\tan \theta = \frac{v^2}{gr} \quad (8.11)$$

From geometry, the radius r of the circle is $L \sin \theta$. Making this substitution, we have

$$\tan \theta = \frac{v^2}{gL \sin \theta}. \quad (8.12)$$

Solving for the speed v , we finally get

$$v = \sqrt{Lg \sin \theta \tan \theta}. \quad (8.13)$$

8.5 The Torsional Pendulum

A *torsional pendulum* (Fig. 8.3) consists of a mass m attached to the end of a vertical wire. The body is then rotated slightly and released; the body then twists back and forth under the force of the twisting wire. As described earlier, the motion is governed by the rotational version of Hooke's law, $\tau = -\kappa\theta$.

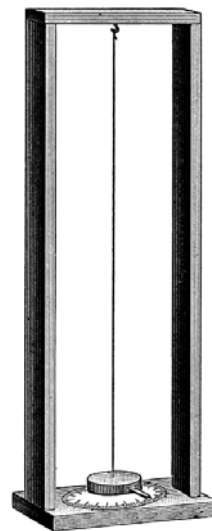


Figure 8.3: A torsional pendulum. (Ref. [1])

8.6 The Physical Pendulum

A *physical pendulum* consists of an extended body that allowed to swing back and forth around some pivot point. If the pivot point is at the center of mass, the body will not swing, so the pivot point should be displaced from the center of mass. As an example,

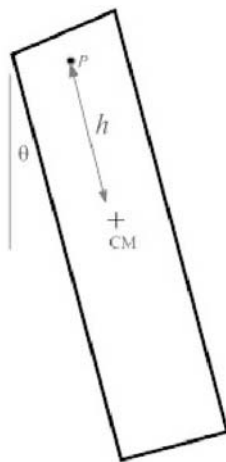


Figure 8.4: A physical pendulum. The object has mass M and is suspended from point P ; h is the distance between P and the center of mass.

you can form a physical pendulum by suspending a meter stick from one end and allowing to swing back and forth.

In a physical pendulum of mass M , there is a force Mg acting on the center of mass. Suppose the body is suspended from a point that is a distance h from the center of mass (Fig. 8.4). Then there is a weight force Mg acting on the center of mass of the body, which creates a torque $-Mgh \sin \theta$ about the pivot point. Then by the rotational version of Newton's second law,

$$\tau = I\alpha \quad (8.14)$$

$$-Mgh \sin \theta = I\alpha, \quad (8.15)$$

where I is the moment of inertia of the body when rotated about its pivot point, and α is the angular acceleration. Like the simple plane pendulum, this is a difficult equation to solve for $\theta(t)$, but it becomes much easier to solve if we restrict the problem to small oscillations θ . If θ is small, we can make the approximation $\sin \theta \approx \theta$, and we have

$$-Mgh\theta \approx I\alpha. \quad (8.16)$$

It can be shown, using the theory of differential equations, that this equation has solution

$$\theta(t) = \theta_0 \cos(\omega t + \delta), \quad (8.17)$$

where θ_0 is the (angular) amplitude of the motion (in radians), $\omega = \sqrt{Mgh/I}$ is the angular frequency of the motion (rad/s), and δ is an arbitrary integration constant (seconds).

The period T of the motion (the time required for one complete back-and-forth cycle) is given by

$$T = \frac{2\pi}{\omega}, \quad (8.18)$$

or

$$T = 2\pi \sqrt{\frac{I}{Mgh}}. \quad (8.19)$$

(See Appendix Q for a table of moments of inertia.)

8.7 Other Pendulums

- *Double pendulum.* A *double pendulum* is formed by attaching one pendulum to the bob of another, so that the two pendulums are attached vertically and both bobs are free to move. The motion of a double pendulum is a classic exercise in an advanced formulation of Newtonian classical mechanics called *Langrangian mechanics*.
- *Ballistic pendulum.* A *ballistic pendulum* is a type of pendulum used to measure the speed of high-speed objects like bullets. A bullet is fired into the pendulum bob, and the pendulum is constructed with a ratchet mechanism that holds the pendulum in place once it reaches its maximum displacement from the vertical. Knowing the masses of the bullet and pendulum bob, the length of the pendulum, and the angle the pendulum reaches when the bullet is fired into it, it is possible to deduce the velocity of the bullet.
- *Foucault pendulum.* A *Foucault pendulum* is a type of simple plane pendulum that is used to demonstrate the rotation of the Earth. As the pendulum swings back and forth in a plane, the Earth rotates underneath the pendulum, causing its trace along the ground to drift with time.

Chapter 9

Waves

Having examined simple harmonic motion, we are now in a position to examine waves. A *wave* is a disturbance in a material medium that propagates itself through the medium.¹ In a harmonic wave, each particle in the medium undergoes simple harmonic motion, but adjacent particles are slightly out of phase with each other, which results in the wave disturbance propagating through the medium while the particles of the medium itself simply oscillate in place.

9.1 Types of Waves

There are two major types of waves:

- *Transverse waves.* Particles of the medium move *perpendicular* to the direction of wave motion. Transverse waves can travel in solids only; they cannot propagate in fluids.
- *Longitudinal waves.* Particles of the medium move *parallel* to the direction of wave motion. Longitudinal waves can propagate in both solids and fluids.

You can create a transverse wave in a long string under tension by giving it a quick flip at one end. The disturbance will propagate down the string, although any point on the string will move up and down, perpendicular to the string.

You can create a longitudinal wave by stretching a Slinky toy (or other spring) and giving it a quick in-and-out “pulse” at one end. You’ll see the coils of the Slinky be alternately close together and spread apart as the disturbance propagates down the length of the spring. A region where the coils are close together is called a *compression*, and a region where the coils are far apart is called a *rarefaction*.

Some waves are neither transverse nor longitudinal. For example, if you examine water waves in the ocean, you will see that particles on the surface move in cycloid-looking paths that have both components both parallel and perpendicular to the wave velocity—so water waves are a combination of transverse and longitudinal waves.

You can create a single *wave pulse* by giving the medium a single displacement at one end; the resulting pulse will then propagate through the medium. You can also follow one pulse by another continuously, resulting in a *wave train*. For example, you can displace one end of the medium with simple harmonic motion, and you will see a continuous wave train propagating through the medium. This will result in a harmonic wave, which can be represented mathematically as

$$y(x, t) = A \cos(kx - \omega t + \delta). \quad (9.1)$$

¹There are some notable exceptions: electromagnetic waves, quantum-mechanical waves, and gravitational waves do not require a physical medium in which to propagate.

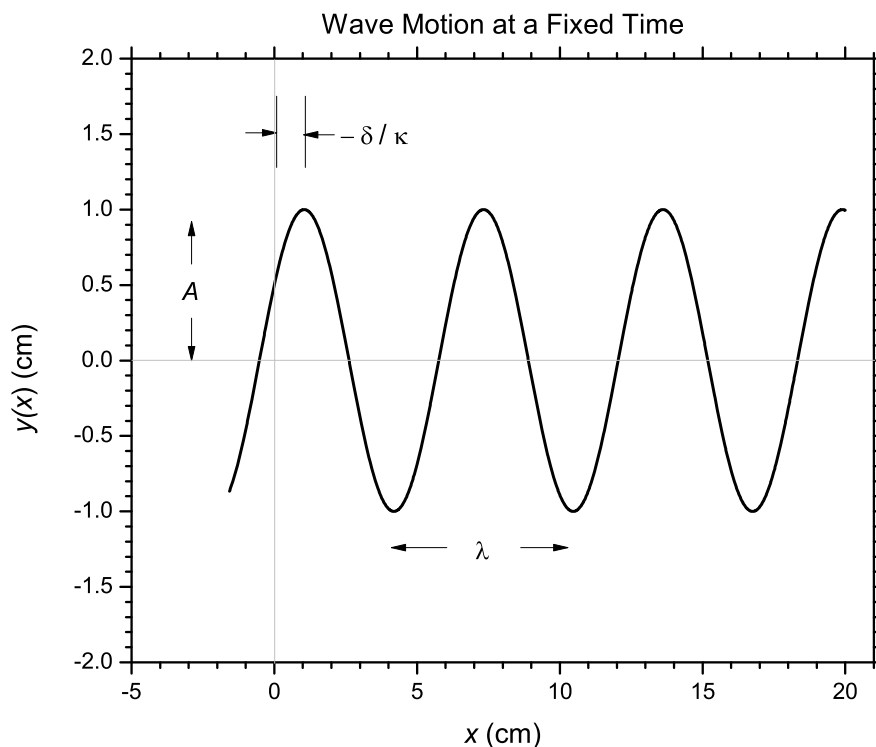


Figure 9.1: Wave motion at a fixed time t . A is the wave amplitude, δ is the phase constant, λ is the wavelength, and $\kappa = 2\pi/\lambda$ is the wave number.

This looks similar to the equation for simple harmonic motion, only it involves *both* position x and time t . Here y is the displacement of the wave at position x and time t , A is the wave amplitude, ω is the angular frequency of the wave, and δ is the phase constant that is determined from the initial conditions. The variable κ is called the *wave number*, and is defined as

$$\kappa = \frac{2\pi}{\lambda}, \quad (9.2)$$

where λ , called the *wavelength* of the wave, is the distance between successive wave crests.² Fig. 9.1 shows a “snapshot” of a harmonic wave at an instant in time, with A , δ , and λ illustrated. As time increases, you would see this wave move to the right. (This analysis applies equally to transverse and longitudinal waves.)

9.2 Wave Speed

The speed of a wave may be thought of as the speed of a single wave crest as it propagates through the medium. Since the wave moves by one wavelength λ in a time equal to the period T , the wave speed is

²Some physicists define the wave number as $1/\lambda$.

$v = \lambda/T$; and since $T = 1/f$, we can write

$$v = f\lambda. \quad (9.3)$$

This equation relates the temporal frequency f of the wave to its “spatial frequency”, or wavelength, λ .

9.3 String Waves

Now let's examine some properties of waves propagating in strings. Although string waves are occasionally of interest (as in some musical instruments), the reason we're interested in them here is that they form a simple system that's easy to visualize, yet illustrates many properties that we'll find later in other kinds of waves.

First, let's look at a formula for the speed v of a wave in a string, in terms of the physical properties of the string (its tension and density). We'll skip the derivation and just present the result:

$$v = \sqrt{\frac{F_T}{m/L}}, \quad (9.4)$$

where v is the wave speed, F_T is the tension in the string (in newtons), and m/L is the mass density of the string (mass per unit length, in kg/m). (We'll see later that the speed of sound waves in a fluid follows a similar formula: $v = \sqrt{B/\rho}$, where B is the bulk modulus and ρ is the density of the medium. The speed of sound waves in a solid is $v = \sqrt{Y/\rho}$, where Y is the Young's modulus.)

9.4 Reflection and Transmission

Next, let's look at what happens when a wave pulse hits a boundary—for example, a boundary with a lighter or heavier string. Generally at the boundary there will be a *reflected wave* that returns in the opposite direction as the incident wave, and there will be a *transmitted wave* that continues into the new medium, in the same direction as the incident wave. The various possibilities are shown in Fig. 9.2.

Note the following points:

- When the incident wave is incident on a “heavier” (denser) medium, the returning reflected wave will be *inverted*.
- When the incident wave is incident on a “lighter” medium, the returning reflected wave will be right-side up.
- The transmitted wave will always be right-side up.
- A fixed end may be regarded as an infinitely heavy medium, and may be thought of as an end that is attached to a heavy wall. In this case there is no transmitted wave.
- A free end may be regarded as a medium of zero density, and may be thought of as an end attached to a ring that is free to move up and down a vertical pole. In this case there is no transmitted wave.
- The transmitted wave will be largest when both media have the same density; in this case there is no reflected wave, and all of the incident wave is transmitted.

We might ask: in string waves, how much of the incident wave is reflected, and how much is transmitted? We can define the *coefficient of reflection* R as the ratio of reflected to incident wave energy, and similarly define a *coefficient of transmission* T as the ratio of transmitted to incident wave energy. Since both strings

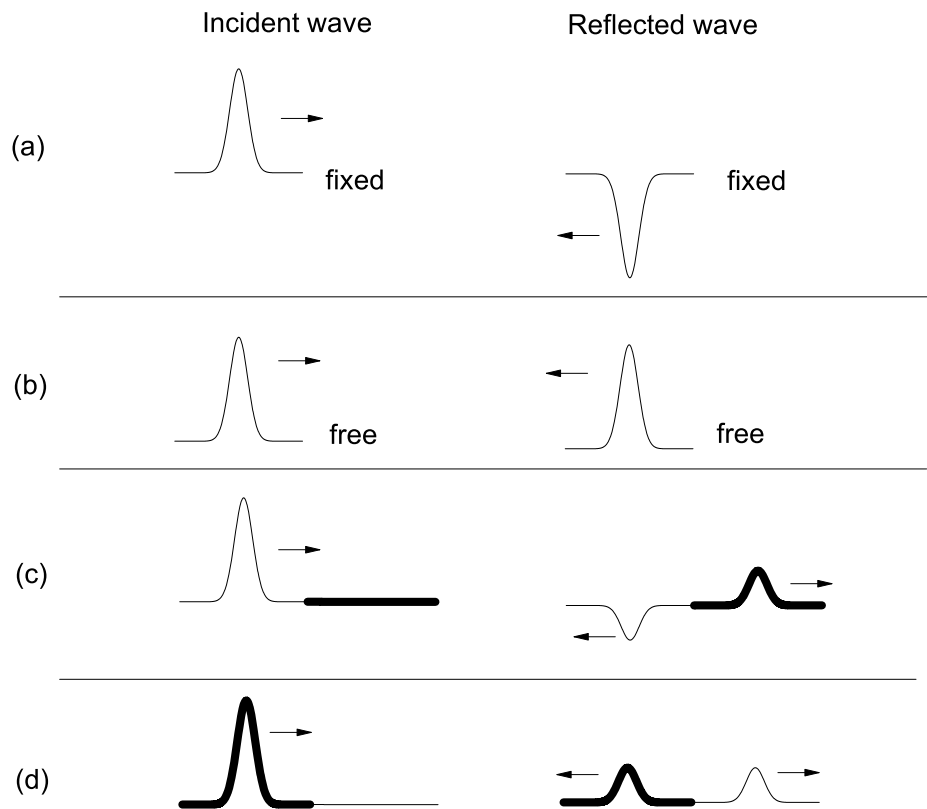


Figure 9.2: Reflection and transmission of waves in a string. (a) Wave in string incident onto fixed end; (b) string wave incident onto free end; (c) wave in light string incident onto heavy rope; (d) wave in heavy rope incident onto light string. When the incident wave hits a “heavier” medium, the reflected wave will be inverted.

are under the same tension, these coefficients depend only on the densities of the two strings. Writing the string density (mass per unit length) as $\rho = m/L$, the coefficients R and T turn out to be (Ref. [8])

$$R = \left(\frac{\sqrt{\rho_1} - \sqrt{\rho_2}}{\sqrt{\rho_1} + \sqrt{\rho_2}} \right)^2 \quad (9.5)$$

$$T = \frac{4\sqrt{\rho_1\rho_2}}{(\sqrt{\rho_1} + \sqrt{\rho_2})^2} \quad (9.6)$$

where the subscripts 1 and 2 refer to the two strings. Note the following about these equations:

1. $R + T = 1$. (This is due to the conservation of energy.)
2. If $\rho_1 = \rho_2$, then $R = 0$ and $T = 1$: if both strings have the same density, then all of the incident wave is transmitted, and none is reflected.
3. If $\rho_2 = 0$, then $R = 1$ and $T = 0$: for a “free” end, all the wave is reflected and none is transmitted.
4. Similarly, if $\rho_2 \rightarrow \infty$, then $R = 1$ and $T = 0$: for a “fixed” end, all the wave is reflected and none is transmitted.

The coefficients R and T show how the initial wave *energy* is divided among the reflected and transmitted waves. The *amplitudes* of the reflected and transmitted waves (A_r and A_t , respectively) are related to the incident wave amplitude A_i by (Ref. [8])

$$\frac{A_r}{A_i} = \frac{\sqrt{\rho_1} - \sqrt{\rho_2}}{\sqrt{\rho_1} + \sqrt{\rho_2}} \quad (9.7)$$

$$\frac{A_t}{A_i} = \frac{2\sqrt{\rho_1}}{\sqrt{\rho_1} + \sqrt{\rho_2}} \quad (9.8)$$

9.5 Superposition

What happens when two waves collide? It turns out that while they overlap (a situation called *superposition*), their displacements y will add algebraically. Given two waves $y_1(x, t)$ and $y_2(x, t)$, the total wave $y(x, t)$ will be the sum of the two: $y = y_1 + y_2$.

An example is shown in Fig. 9.3, where two wave pulses are shown colliding with each other. During the time that the wave pulses overlap, they add algebraically. Afterwards, the two pulses continue, as if they just passed right through each other.

This ability of waves to pass through each other is fortunate, and we observe it in everyday life. For example, you can talk with someone directly across from you, at the same time people to your left and right can talk to each other. The sound waves pass right through each other, and each person is able to hear and understand his partner without difficulty. The same is true of light waves: each person is able to see the other three, because the light waves are able to pass through each other.

9.6 Interference

Closely related to the idea of superposition is the concept of wave *interference*. When two waves overlap and their displacements y are in the same direction, the two waves will, by superposition, add together to make

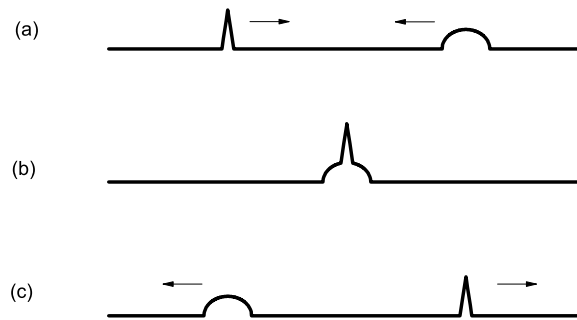


Figure 9.3: Colliding wave pulses. (a) Before the collision. (b) During the collision, the wave pulses overlap, and amplitudes add algebraically. (c) After the collision, the wave pulses have passed through each other unchanged.

a bigger wave. This situation is called *constructive interference*—the waves add together constructively. On the other hand, if the waves overlap and their displacements are in the *opposite* direction, the two waves will tend to cancel each other out, resulting in a smaller wave (or even no wave at all). This situation is called *destructive interference*.

An example of wave interference is shown in Fig. 9.4. The figure shows two wave pulses of the same size and shape headed toward each other. Fig. 9.4(a) shows constructive interference, and Fig. 9.4(b) shows destructive interference. Notice something interesting that happens in the case of destructive interference: although the waves momentarily cancel completely and leave no wave at all, the particles in the string are still in motion, so new waves will emerge from the flat string and continue on their way.

9.7 Wave Energy

Waves carry energy, but not mass. Each particle of the wave medium oscillates in place around its own equilibrium position, so no mass is transported. The wave disturbance does move, though, and carries energy with it. How much energy does a wave transport?

Suppose we have a harmonic wave traveling through a medium. Each particle of the medium oscillates with simple harmonic motion, and has energy $E = kA^2/2$, where k is the spring constant and A is the amplitude. By Eq. (5.7), we know the spring constant is related to the frequency by $k = m\omega^2$. Substituting this into the expression for energy gives

$$E = \frac{1}{2}m\omega^2 A^2. \quad (9.9)$$

Now if the wave has surface area S and moves with velocity v , then in time t it will sweep out a volume Svt . Since the mass m of a small volume of the medium is the mass divided by the volume, we have $m = \rho Svt$,

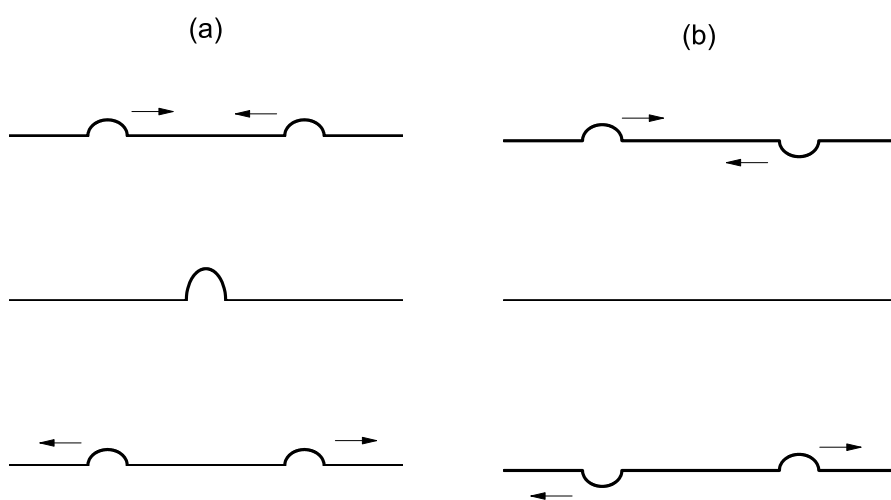


Figure 9.4: Interference in two colliding wave pulses. (a) Constructive interference; when the two pulses overlap, their displacements add constructively, giving a large pulse equal to the sum of the original two. (b) Destructive interference; when the two pulses overlap, they cancel out and momentarily add to zero.

where ρ is the density of the medium. Substituting this for m into Eq. (9.9), we have

$$E = \frac{1}{2}\rho S v t \omega^2 A^2. \quad (9.10)$$

This says that the energy E carried by the wave is proportional to the square of the amplitude A , and also to the square of the frequency ω . The *power* P (in watts) is the energy per unit time, or

$$P = \frac{E}{t} = \frac{1}{2}\rho S v \omega^2 A^2. \quad (9.11)$$

Now dividing the power by the surface area S gives an expression for the wave *intensity* I (watts per square meter):

$$I = \frac{P}{S} = \frac{1}{2}\rho v \omega^2 A^2. \quad (9.12)$$

9.8 Wave Intensity

Another issue that often arises is how wave intensity I varies with the distance r from the source of the waves. The answer is: it depends upon the shape of the waves. The power emitted by the source will be distributed along a surface at distance r , and the shape of that surface will depend on the shape of the waves.

One common case is *spherical waves*, which are produced by a point source of spherical source. For spherical waves, the power P emitted by the source is spread over the surface of a sphere of radius r . If the power is radiated *isotropically* (that is, equally in all directions), then the intensity in any direction at a distance r from the source will be $I = P/(4\pi r^2)$, so $I \propto 1/r^2$. Since the intensity is proportional to the square of the amplitude, this implies the wave *amplitude* drops off as $A \propto 1/r$. In summary, for spherical waves,

$$I \propto \frac{1}{r^2}; \quad A \propto \frac{1}{r}. \quad (9.13)$$

Another case is *cylindrical waves*, which are produced by a line or cylindrical source. In this case the power is distributed over the surface of a cylinder of radius r , and we have

$$I \propto \frac{1}{r}; \quad A \propto \frac{1}{\sqrt{r}}. \quad (9.14)$$

When either of these types of waves is observed very far from the source, they approximate *plane waves*, where the wave fronts are planes. For plane waves, the intensity I and amplitude A are both constant and independent of r :

$$I = \text{const.}; \quad A = \text{const.} \quad (9.15)$$

9.9 Ocean Waves

The speed of ocean waves is a function of their wavelength and the ocean depth. Ocean wave speed is given by the expression³

$$v = \sqrt{\frac{g\lambda}{2\pi} \tanh\left(2\pi \frac{d}{\lambda}\right)}, \quad (9.16)$$

³N. Mayo, "Ocean Waves—Their Energy and Power," *Physics Teacher*, 35, 352 (September 1997).

where v is the wave speed, λ the wavelength, d the ocean depth, and g the acceleration due to gravity. If the waves are in *deep* water ($d > \lambda/2$), then the hyperbolic tangent in Eq. (9.16) is approximately 1 (i.e. $\tanh x \approx 1$ for large x), and this reduces to

$$v \approx \sqrt{\frac{g\lambda}{2\pi}} \quad (\text{deep waves, } d > \lambda/2). \quad (9.17)$$

On the other hand, for *shallow* waves ($d < \lambda/20$), the hyperbolic tangent in Eq. (9.16) reduces to its argument (i.e. $\tanh x \approx x$ for small x), and we have

$$v \approx \sqrt{gd} \quad (\text{shallow waves, } d < \lambda/20). \quad (9.18)$$

Tsunami waves are waves created by earthquakes. They are unlike normal ocean waves; they have very long wavelengths (often exceeding 100 km or 60 miles), and they travel at very high speed (typically well in excess of 500 miles per hour, depending on depth and wavelength) (Eq. (9.16)). The amplitude of a tsunami wave is very small while the wave is in the deep ocean; a tsunami may pass under a ship without the passengers even noticing. But when it enters shallow water near shore, a tsunami wave decreases in both speed and wavelength, resulting in a very destructive wave of very large amplitude.

9.10 Seismic Waves

An example of waves encountered in nature is *seismic waves*, which are waves in the Earth's crust and interior that are produced by earthquakes. Geologists have observed two types of seismic waves that travel in the interior of the Earth:

- *P waves* (for “primary” or “pressure” waves) are longitudinal waves, and can travel in both the solid and liquid parts of the interior of the Earth.
- *S waves* (for “secondary” or “shear” waves) are transverse waves, and can travel only in the solid parts of the Earth.

The S waves are the slower of the two; they travel at about 60% of the speed of P waves. This is actually why they are called “primary” and “secondary” waves: the P waves, being faster, arrive first at a seismic observing station. P waves travel with a speed that varies from less than 5 km/s at the Earth's crust to about 13 km/s through the core. From the time delay between the arrival of the P waves and S waves, a seismic observing station may infer the distance to the earthquake's *epicenter* (the point on the Earth's surface directly above the point of origin of the earthquake). Measurements from several observing stations allow a determination of the position of the epicenter through triangulation.

Also, since S waves cannot travel through liquids, observing seismic waves has allowed geologists to infer something about the structure of the interior of the Earth—for example, that there the core consists of a solid *inner core*, surrounded by a liquid *outer core*.

In addition to P waves and S waves, geologists have observed two types of waves that propagate only at the surface of the Earth's crust: *Rayleigh waves* ripple along the Earth's surface like water waves, and *L waves* (or *Love waves*) are a kind of transverse wave whose displacement is in the plane of the Earth's surface.

Seismic wave energy is measured on a logarithmic scale called the *moment magnitude scale*. An earthquake energy of E joules is said to have a magnitude M given by

$$M = \frac{2}{3} \log_{10} E - 6.0. \quad (9.19)$$

For small to medium earthquakes, this moment magnitude scale gives numbers close to those on the older Richter scale that it replaces.

Chapter 10

Standing Waves

Supposes you attach one end of a string to a wall, and hold the other end in your hand. Now give your end a quick “flip”, and you will see a wave pulse travel down to the wall, get inverted, and the reflected wave will come back to you.

Now suppose you set up a continuous *wave train* at your end. The waves will travel to the wall, get inverted and reflected back toward you. On the way back, they will interfere with the waves coming in the opposite direction, and you will get a complicated-looking jumble of interfering waves.

But suppose you time things just right, with just the right frequency, so that the returning reflected waves interfere constructively with the waves coming the other way. In this case the waves all add together nicely, and you get a pattern of *standing waves*. Standing wave patterns look like the patterns in Fig. 10.1; you’ll see a set of “segments” vibrating up and down, where each segment is a half wavelength. At the points between segments, the string does not move at all; these points are called the *nodes*. Halfway between the nodes are the points of maximum displacement; these are the *antinodes*.

It’s important to realize that if you drive one end of the string with simple harmonic motion, you will, in general, not get standing waves—you’ll get a jumbled mess at first, that will eventually settle into non-standing waves that oscillate at the forcing frequency. Only at certain specific frequencies will you get standing waves.

So what frequencies will give standing waves? That depends on whether the string is fixed at both ends, or just one end, or if both ends are free.

10.1 Fixed or Free at Both Ends

If the string is fixed at both ends and the ends are a distance L apart, then you can see from examining Fig. 10.1 that an integer number of segments have to fit into the distance L . Since each segment is a half wavelength, the condition for standing waves in this case is that an integer number of half-wavelengths must fit into length L :

$$L = n \frac{\lambda}{2} \quad (n = 1, 2, 3, 4, \dots) \quad (10.1)$$

Now since the wave speed $v = f\lambda$, we can substitute for λ and solve for f to find an expression for the frequencies that give rise to standing waves:

$$f_n = n \frac{v}{2L} \quad (n = 1, 2, 3, 4, \dots) \quad (10.2)$$

As shown in Fig. 10.1, there is a sequence of standing waves, one pattern for each integer $n = 1, 2, 3, 4, \dots$. The standing wave f_1 is called the *first harmonic*; the next one (f_2) is called the *second harmonic*, and so

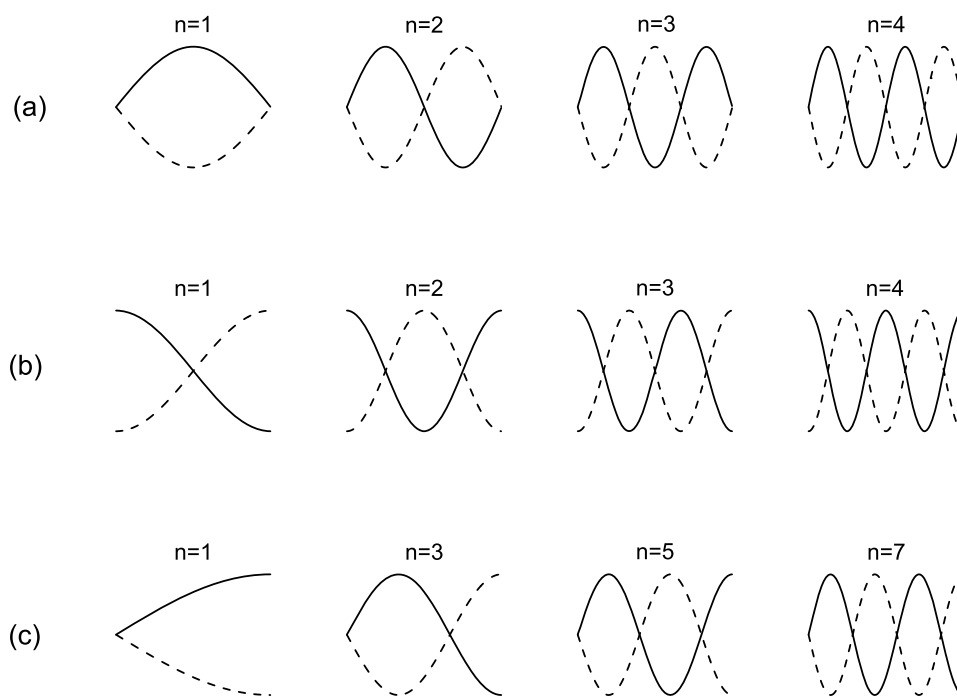


Figure 10.1: The first four standing waves in a string (a) fixed at both ends; (b) free at both ends; (c) fixed on the left end and free on the right. The stationary points with no displacement are the *nodes*; in between them are the points of maximum displacement, the *antinodes*.

on, so f_n is the n -th harmonic. (Sometimes a different nomenclature is used: f_1 is called the *fundamental frequency*, f_2 is called the *first overtone*, f_3 is the *second overtone*, and so on, so f_n is the $(n-1)$ -th overtone.)

It turns out (as you can see from examining Fig. 10.1(b)) that this same condition (Eq. 10.2) also applies to waves that are *free* at both ends: an integer number of half-wavelengths must fit into length L .

10.2 Fixed at One End and Free at the Other

A different situation occurs when the wave is fixed at one end and free at the other (Fig. 10.1(c)). From examining the figure, you can see the pattern: an odd number of half-segments has to fit into distance L . Since each segment is a half wavelength, this means that an odd number of quarter-wavelengths must fit into length L :

$$L = n \frac{\lambda}{4} \quad (n = 1, 3, 5, 7, \dots) \quad (10.3)$$

Again using the relation $v = f\lambda$ and solving for f , we find the condition for standing waves in this case is

$$f_n = n \frac{v}{4L} \quad (n = 1, 3, 5, 7, \dots) \quad (10.4)$$

Although we've been talking about string waves, this analysis refers to both transverse and longitudinal waves (sound waves, for example). As we'll see later, musical instruments work by creating standing sound waves which satisfy these same conditions.

10.3 Vibrations of Rods and Plates

A rod may be set vibrating (longitudinally) by holding or clamping it at some point and stroking it with rosin. There will be a node at the point where the rod is clamped, and antinodes at each end. For example, clamping the rod at its center point will create standing waves free at both ends (where there are antinodes) and fixed in the center (where the rod is clamped), resulting in an $n = 1$ standing wave, as shown in Figure 10.1(b), $n = 1$. Clamping the rod at $1/4$ its length from one end again creates a node at the clamped point and antinodes at the two ends, resulting in an $n = 2$ standing wave (Figure 10.1(b), $n = 2$).

Standing waves can also be created in two-dimensional plates or membranes. Figure 10.2 shows the standing wave modes of a circular membrane such as a drum head. Notice in this case that the frequencies of the standing wave modes are *not* integer multiples of the fundamental frequencies, so they are *not* harmonics.

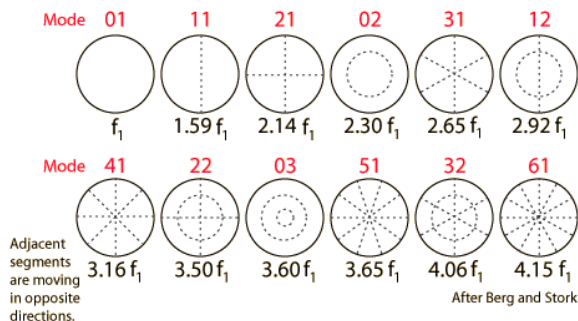
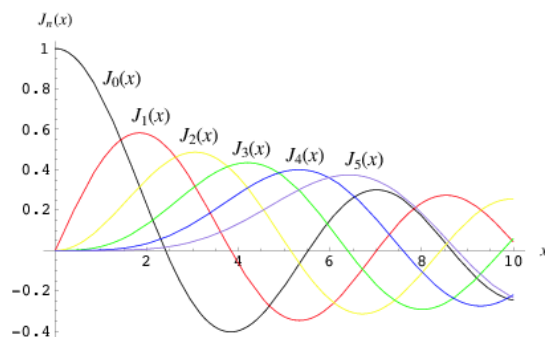


Figure 10.2: Modes of vibration of a circular membrane, showing nodal lines. (Figure from D. Livelybrooks, Univ. of Oregon.)

Figure 10.3: Bessel functions $J_m(x)$. (Credit: Wolfram MathWorld.)

For the circular membrane, the vibration modes are characterized by *two* integers, m and n . The frequency of mode mn is given by

$$f_{mn} = \frac{\alpha_{mn}}{\alpha_{01}} f_{01}, \quad (10.5)$$

where α_{mn} is the n -th zero of a special function called the *Bessel function* $J_m(x)$ (Figure 10.3). In other words, α_{mn} is the value of x at the n -th time the function $J_m(x)$ crosses the x axis for $x > 0$.

The first few zeros of the first few Bessel functions are given in Table 10-1.

Table 10-1. Zeros α_{mn} of the Bessel functions $J_m(x)$. (Credit: Wolfram MathWorld.)

n	$J_0(x)$	$J_1(x)$	$J_2(x)$	$J_3(x)$	$J_4(x)$	$J_5(x)$
1	2.4048	3.8317	5.1356	6.3802	7.5883	8.7715
2	5.5201	7.0156	8.4172	9.7610	11.0647	12.3386
3	8.6537	10.1735	11.6198	13.0152	14.3725	15.7002
4	11.7915	13.3237	14.7960	16.2235	17.6160	18.9801
5	14.9309	16.4706	17.9598	19.4094	20.8269	22.2178

Example. The frequency of mode $m = 3$, $n = 2$ is

$$f_{32} = \frac{\alpha_{32}}{\alpha_{01}} f_{01} = \frac{9.7610}{2.4048} f_{01} = 4.0589 f_{01}.$$

Part III

Acoustics

Chapter 11

Sound

Sound consists of longitudinal waves that propagate through some medium and may be detected by the human ear. We often think of sound waves as propagating through the air, but sound waves may also move through other materials like helium, water, or steel. In this chapter we'll examine a few of the basic properties of sound waves.

11.1 Speed of Sound

First of all, how fast do sound waves travel? You've probably noticed that sound waves have a noticeable travel time—for example, when you're watching a baseball game far from home plate, there is a definite delay between *seeing* a batter hit the ball, and *hearing* the sound. Experimentally, we find the nominal speed of sound in air to be (at 20°C)

$$v_{\text{snd}} = 343 \text{ m/s} \quad (11.1)$$

It turns out that the speed of sound is strongly dependent on temperature. An empirical formula that corrects for this temperature variation gives the speed of sound in air as

$$v_{\text{snd}} \approx (331 + 0.60T_c) \text{ m/s}, \quad (11.2)$$

where T_c is the air temperature, in °C. Notice that if $T_c = 20^\circ\text{C}$, we get 343 m/s.

If we convert units, we find that this is equal to about $\frac{1}{5}$ mile per second. This gives the rule you may have learned in childhood for estimating the distance of a lightning flash: after you see the lightning, count how many seconds go by before you hear the thunder, then divide by 5 to find how many miles away the lightning was. (Light travels about 900,000 times faster than sound, so the lightning reaches you almost instantly, and you don't need to consider the light travel time.)

What about the speed of sound in other materials? Recall from Eq. (9.4) that the speed of waves in a string is the square root of the tension divided by the density: $v = \sqrt{F_T/(m/L)}$. The speed of sound waves in fluids follows a similar formula, known as the *Newton-Laplace equation*:

$$v_{\text{snd}} = \sqrt{\frac{B}{\rho}}, \quad (11.3)$$

where B is called the *bulk modulus* of the material (a measure of its compressibility), and ρ is the density of the material. Table 11-1 shows the bulk moduli, densities, and speeds of sound for several different fluids. (For the speed of sound in *solids*, you use the *Young's modulus* Y in place of the bulk modulus B : $v_{\text{snd}} = \sqrt{Y/\rho}$.)

Table 11-1. Speed of sound in several fluids. (All data are for 20°C.)

Medium	Bulk modulus B (Pa)	Density ρ (kg/m ³)	Speed of sound
Air	1.42×10^5	1.204	343
Helium	1.69×10^5	0.1663	1008
SF ₆	1.35×10^5	6.069	149
Water	2.2×10^9	1000	1497

A common laboratory demonstration is to inhale some helium gas and then try to talk; the amusing result is an abnormally high-pitched voice. The opposite effect can be demonstrated by inhaling sulfur hexafluoride (SF₆), which results in an abnormally low voice. (You should *not* attempt to do this yourself, as both demonstrations are potentially dangerous.) As you can see from the table, all three gases have similar bulk moduli; they differ mainly by their densities, which results in different speeds of sound for each gas. It is these differences in the sound speed that is responsible for the high and low pitches of one's voice in each gas.

The bulk modulus and density of a gas are also functions of temperature. We can find the an explicit expression for the speed of sound in a gas as a function of temperature as follows: the bulk modulus B of an ideal gas is given by

$$B = \gamma p, \quad (11.4)$$

where p is the pressure of the gas, and γ is the ratio of the heat capacity at constant pressure (C_P) to the heat capacity at constant volume (C_V):

$$\gamma = \frac{C_P}{C_V}. \quad (11.5)$$

It can be shown from thermodynamics that:

- For a monatomic gas: $\gamma = \frac{5}{3} = 1.67$
- For a diatomic gas, or other gas with linear molecules: $\gamma = \frac{7}{5} = 1.40$
- For a gas with nonlinear molecules: $\gamma = \frac{4}{3} = 1.33$

Now substituting Eq. (11.4) into the Newton-Laplace equation (11.3), we have

$$v_{\text{snd}} = \sqrt{\frac{\gamma p}{\rho}}. \quad (11.6)$$

Now using the ideal gas law

$$pV = Nk_B T \quad (11.7)$$

(where V is the volume of gas, N is the number of atoms or molecules of gas, $k_B = 1.3806488 \times 10^{-23}$ J K⁻¹ is the Boltzmann constant, and T is the absolute temperature in kelvins) to substitute for the pressure p , we have

$$v_{\text{snd}} = \sqrt{\frac{\gamma N k_B T}{\rho V}}. \quad (11.8)$$

Now ρV is the total mass of gas, which we'll call m , so we have

$$v_{\text{snd}} = \sqrt{\frac{\gamma N k_B T}{m}}. \quad (11.9)$$

The mass per atom (or molecule) is $m_a = m/N$, so we have

$$v_{\text{snd}} = \sqrt{\frac{\gamma k_B T}{m_a}} = \sqrt{\frac{\gamma R T}{M}}, \quad (11.10)$$

where $R = k_B N_A = 8.3144621 \text{ J mol}^{-1} \text{ K}^{-1}$ is the molar gas constant and N_A is Avogadro's number, and $M = m/(N/N_A)$ is the molar mass of the gas (kilograms per mole). Since the molecular (or atomic) weight is in grams per mole, this means that M is just the molecular (or atomic) weight divided by 1000.

Using Eq. (11.10), we can see where the empirical relation for the speed of sound in air, Eq. (11.2), comes from. Air consists of about 78% nitrogen (N_2), 21% oxygen (O_2), and 1% argon (Ar). Since the gases are mostly diatomic, we will take $\gamma = 1.40$. To find the mass per molecule, we'll compute a weighted average based on composition. Since N_2 has a molecular weight of 28, O_2 has a molecular weight of 32, and Ar has an atomic weight of 40, we compute the weighted average molecular weight of air to be

$$m_a = (0.78 \times 28) + (0.21 \times 32) + (0.01 \times 40) = 28.96. \quad (11.11)$$

To convert this to mass in kilograms, we multiply this by the atomic mass unit $u = 1.660538921 \times 10^{-27} \text{ kg}$ to get $m_a = 4.8089 \times 10^{-26} \text{ kg}$. Substituting these results into Eq. (11.10), we get

$$v_{\text{snd}} = \sqrt{\frac{\gamma k_B T}{m_a}} \quad (11.12)$$

$$= \sqrt{\frac{(1.40)(1.3806488 \times 10^{-23} \text{ J/kg})T}{4.8089 \times 10^{-26} \text{ kg}}} \quad (11.13)$$

$$= 20.0472 \sqrt{T} \quad (11.14)$$

in SI units. Now T is the absolute temperature, and let's let T_c be the temperature in degrees Celsius. Since the two are related by $T = T_c + 273.15$, we have

$$v_{\text{snd}} = 20.0472 \sqrt{T_c + 273.15} \quad (11.15)$$

$$= (20.0472) \sqrt{273.15} \sqrt{\frac{T_c}{273.15} + 1} \quad (11.16)$$

$$= 331.32 \sqrt{1 + \frac{T_c}{273.15}} \quad (11.17)$$

We now use the series expansion (valid for $|x| < 1$; see Appendix D)

$$(1+x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{128}x^4 + \frac{7}{256}x^5 - \dots \quad (11.18)$$

$$\approx 1 + \frac{1}{2}x \quad (11.19)$$

and we have

$$v_{\text{snd}} \approx 331.32 \left(1 + \frac{1}{2} \frac{T_c}{273.15} \right) \quad (11.20)$$

$$= 331.32 + 0.6065 T_c \quad (11.21)$$

$$\approx 331 + 0.60 T_c \quad (11.22)$$

and we have just derived the empirical relation, Eq. (11.2).

11.2 Frequency of Sound

Sound frequencies may be divided into the following categories, depending on whether or not they are within the range of human hearing:

- *Infrasonic.* ($f < 20$ Hz) These sounds are at frequencies too low to be audible to humans.
- *Audible.* ($20 \text{ Hz} \leq f \leq 20,000$ Hz) This is the range of human hearing.
- *Ultrasonic.* ($f > 20,000$ Hz) These sounds are at frequencies too high to be audible by humans.

These are only approximate ranges. In particular, there is a strong correlation between the highest audible frequency and the person's age; as we get older, we become less able to hear very high-frequency sounds.

Infrasonic sounds are inaudible to humans, but can be heard (and produced) by some animals like whales and elephants. Some natural phenomena like earthquakes also produce infrasonic sounds.

Ultrasonic sounds are also inaudible to humans, but can be heard by some other animals, like dogs, bats, and dolphins. The familiar dog whistle produces a high-pitched sound that is inaudible to humans, but can be heard by dogs. Ultrasound has several practical uses: it is used in some cleansing processes, and for medical imaging.

Table 11-2 shows the hearing ranges (frequencies) audible to different animals.

Table 11-2. Hearing ranges for various animals. [6]

Species	Hearing range (Hz)
Turtle	20 – 1,000
Goldfish	100 – 2,000
Frog	100 – 3,000
Pigeon	200 – 10,000
Sparrow	250 – 12,000
Human	20 – 20,000
Chimpanzee	100 – 20,000
Rabbit	300 – 45,000
Dog	50 – 46,000
Cat	30 – 50,000
Guinea pig	150 – 50,000
Rat	1,000 – 60,000
Mouse	1,000 – 100,000
Bat	3,000 – 120,000
Dolphin (<i>Tursiops</i>)	1,000 – 130,000

Chapter 12

The Doppler Effect

You have probably noticed that the frequency of sound emitted by a moving source depends on its speed; for example, when you're standing by the side of a road near fast-moving traffic, the engine sounds decrease in frequency as the car passes you. (This is especially noticeable at the Indianapolis 500, for example.) This effect is called the *Doppler effect*, after Christian Doppler, an Austrian physicist who first described the effect in the 19th century.

This change in frequency is observed whether the source or the observer is moving. If the source and observer are getting closer together, the frequency is *higher* than if both were stationary; if they are getting farther apart, the frequency is *lower*.

A little thought reveals why this is. If the *source* of the sound is moving toward a stationary observer, then the source will have moved in between emitting wave fronts, causing effective wavelength to be shorter, resulting in a higher frequency heard by the observer. On the other hand, if the *observer* of the sound is moving toward a stationary source, then the observer runs into the wavefronts faster than if he were stationary, so he hears a higher frequency.

The frequency shift may be described by the following equation, which covers either the source or the observer moving (or both):

$$f' = f \left(\frac{v_{\text{snd}} \pm v_{\text{obs}}}{v_{\text{snd}} \mp v_{\text{source}}} \right). \quad (12.1)$$

Here f is the frequency emitted by the source, and f' is the frequency heard by the observer. Three speeds go into this equation, and they are all measured with respect to the air: v_{snd} is the speed of sound (nominally 343 m/s); v_{obs} is the speed of the observer, and v_{source} is the speed of the source of the sound. All of these speeds are taken to be positive; the directions are taken into account with the $\pm$ and $\mp$ signs. The rule for using these signs is:

“Top sign toward, bottom sign away.”

In other words, if the source and observer are moving toward each other, we use the top signs: $+$ in the numerator and $-$ in the denominator. If they are moving away from each other, we use the bottom signs: $-$ in the numerator and $+$ in the denominator. To be fully explicit:

- If the observer is moving *toward* the source, use $+$ in the numerator.
- If the observer is moving *away* from the source, use $-$ in the numerator.
- If the source is moving *toward* the observer, use $-$ in the denominator.
- If the source is moving *away* from the observer, use $+$ in the denominator.

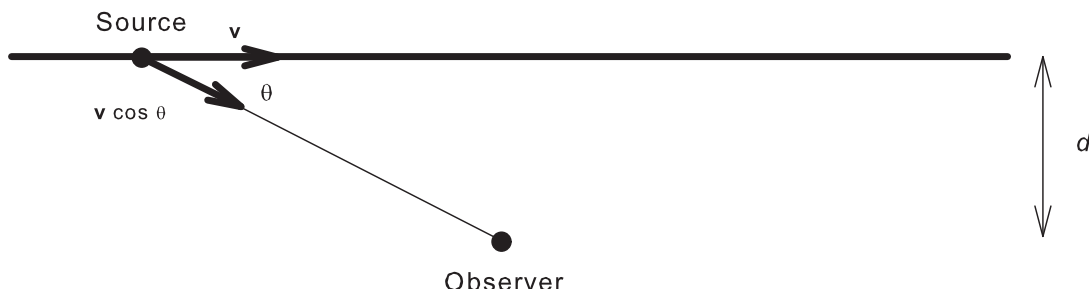


Figure 12.1: Doppler shift for a source moving at an angle relative to the observer. Here the source is moving along a straight line track with velocity $\mathbf{v}$, and the observer is standing off to one side of the track. For the purpose of computing the Doppler shift, the effective speed of the source is the component of $\mathbf{v}$ in the direction of the observer, $v \cos \theta$.

For example, suppose a fire engine emits a sound with a frequency of 2000 Hz, and is moving directly toward you at 50 m/s. You are stationary. What frequency do you hear? In this case $f = 2000$ Hz, $v_{\text{snd}} = 343$ m/s, $v_{\text{source}} = 50$ m/s, and $v_{\text{obs}} = 0$. Since the fire engine is moving toward you, you choose the top signs, so the frequency you hear is $f' = (2000 \text{ Hz})[(343 + 0)/(343 - 50)] = 2341$ Hz.

Eq. (12.1) covers the case where the source and observer are moving *directly* toward or away from each other. But what if they are moving at some angle relative to each other, rather than directly toward or away from each other? In that case, the velocities v_{obs} and v_{source} that you use in Eq. (12.1) are the *components* of the velocity along a line connecting the source and the observer. Fig. 12.1 shows an example: a source of sound is moving along a straight track, and the observer is standing off to one side. At any point, the v_{source} we use in Eq. (12.1) is the component of the source's velocity along the line connecting the source to the observer at that point, or $v_{\text{source}} \cos \theta$. If we perform this calculation using $v_{\text{source}} = 50$ m/s and $d = 10$ m for each point along the track, we get the plot shown in Fig. 12.2.

It is a common misconception that in the case of a moving source, the frequency increases as the object moves toward the observer, and decreases as it moves away. As you can see from Fig. 12.2, this is not the case: the frequency decreases monotonically.

12.1 Relativistic Doppler Effect

Light waves exhibit the Doppler effect just as sound waves do, but the analysis is different. We'll examine light waves in more detail later, but for now we can just note that light waves are a type of transverse wave that can travel through a vacuum. In discussing the Doppler effect for sound, we specified the speeds of both the source and the observer relative to the reference frame of the *air*. However, there is no such reference frame for light waves. According to Einstein's special theory of relativity, there is no "universal" reference frame with respect to which we can measure speeds of bodies—and furthermore, the theory says that the speed of light is constant, regardless of the speed of the person making the measurement. So in the case of light waves, it makes no sense to talk about the speeds of the source or the observer with respect to some fixed reference frame, since there is no such frame—we can only talk about the speeds of the source and observer *relative to each other*. This means that the Doppler shift equation for light has only two speeds in it: the speed of light c , and the relative speed between the source and observer, v . The Doppler equation for light waves (called

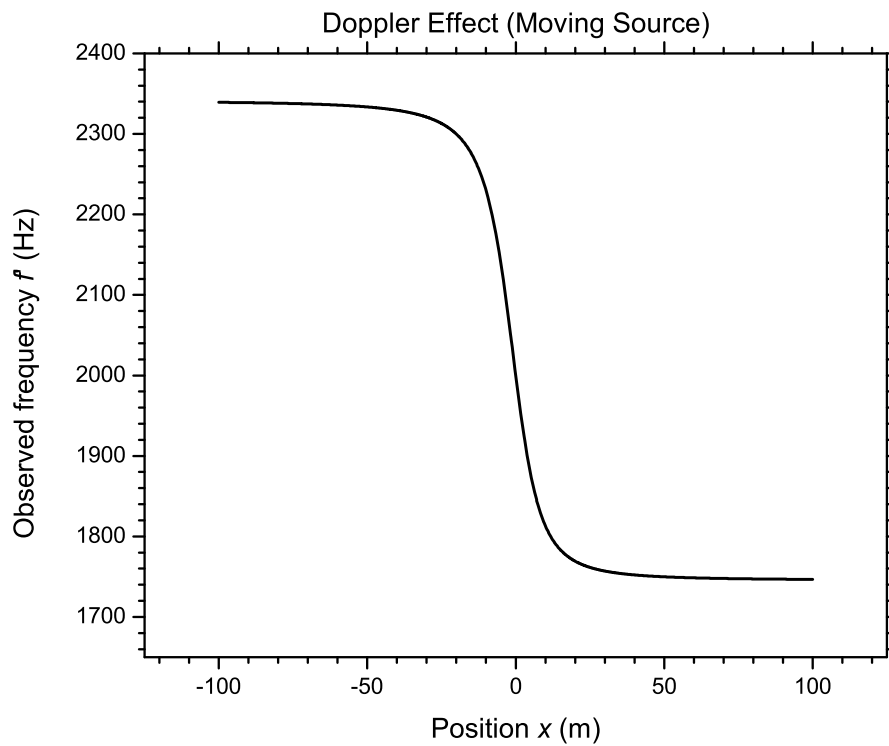


Figure 12.2: Doppler shift for a moving source. In this example, the source is moving at a speed of 50 m/s along a straight line, the stationary observer is a perpendicular distance of 10 m from the source's path at $x = 0$, and the frequency of the sound emitted by the source is 2000 Hz.

the *relativistic* Doppler equation) is

$$f' = f \sqrt{\frac{c \pm v}{c \mp v}}, \quad (12.2)$$

where the sign conventions are the same as for the Doppler effect described earlier. This effect means that if the source and observer of light waves are moving *toward* each other, the light waves appear *bluer* than they would if the source and observer were not moving relative to each other; this effect is called a *blueshift*. Similarly, if the source and observer are moving away from each other, the light appears redder than it would otherwise, an effect called a *redshift*.

Astronomers often observe this effect in astronomical bodies. For example, because of the Sun's rotation, lines in the Sun's spectrum are blueshifted on the edge of the Sun moving toward us, and redshifted on the edge moving away from us.

It was discovered decades ago that all distant galaxies have redshifted light, so they are all moving away from us. Furthermore, the farther the galaxy, the greater the redshift—meaning that the farther the galaxy, the faster it's moving away from us. The American astronomer Edwin Hubble first noted this, and postulated what is now called *Hubble's law*; it relates the speed v with which a galaxy is moving away from us to its distance D from us:

$$v = H_0 D, \quad (12.3)$$

where H_0 is a proportionality constant called the *Hubble constant*. Observations by several NASA spacecraft have recently determined the value of the Hubble constant to be about $H_0 = 71$ (km/s)/Mpc. (A *parsec* (pc) is about 3.26 light-years, or about 3.09×10^{16} meters, and so a *megaparsec* (Mpc) is a million times that.)

Why are all the galaxies moving away from us like this? It's because the Universe is expanding, which is causing every distant galaxy to move away from every other one, much like dots drawn on a balloon moving farther apart as the balloon is inflated. This expansion began 13.7 billion years ago with the Big Bang, the huge explosion in which the Universe was created, and is continuing to this day.

Chapter 13

Sound Intensity

13.1 Intensity

Let's now look at another property of sound: its *loudness*. The loudness of sound is just the intensity of the sound waves, in watts per square meter. The sound intensity I is the power P of the sound source (e.g. a loudspeaker), divided by the area over which this power is spread. For example, if the source of sound waves is an isotropic point source, then spherical sound waves are emitted equally in all directions. At a distance r from the source, the source's power will be spread over the surface of a sphere of radius r , so the sound intensity at distance r will be

$$I = \frac{P}{A} = \frac{P}{4\pi r^2}. \quad (13.1)$$

13.2 Decibels

Our ears are capable of hearing sounds over a tremendous range of intensities. It has been said that if our ears were any more sensitive than they are, we would be able to hear the sound of individual air molecules hitting our eardrums. But we can also hear very loud sounds, like from a jet engine. In order to accommodate this large range of intensities, our ears tend to respond *logarithmically* to sounds; this has motivated the creation of a *logarithmic* loudness scale, where sound level is proportional to the logarithm of the intensity.

Simply taking the logarithm of the intensity doesn't work dimensionally, though—when you take the logarithm of a quantity, it should be dimensionless. We therefore take the logarithm of a *ratio* of intensities to get the *sound level*:

$$B = \log_{10} \frac{I}{I_0}, \quad (13.2)$$

where B is the sound level in units of *bels* (B) (named after Alexander Graham Bell), I is the sound intensity, and $I_0 = 10^{-12} \text{ W/m}^2$ is called the *threshold of hearing*, and is roughly the lowest-intensity sound that an average person can hear. The *actual* softest audible sound varies from person to person, changes with age, and is also a function of frequency. But for the purpose of defining the bel, we always use 10^{-12} W/m^2 for I_0 . Also, notice that by convention, the *common* (base 10) logarithm is used in defining the bel.

In practice, the bel is rarely used; the more common unit is $\frac{1}{10}$ bel, or the *decibel* (dB). The sound level in decibels (β) is given by

$$\beta = 10 \log_{10} \frac{I}{I_0}. \quad (13.3)$$

The threshold of hearing $I = I_0$ corresponds to a sound level of 0 dB. A sound intensity of 1 W/m^2 corresponds to a sound level of 120 dB, and is about where most people start finding the sound to be painfully loud; 120 dB is called the *threshold of pain*.

One useful fact to note about decibels is that each time you *double* the intensity I (W/m^2), you *add* 3 dB to the sound level. This is because in going from intensity I to $2I$, the sound level becomes

$$\beta' = 10 \log_{10}(2I/I_0) \quad (13.4)$$

$$= 10 \log_{10} 2 + 10 \log_{10}(I/I_0) \quad (13.5)$$

$$\approx 3.010 + \beta. \quad (13.6)$$

Similarly, when you *halve* the sound intensity I , you *subtract* 3 dB from the sound level.

When computing sound levels, you cannot do the computations in decibel units. Instead, you must do the calculations in intensity units (W/m^2), then convert to dB at the end. For example, suppose you are 35 meters away from a 10-watt isotropic sound source. How loud a sound do you hear? You first find the intensity: $I = P/(4\pi r^2) = (10 \text{ W})/[4\pi(35 \text{ m})^2] = 6.496 \times 10^{-4} \text{ W/m}^2$. Now convert the result to dB to find the sound level: $\beta = 10 \log_{10}(I/I_0) = 10 \log_{10}(6.496 \times 10^{-4}/10^{-12}) = 88 \text{ dB}$.

13.3 Nepers

A less common unit for measuring sound level is the *neper* (Np). Like the decibel, the neper is a logarithmic scale; but unlike the decibel, it is a measure of the ratio of *amplitudes* (not intensities), and uses the natural logarithm instead of the common logarithm. Since the amplitude is proportional to the square root of the intensity ($A \propto \sqrt{I}$), the sound level γ in nepers is given by $\gamma = \ln \sqrt{I/I_0}$, or

$$\gamma = \frac{1}{2} \ln \frac{I}{I_0}. \quad (13.7)$$

You may convert between decibels and nepers using the relationship

$$\gamma \text{ (Np)} = \beta \text{ (dB)} \times \frac{\ln 10}{20} \quad (13.8)$$

Every doubling of the intensity I (W/m^2) corresponds to adding about $\frac{1}{3}$ Np to the sound level, and halving I means subtracting about $\frac{1}{3}$ Np from the sound level.

In terms of nepers, the threshold of hearing = 0 dB = 0 Np; the threshold of pain = 120 dB = 14 Np.

Chapter 14

Music

Music is a sequence of sounds created for enjoyment or artistic expression. The sounds may be produced by the human voice (singing), or by any number of musical instruments. Music is a vast field, and we can only hope to touch on some of the most basic ideas of music theory here; the interested reader is referred to the references in Appendix A for more information.

14.1 Pitch

To begin, music consists of a sequence of sounds of short duration (called *notes*); each of these notes is at a specific frequency (called *pitch*). Not just any frequencies are used, though; musical notes are selected from a set of discrete frequencies.

We find that if we hear a sound at frequency f , then to our ears a sound at twice that frequency ($2f$) sounds “similar”, but higher. To get musical notes, the interval between frequency f and $2f$, known as an *octave*, is divided into twelve equal parts (in a logarithmic sense) so that each note is higher in frequency than the next lower note by a factor of $\sqrt[12]{2} \approx 1.059463$. Each factor of $\sqrt[12]{2}$ change in frequency is called a *half step*, and two half steps make a *whole step*. The complete set of 12 notes in an octave (each separated in pitch by a half step) is called the *chromatic scale*.

Early musicians discovered that musical compositions sounded better when they used only certain subsets of these 12 notes, rather than all 12. One of the best-known of these subsets (or *scales*) consists of 7 of the 12 notes in an octave; these notes were named (in order of increasing pitch) C, D, E, F, G, A, and B. In this scale, called the *C major scale*, notes B and C (of the next octave) are one half step apart in frequency, as are notes E and F; the others are a whole step apart.

Each octave contains the 12 notes in the chromatic scale, which are given the following names, in order of increasing pitch:

Table 11-1. The musical notes.

C	F $\sharp$ / G $\flat$
C $\sharp$ / D $\flat$	G
D	G $\sharp$ / A $\flat$
D $\sharp$ / E $\flat$	A
E	A $\sharp$ / B $\flat$
F	B

In this table we find the seven notes of the C major scale, along with the remaining five notes, which are named using the symbols $\sharp$ and $\flat$ to indicate that they fall in between the notes of the C major scale. The symbol $\sharp$ (called “sharp”) indicates a raising in pitch by one half step over the note to which it is attached; similarly, the symbol $\flat$ (called “flat”) indicates a lowering of pitch by one half step. For example, $C\sharp$ is one half step higher in pitch than C, and $B\flat$ is one half step lower in pitch than B. (The symbols $\sharp$ and $\flat$ are collectively called *accidentals*.)

Notice that several notes are known by two equivalent names. For example, $C\sharp$ and $D\flat$ refer to the same note—the one between notes C and D. Also, since notes B and C are separated by just one half step, we have $B\sharp = C$ and $C\flat = B$; similarly, E and F are separated by one half step, so $E\sharp = F$ and $F\flat = E$.

When it is necessary to indicate a specific octave, it is written as a subscript after the note. The note A_4 (near the middle of the piano keyboard) is assigned a frequency of exactly 440 Hz. Since the notes in each octave have twice the frequency of the same note in the next lower octave, we find the frequencies of note A in higher octaves by repeatedly multiplying by 2: $A_5 = 880$ Hz, $A_6 = 1760$ Hz, $A_7 = 3520$ Hz, $A_8 = 7040$ Hz, and $A_9 = 14080$ Hz. Similarly for A in lower octaves, we repeatedly divide 440 Hz by 2: $A_3 = 220$ Hz, $A_2 = 110$ Hz, $A_1 = 55$ Hz, and $A_0 = 27.5$ Hz. Human hearing covers ten octaves in pitch, going roughly from note E_0 to E_{10} . The piano's range is $7\frac{1}{4}$ octaves, from A_0 to C_8 .

Beginning with the frequency of note $A_4 = 440$ Hz, we successively multiply and divide by $\sqrt[12]{2}$ to find the frequencies of all the other notes, as shown in Table 14-2.

Table 14-2. Frequencies (in hertz) of all the musical notes that are audible to the human ear. Middle C is shown in bold, and the musical standard A_4 is shown in italics.

Note	Octave										
	0	1	2	3	4	5	6	7	8	9	10
C		32.70	65.41	130.81	261.63	523.25	1046.50	2093.00	4186.01	8372.02	16744.04
$C\sharp / D\flat$		34.65	69.30	138.59	277.18	554.37	1108.73	2217.46	4434.92	8869.84	17739.69
D		36.71	73.42	146.83	293.66	587.33	1174.66	2349.32	4698.64	9397.27	18794.55
$D\sharp / E\flat$		38.89	77.78	155.56	311.13	622.25	1244.51	2489.02	4978.03	9956.06	19912.13
E	20.60	41.20	82.41	164.81	329.63	659.26	1318.51	2637.02	5274.04	10548.08	21096.16
F	21.83	43.65	87.31	174.61	349.23	698.46	1396.91	2793.83	5587.65	11175.30	
$F\sharp / G\flat$	23.12	46.25	92.50	185.00	369.99	739.99	1479.98	2959.96	5919.91	11839.82	
G	24.50	49.00	98.00	196.00	392.00	783.99	1567.98	3135.96	6271.93	12543.85	
$G\sharp / A\flat$	25.96	51.91	103.83	207.65	415.30	830.61	1661.22	3322.44	6644.88	13289.75	
A	27.50	55.00	110.00	220.00	<i>440.00</i>	880.00	1760.00	3520.00	7040.00	14080.00	
$A\sharp / B\flat$	29.14	58.27	116.54	233.08	466.16	932.33	1864.66	3729.31	7458.62	14917.24	
B	30.87	61.74	123.47	246.94	493.88	987.77	1975.53	3951.07	7902.13	15804.27	

In general, a note n half steps above A_4 has a frequency of

$$2^{n/12} \times 440 \text{ Hz}, \quad (14.1)$$

where n is negative for notes below A_4 .

Note C_4 (in the middle of the piano keyboard) is called *middle C*. Since it's 9 half steps below A_4 , middle C has a frequency of $2^{-9/12} \times 440 \text{ Hz} = 261.6256 \text{ Hz}$.

14.2 Musical Scales

As mentioned earlier, early musicians discovered that musical compositions sound best when they don't use all 12 notes of the chromatic scale; instead, restricting the notes to certain subsets of the 12 (called *scales*)

results in more pleasant-sounding music.

In Western music, the most common of these scales are called *major scales*, and the best-known of these is the *C major scale*, which has already been described: it consists of the notes C, D, E, F, G, A, and B. In this scale, the first two notes (C and D) are separated in pitch by a whole step, as are the second and third notes (D and E). The third and fourth notes are separated by a half step. Continuing through the whole scale, we find that the separations between the notes in pitch are two whole steps, then one half step, then three whole steps, then another half step at the end when going from B to C of the next octave. For shorthand, let's write "W" for a whole-step interval between notes, and "H" for a half-step interval; then the intervals between notes in the C major scale can be written as WWHWWWH.

There are 11 other major scales besides the C major scale. To get them, we simply start with a different note in the chromatic scale, then follow the same WWHWWWH interval pattern; the scale is named for the note we started with. For example, for the C $\sharp$ major scale, we begin with C $\sharp$, then go up a whole step in pitch to get the next note in the scale, D $\sharp$. Then we go up another whole step to get the next note, F. Then up a half step to get the next note (F $\sharp$), and so on until we find all seven notes in the scale. Similarly, for the D major scale, we start with the note D and follow the same WWHWWWH pattern to find the seven notes of the D major scale. We can repeat the process for all 12 notes in the chromatic scale; the results are shown in Table 14-3.

Table 14-3. The major scales. The last column shows the number of accidentals in that scale.

Major Scale	Notes							# Acc.
C	C	D	E	F	G	A	B	0
G	G	A	B	C	D	E	F $\sharp$	1 $\sharp$
D	D	E	F $\sharp$	G	A	B	C $\sharp$	2 $\sharp$
A	A	B	C $\sharp$	D	E	F $\sharp$	G $\sharp$	3 $\sharp$
E	E	F $\sharp$	G $\sharp$	A	B	C $\sharp$	D $\sharp$	4 $\sharp$
B (=C $\flat$)	B	C $\sharp$	D $\sharp$	E	F $\sharp$	G $\sharp$	A $\sharp$	5 $\sharp$
F $\sharp$ (=G $\flat$)	F $\sharp$	G $\sharp$	A $\sharp$	B	C $\sharp$	D $\sharp$	E $\sharp$ (=F)	6 $\sharp$
C $\sharp$ (=D $\flat$)	C $\sharp$	D $\sharp$	E $\sharp$ (=F)	F $\sharp$	G $\sharp$	A $\sharp$	B $\sharp$ (=C)	7 $\sharp$
F	F	G	A	B $\flat$	C	D	E	1 $\flat$
B $\flat$	B $\flat$	C	D	E $\flat$	F	G	A	2 $\flat$
E $\flat$	E $\flat$	F	G	A $\flat$	B $\flat$	C	D	3 $\flat$
A $\flat$	A $\flat$	B $\flat$	C	D $\flat$	E $\flat$	F	G	4 $\flat$
D $\flat$ (=C $\sharp$)	D $\flat$	E $\flat$	F	G $\flat$	A $\flat$	B $\flat$	C	5 $\flat$
G $\flat$ (=F $\sharp$)	G $\flat$	A $\flat$	B $\flat$	C $\flat$ (=B)	D $\flat$	E $\flat$	F	6 $\flat$
C $\flat$ (=B)	C $\flat$ (=B)	D $\flat$	E $\flat$	F $\flat$ (=E)	G $\flat$	A $\flat$	B $\flat$	7 $\flat$

Notice that 15 scales are listed in this table; several of them (such as B and C $\flat$) are really the same scale, but with the notes "spelled" differently (recall that some notes have two names, such as A $\sharp$ =B $\flat$), so there are really only 12 different major scales, each one beginning with a different note in the chromatic scale and following the WWHWWWH pattern.

Notice also in Table 14-3 that each major scale can be uniquely identified by the total number of accidentals (sharps and flats) of all the notes in that scales, as shown in the last column. (That's actually the reason for showing the "duplicate" scales in this table, so that this pattern will be clear.) Music written by selecting notes from one of these scales is said to be written in that *key*. For example, a musical composition written using notes selected from the C major scale is said to be written "in the key of C major". This selection of

notes is not strictly adhered to, though; while the notes in a composition are generally selected from the seven in the key being used, the composer may occasionally use other notes for effect.

Since each major key can be uniquely identified by the number of accidentals, the key in which a composition may be indicated by writing the appropriate number of sharps or flats immediately after the clef sign. For example, suppose we wish to write a composition in the key of G major. From Table 14-3, we see that the key of G major contains only one “sharp” note, F $\sharp$. So we indicate a key of G major by writing a single $\sharp$ sign on the F line immediately after the clef sign; this is called the *key signature*. The performer who plays the music will see that the key signature shows a single $\sharp$ on the F line, and will know that the key is therefore G major and that all written F notes should be played as F $\sharp$.

The major scales we’ve just seen are just one of many such scales, each of which gives a different “feel” to the music. For example, there are several *minor scales*; music written in a minor scale has a distinctively dark, “sad” sound to it, and may remind the listener of “spooky” or “funeral” music. There is a *whole tone scale* that is often used for jazz music, and has a whole step between each note in the scale. The *pentatonic scale* is widely used in Eastern music and for many other forms of music around the world.

Table 14-4 shows some of these scales, and their corresponding pitch interval patterns. Remember that each scale shown represents 12 different keys, each one starting with a different note in the chromatic scale, and each one having a bit of a different feel to it.

Table 14-4. Several musical scales and modes, and their pitch interval patterns. (H=half step, W=whole step, 3=three half steps.)

Name	Pattern
Major scale	WWHWWWH
Natural minor scale	WHWWHWW
Harmonic minor scale	WHWWH3H
Melodic minor scale	WHWWWWH
Whole tone scale	WWWWWW
Pentatonic scale	WW3W3
Ionian mode	WHWWHHW
Phrygian mode	HWWHWW
Lydian mode	WWWHWW
Mixolydian mode	WWHWHHW
Aeolian mode	WHWWHWW
Locrian mode	HWWHWW

14.3 Music Notation

Suppose we wish to record a musical composition so that a musician can play it. How do we write out the notes to be played? We could just list the notes to be played (B₄, D₃, etc.), but musicians would find that difficult to read. Also, there needs to be some way to show the *duration* of each note, and to indicate when the performer should pause while playing the composition. To deal with these issues, musicians have developed a special graphical system of musical notation to record music and indicate how it should be played.

The notation begins with five horizontal lines (called a *staff*), which essentially form a plot of frequency vs. time, with increasing frequency (pitch) going up, and increasing time to the right. Each note is written either *on* one of the lines, or in the space *between* lines. A *clef sign* is written at the beginning of the staff to indicate which lines correspond to which notes.

Two clef signs are in common use. A *treble clef* is used when writing music for women's voices, or for instruments that play high notes, generally in the range above middle C. The treble clef sign is a stylized script letter G that curlicues around the line for the note G_4 . Each of the notes in the C major scale (C, D, E, F, G, A, and B) is written on or between lines of the staff; the other notes are written using $\sharp$ and $\flat$ signs next to these notes. Fig. 14.1 shows a treble clef symbol on the far left, followed by ovals (*whole notes*) that show how each of the notes is written for one octave. (Notice that for the first two notes, the staff has been extended downward by a short *ledger line* to write middle C and $C_4\sharp$.)

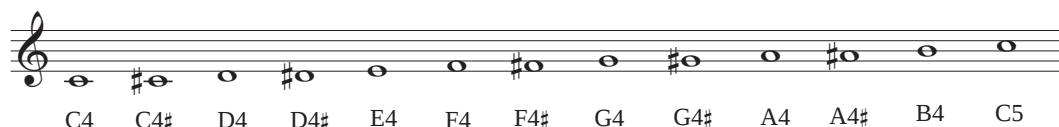


Figure 14.1: Treble clef showing the chromatic scale for octave 4 (plus C_5). The lowest note (far left) is middle C.

A *bass clef* is used for men's voices, or for instruments that play low notes, generally below middle C. The bass clef sign is a stylized script letter F, with two dots on either side of the line for the note F_3 . Again every line or space corresponds to one of the notes in the C major scale, with $\sharp$ and $\flat$ signs written for the other notes. Fig. 14.2 shows a bass clef symbol on the far left, along with whole notes showing each note for one octave.



Figure 14.2: Bass clef showing the chromatic scale for octave 3, plus middle C (C_4) on the far right.

A few other clef signs are in use. For example, viola music is written using an *alto clef*, and there is a *tenor clef* that is common in vocal music. The main point of the different clefs is to shift the notes that are assigned to the lines of the staff in order to minimize the number of ledger lines that are needed. Music is easier to read if most of the notes lie within the five lines of the staff.

The *duration* of each note in time is indicated by various symbols, as shown in Fig. 14.3. A *whole note*, drawn as an oval as shown on the far left, is the longest duration. Other notes are fractions of a whole note, as shown in the figure: a *half note* has half the duration of a whole note, a *quarter note* has one-fourth the duration of a whole note, and so on. So each written note indicates a specific pitch (by its position on the staff) and a specific duration in time (by the symbol used).



Figure 14.3: Note A_4 , showing the symbols for different note durations. The stems may point either upward (as shown here) or downward; generally they point upward for notes near the bottom of the staff, and downward for notes near the top.

Similarly, there are symbols for pauses, or *rests*. Fig. 14.4 shows the various symbols for rests of different

durations. The longest rest (a *whole rest*) is shown on the far left. Next is a *half rest*, which has half the duration of a whole rest, and so on. The rests are always placed on the staff as shown in the figure; they are not drawn higher or lower on the staff, since there is no pitch to be indicated.¹

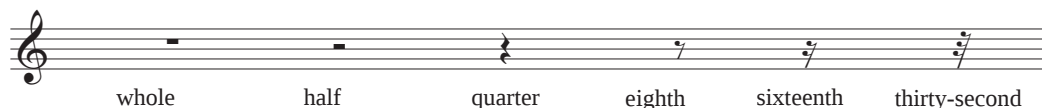


Figure 14.4: Symbols for different rest durations.

14.4 Timing

When we specified the durations of different notes, we specified them relative to the length of a whole note. But what is the duration of a whole note, in seconds? That's not necessarily specified—music may be played faster or slower, so the duration of a whole note is somewhat flexible. But if he wishes, a composer may indicate a specific rate, or *tempo*, at which the music is to be played (typically in units of quarter notes per minute).

For convenience, musical notes are grouped into *measures* of equal time; these are indicated by vertical lines dividing the staff. A *time signature* immediately follows the key signature, and indicates how the timing of the composition works. The time signature is written as something resembling a fraction of the form p/q , where p is the number of “beats” of music per measure, and q indicates which note represents one beat. Some common time signatures are shown in Table 14-5.

Table 14-5. Some time signatures.

Time Signature	Beats per measure	1 beat =
$\frac{4}{4}$ (or C)	4	quarter note
$\frac{2}{2}$ (or C)	2	half note
$\frac{2}{4}$	2	quarter note
$\frac{3}{4}$	3	quarter note
$\frac{6}{8}$	6	eighth note

14.5 An Example

Figure 14.5 shows a simple example of musical notation—the beginning of a popular song, *Old MacDonald Had a Farm*.

Let's break this down and see how it all works. Starting at the far left, you see the treble clef, which indicates which lines on the staff correspond to which notes. Immediately after the treble clef is the key signature, which is two $\sharp$ (sharp) signs on the line for note F_5 and the space for C_5 . As shown in Table 14-3, the key with two $\sharp$ signs is the key of D major, and in that key the sharp notes are $F\sharp$ and $C\sharp$. In the key of D major, all written F notes are to be played as $F\sharp$, and all written C notes are to be played as $C\sharp$.

¹One famous *avant-garde* musical composition is *4'33"* by the American composer John Cage. It consists entirely of rests, and contains no musical notes—it is just 4 minutes and 33 seconds of silence.

Old MacDonald Had a Farm



Figure 14.5: The first 12 notes of *Old MacDonald Had a Farm*, in D major.

After the key signature comes the time signature, which is $\frac{4}{4}$ in this case—meaning four quarter notes (or the equivalent) per measure. After the time signature we see the notes: D₅, D₅, D₅, A₄; B₄, B₄, A₄; F₅♯, E₅, E₅; and D₅. The words (*lyrics*) are written below the staff.

14.6 Musical Instruments

Musical instruments produce musical notes by creating standing waves of some sort. In *string instruments* (violin, cello, guitar, harp, etc.), a string under tension is caused to vibrate, either by being plucked or having a bow drawn across it. The string is held fixed at both ends, and standing waves are created in the string, which produces a sound. Recall that the frequencies f_n of standing waves fixed at both ends are given by

$$f_n = n \frac{v}{2L} \quad (n = 1, 2, 3, 4, \dots) \quad (14.2)$$

where v is the wave speed and L is the distance between the ends. Only the first harmonic ($n = 1$) standing wave is played on a string instrument. Recall also that the speed of waves in a string is given by $v = \sqrt{F_T/(m/L)}$ (where F_T is the tension and m/L is the string density), so the frequency of the first harmonic will be

$$f_1 = \frac{1}{2L} \sqrt{\frac{F_T}{m/L}}. \quad (14.3)$$

The performer can shorten the effective length L of the string, typically by pressing the string against the neck of the instrument. Since m/L is constant, we have $f_1 \propto 1/L$, and shortening the string will increase the pitch f_1 and play a higher note. String instruments will have several strings with different thicknesses; the thicker strings have a higher mass density m/L , so they play a lower pitch. In order to tune the instrument before playing, a set of knobs allows the player to change the tension F_T in each string to make sure it plays each note at the proper frequency; a higher tension gives a higher pitch.

In *brass instruments* (e.g. trumpet, trombone, French horn, tuba), the performer sets up standing sound waves in the instrument by blowing into a mouthpiece. The player's lips vibrate or “buzz” at a frequency that produces standing waves; different notes are produced by changing the length of tubing (using valves, or a slide for the trombone), and by changing the tension in the player's lips. In some brass instruments, like the trombone, the player can play the first harmonic by buzzing the lips very loosely in the mouthpiece; higher harmonics are produced by increasing the lip tension. In other instruments, like the French horn, the first harmonic cannot be played—only higher harmonics. This makes the French horn a tricky instrument to play—only slight changes in lip tension will change the note from one harmonic from the next.

In *woodwind instruments* (e.g. clarinet, oboe, bassoon, flute, recorder), as in brass instruments, the player sets up standing sound waves in the instrument. In this case, the vibrations are often produced with a reed, and the performer changes notes by opening or closing combinations of holes along the side of the instrument, using the fingers or a complex system of keys. Woodwind instruments generally play the first few harmonic standing waves; which ones can be played depend on the shape of the bore of the instrument.

Percussion instruments are instruments like drums, which produce a sound when a membrane or other surface is struck and allowed to vibrate, creating standing waves in the membrane. The timpani is a drum in which the tension in the membrane can be changed to produce a few different notes.

Some musical instruments are *transposing* instruments; for these instruments, the written notes are not the same as the notes that are actually played. For example, music for the French horn is written seven half steps higher than it is actually played. So when a French horn player plays a written middle C (C_4), the note that actually comes out of the instrument will be seven half steps lower, F_3 ; such a horn is said to be “pitched in F”, and is called an *F horn*. There is a lighter French horn favored by some players that is better for playing high notes; it plays a $B\flat$ for a written C, and is called a *$B\flat$ horn*. The horn most commonly seen in orchestras, with its very complex-looking system of tubing, is a *double horn*. The double horn contains tubing for *both* an F horn and a $B\flat$ horn, and allows the player to switch between the two sides using a thumb valve. The player will play lower notes on the F side of the horn, then use the thumb valve to switch to the $B\flat$ side for high notes, since they’re easier to play on that side. Today there’s an even more complex *triple horn*, which includes a third *descant horn* side for playing very high notes.

Transposition is partly for historical reasons, and partly to allow performers to play similar instruments more easily. For example, a trumpet player can play a French horn or tuba without having to learn a different fingering for each instrument. However, if a performer wishes to play music written for an instrument other than the one he is playing (a horn player playing music written for trombone, for example), he may need to mentally transpose the music while playing in order to play in the same key as the rest of the orchestra.

As mentioned earlier, music is a very large subject, and here we’ve only barely touched on the very basics of music theory and musical notation. There’s much more to this subject: chords, harmony, timbre, intervals, non-Western music, etc.—and there’s much more to musical notation than the bare outlines we’ve seen here. The interested reader is referred to books on music theory for more information.

Part IV

Electricity and Magnetism

Chapter 15

Electricity

The phenomenon of *electricity* has been known since ancient times. Long ago people discovered that rubbing fossilized tree resin (called *amber*) with fur could cause it to attract bits of light material. (In fact, the Greek word for amber, *ηλεκτρον*, is where we get our word “electricity”.)

Experiments by French scientist Charles du Fay in the early 18th century showed that there were two types of electricity: one he called “vitreous”, acquired by glass when rubbed with silk; and the other he called “resinous”, acquired by amber when rubbed with fur. He also discovered that two objects with vitreous charge repelled each other, as did two resinous-charged objects, but that a vitreous-charged object and resinous-charged object attracted each other.

Another of many early scientists studying electricity was the American scientist and statesman Benjamin Franklin. Franklin held the view that electricity was a fluid, and that the two types of electricity were actually an excess of electric fluid in one material and a deficiency of fluid in the other. But which was which? Franklin took a 50-50 shot in the dark—and missed! He called the vitreous charge “positive”, and the resinous charge “negative”, believing these to be an excess and deficiency of electric fluid (respectively). We now know it’s the other way around. What Franklin thought of as an electric fluid is actually a flow of particles called *electrons*, and it is an excess of electrons that is what we call “negative” charge; positive charge is a deficiency of electrons. Franklin’s unfortunate choice continues to be a source of some confusion in discussing electric current, as we’ll see later.

Benjamin Franklin is also famous for his (quite dangerous) “kite experiment”, in which he flew a kite into an electrically charged storm cloud. Electricity from the cloud conducted down the wet kite string to a key at the other end, and Franklin was able to produce sparks by bringing his knuckle near the key. The experiment showed that lightning is a form of electricity. (For his contributions to the theory of electricity, the unit of charge in electrostatic units is named the *franklin* in Benjamin Franklin’s honor.)

15.1 Electric Charge

Our modern understanding of electricity may be summarized as follows:

- There are two types of electricity, called *positive* (+) and *negative* (–).
- Like-charged bodies (+ and +, or – and –) repel; unlike-charged bodies (+ and –) attract.
- Electric charge is *quantized*; that is, the charge on a body must always be a multiple of the so-called *elementary charge* e . No charge can ever be smaller than e .
- Electric charge is *conserved*: that is, the total charge in a closed system is always constant.

15.2 Coulomb's Law

Using a torsion balance, the 18th century French physicist Charles-Augustin de Coulomb discovered the law that determines the amount of force between two charged bodies—a law now called *Coulomb's law*. It states that if two point charges q_1 and q_2 are separated by a distance r , then the force between them will be proportional to the product of the charges and inversely proportional to the square of the distance between them:

$$F = \frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2}. \quad (15.1)$$

Here F is the force (in newtons), r is the separation distance (meters), and q_1 and q_2 are the charges measured in units of *coulombs* (C). A coulomb is a very large unit of charge; charges we encounter in the laboratory will typically be on the order of microcoulombs (μC) or nanocoulombs (nC).

The constant ϵ_0 in Eq. (15.1) is called the *permittivity of free space*,¹ and is equal to²

$$\epsilon_0 = 8.85418781762038985 \dots \times 10^{-12} \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}. \quad (15.2)$$

The proportionality constant $1/(4\pi\epsilon_0)$ is called the *Coulomb constant* (k_c). It is equal to exactly

$$k_c = \frac{1}{4\pi\epsilon_0} = 8.9875517873681764 \times 10^9 \text{ N m}^2 \text{ C}^{-2}. \quad (15.3)$$

Coulomb's law (Eq. 15.1) implicitly makes use of a property in arithmetic that mirrors the properties of electric charges. Multiplying two numbers of like sign gives a positive number, and multiplying two numbers of unlike sign gives a negative number. This property mirrors the behavior of electric charges: two charges of like sign repel, and two charges of unlike sign attract. So in Coulomb's law (Eq. (15.1)), we can interpret a *positive* force as repulsion, and a *negative* force as attraction.

15.3 Atomic View of Electricity

As you will have already learned, all ordinary matter consists of *atoms*. At the center of the atom is a tiny, massive *nucleus*, which is surrounded by shells of very light *electrons*. The nucleus consists of electrically neutral (uncharged) *neutrons* along with positively-charged *protons* that carry a charge equal to the elementary charge, $e = 1.6021766208 \times 10^{-19} \text{ C}$. The electrons surrounding the nucleus carry a *negative* charge, also equal to the elementary charge. In other words, neutrons have charge 0, protons have charge $+e$, and electrons have charge $-e$.

In ordinary matter, it is only the *electrons* that move around and produce electric charge and electric currents. The protons are massive (about 1800 times heavier than the electrons) and tucked away in the center of the atom, so they barely move. When we rub a piece of amber with a piece of fur, for example, we're removing a small number of the outermost electrons from atoms in the fur, and depositing them onto the amber. This leaves the fur with a deficiency of electrons (giving it a positive charge) and the amber with extra electrons (giving it a negative charge). Very few electrons are involved in this type of charging: if only one fur atom in a *quintillion* loses an electron to the amber, it will produce an easily measurable electric charge, enough to allow the amber to pick up bits of paper, for example.

So keep this in mind: whenever you're charging objects or creating electric currents in the laboratory, it is always the negatively-charged *electrons* that are moving.³

¹ ϵ_0 is pronounced "epsilon-nought."

²Because of the way SI units are defined, the constant ϵ_0 is a transcendental number that may be computed to as many digits as desired; its exact value is $1/(299792458^2 \times 4\pi \times 10^{-7}) \text{ C}^2 \text{ N}^{-1} \text{ m}^{-2}$.

³Two hydrogen atoms walk into a bar. One says, "I've lost my electron." The other says, "Are you sure?" The first replies, "Yes, I'm positive..."

15.4 Materials

Different materials behave differently depending on their ability to allow electrons to flow through them. We classify materials as follows:

- *Conductors* are materials in which electrons can flow very easily. You can think of a conductor as a lattice of positive ions, surrounded by a kind of “gas” of free electrons that belong to no particular atom. The free electrons are free to move throughout the conductor. Familiar conductors are metals such as copper, gold, and silver.
- *Insulators* (or *dielectrics*) are materials in which each atom holds on to all of its atoms strongly, so they are *not* free to move through the material. Examples of insulators are rubber, wood, plastics, and ceramics.
- *Semiconductors* are between conductors and insulators. They are insulators that can be coaxed into giving up a conduction electron under the right conditions, such as a sufficiently strong electric field. Common semiconductors are the elements silicon and germanium.
- *Superconductors* are exotic materials that form a special class of conductor. While ordinary conductors always offer some sort of resistance to the flow of electrons, superconductors offer no such resistance. This means, for example, that if you form a superconductor into a ring and start electrons flowing in it, they will continue flowing forever. Traditional superconductors are made by cooling an ordinary conductor like mercury down to very low temperatures; below some critical temperature, the material will suddenly transition from an ordinary conductor to a superconductor. Experiments in the 1980s discovered a new class of superconductors called *high-temperature superconductors* that are made of exotic ceramic-like materials. These still need to be cooled to become superconducting, but not nearly as much. For example, mercury doesn't become superconducting until it's cooled down to 4.1 K, which requires liquid helium temperatures and is difficult to do. But the high-temperature superconductor $\text{YBa}_2\text{Cu}_3\text{O}_7$ only needs to be cooled to 90 K, which can easily be achieved by cooling with liquid nitrogen.

An exotic form of hydrogen called *metallic hydrogen* is thought to exist at the very high pressures (more than 4 million atmospheres) in the interior of the planets Jupiter and Saturn. Scientists are currently attempting to create metallic hydrogen in the laboratory, so far without success. Metallic hydrogen is thought to be either a solid or a superfluid⁴, and theory suggests it may possibly be a room-temperature superconductor. Its creation in the laboratory could have significant commercial applications.

15.5 Coulomb's Law in Two or Three Dimensions

Coulomb's law in the form shown in Eq. (15.1) works fine for a one-dimensional problem involving two point charges: the sign of the force F is sufficient to indicate the direction of the force. But when we work in two or three dimensions (for example, point charges on the vertices of a triangle) we must use *vectors* to determine the force in each charge. In vector form, Coulomb's law is

$$\mathbf{F}_{12} = -\frac{1}{4\pi\epsilon_0} \frac{q_1 q_2}{r^2} \hat{\mathbf{r}}_{12}, \quad (15.4)$$

where $\mathbf{F}_{12}$ is the force on charge q_1 due to charge q_2 , and $\hat{\mathbf{r}}_{12}$ is a *unit vector* (a vector of magnitude 1) that points in the direction from charge q_1 to charge q_2 . Note the minus sign: if both charges are positive, for example, then the force points *opposite* $\hat{\mathbf{r}}_{12}$ —that is, the force on q_1 will be away from q_2 .

⁴See chapter 59.

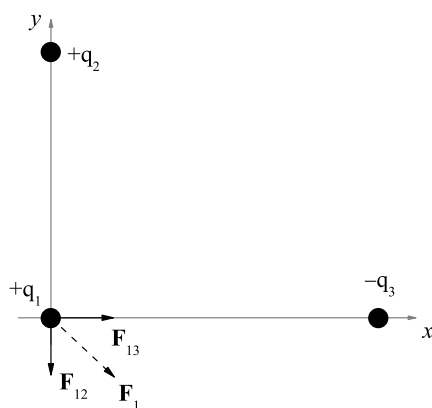


Figure 15.1: Example of Coulomb's law in two dimensions. Here charges q_1 , q_2 , and q_3 at the vertices of an isosceles right triangle; q_1 and q_2 are positive, and q_3 is negative. The total force $\mathbf{F}_1$ on charge q_1 is the vector sum of the force $\mathbf{F}_{12}$ on q_1 due to q_2 and the force $\mathbf{F}_{13}$ on q_1 due to q_3 : $\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13}$.

If you know the angle θ of the unit vector $\hat{\mathbf{r}}_{12}$ (measured counterclockwise from the $+x$ direction), then the unit vector in rectangular (cartesian) form is

$$\hat{\mathbf{r}}_{12} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}, \quad (15.5)$$

where $\mathbf{i}$ and $\mathbf{j}$ are unit vectors in the x and y directions, respectively.

If there are multiple charges present, then the *total* force on charge q_1 is the vector sum of all the forces on charge q_1 . For example, consider Fig. 15.1, which shows charges q_1 , q_2 , and q_3 at the vertices of an isosceles right triangle. The total force $\mathbf{F}_1$ on charge q_1 is the vector sum of the force $\mathbf{F}_{12}$ on q_1 due to q_2 and the force $\mathbf{F}_{13}$ on q_1 due to q_3 : $\mathbf{F}_1 = \mathbf{F}_{12} + \mathbf{F}_{13}$.

Appendix O gives a brief review of vector arithmetic.

Chapter 16

The Electric Field

Except for fairly simple problems involving a few charges, it's usually not particularly convenient to use Coulomb's law (Eq. (15.1)) directly. One would have to compute all the pairs of forces between each of the charges making up each of the bodies, which could become a fairly complex calculation. Instead, we introduce the idea of an *electric field* as a kind of intermediate quantity. We think of one body as producing an electric field at each point in space; we can then look at how a second body responds to that electric field. One reason this is convenient is that we often know what the electric field looks like without necessarily knowing anything about the distribution of charges that produced the field.

Now let's define the electric field. The electric field is a *vector field*—it assigns a vector to every point in space. So let's imagine you're standing in a room and wish to find the electric field vector at some point within the room. You take a small *positive* point charge q_0 (say a proton) and place it at that point, and measure the electric force on it. Then the electric field $\mathbf{E}$ is the force $\mathbf{F}$ divided by the test charge q_0 :

$$\mathbf{E} = \frac{\mathbf{F}}{q_0}. \quad (16.1)$$

The electric field vector has units of newtons per coulomb (N/C).

A typical situation is that we will already know the electric field by some other calculation; then Eq. (16.1) indicates that the force on a charge q in the electric field $\mathbf{E}$ is $\mathbf{F} = q\mathbf{E}$.

16.1 Electric Field due to a Point Charge

The electric field due to a point charge q can be found by using Coulomb's law. Let's put a small positive test charge q_0 at some distance r from the charge q ; then by Coulomb's law, the force on q_0 is $F = (1/4\pi\epsilon_0)(qq_0/r^2)$. Dividing by q_0 gives us the electric field due to charge q :

$$E = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}. \quad (16.2)$$

16.2 Electric Field Lines

To help visualize the shape of the electric field, it can be helpful to draw diagrams of *electric field lines*. These lines have the following properties:

- The electric field lines are directed lines (with arrows) that point *from* positive (+) charge *to* negative (−) charge.

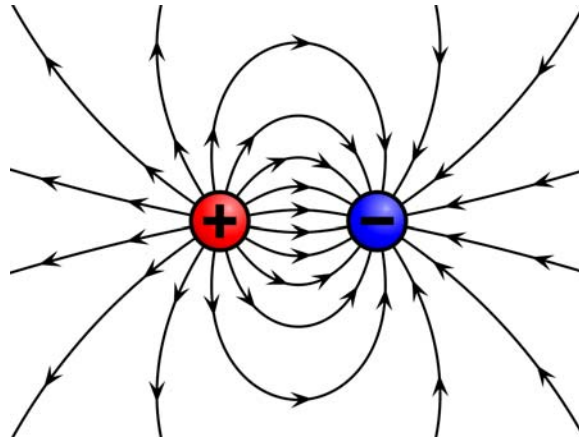


Figure 16.1: Electric field lines for an electric dipole. The dipole moment vector $\mathbf{p}$ points to the left in this case. (©GNU-FDL, Wikimedia Commons [11].)

- At any point along a field line, the electric field vector $\mathbf{E}$ is tangent to the field line.
- We cannot possibly draw *all* field lines (because they fill all space), so we draw only a few. The number of field lines you draw is somewhat arbitrary — we just draw enough to visualize the field without making the diagram too crowded.
- The number of field lines terminating on a charge should be proportional to the charge.
- The closer together the field lines are, the stronger the electric field.

16.3 The Electric Dipole

As an example, consider Fig. 16.1, which shows two charges of equal magnitude and opposite sign, separated by a fixed distance; such an arrangement is called an *electric dipole*.

An electric dipole may be characterized by a quantity called the *dipole moment*. The dipole moment $\mathbf{p}$ of an electric dipole is defined as

$$\mathbf{p} = q\mathbf{d}, \quad (16.3)$$

where q is the magnitude of either of the charges in the dipole, and $\mathbf{d}$ is a vector whose length is equal to the distance between the charges, and which points from the negative charge to the positive charge (opposite the direction of the electric field line between the charges). The dipole moment essentially measures how electrically “polarized” a pair of charges is, with larger values when more charge is separated by a greater distance. Electric dipole moment is measured in units of coulomb-meters (C m).

16.4 Electric Flux

Electric flux may be thought of as being proportional to the total number of electric field lines passing through a given area. Given an area A embedded in an electric field $\mathbf{E}$, the electric flux Φ_E passing through plane S of area A is equal to the product of $\mathbf{E}$ and the component of A perpendicular to the field:

$$\Phi_E = \mathbf{E} \cdot \hat{\mathbf{n}}A = EA \cos \theta. \quad (16.4)$$

Here $\hat{\mathbf{n}}$ is a unit vector perpendicular to surface S and A is the total area of S . If S is a curved surface instead of a plane, then the electric flux is more generally

$$\Phi_E = \int_S \mathbf{E} \cdot \hat{\mathbf{n}} dA \quad (16.5)$$

where dA is an infinitesimally small piece of area of S , and the integral is over the entire area of S . In other words, we imagine dividing surface S into many tiny squares, each of which has area dA . For each square, we draw a normal unit vector $\hat{\mathbf{n}}$ at that square, and we compute $\mathbf{E} \cdot \hat{\mathbf{n}}$, which is the component of the electric field $\mathbf{E}$ that is perpendicular to that square. We then multiply that result by the area of the square dA to get the electric flux through area dA . We add each of those fluxes together over the entire area A of surface S .

16.5 Gauss's Law

One important application of electric flux is its appearance in *Gauss's law*, which is one of the four fundamental equations of classical electromagnetism known as *Maxwell's equations*. Gauss's law states that if we draw an imaginary *closed* surface in space, then the total electric flux through that closed surface S is proportional to the total amount of charge enclosed inside that surface:

$$\Phi_E = \oint_S \mathbf{E} \cdot \hat{\mathbf{n}} dA = \frac{q_{\text{encl}}}{\epsilon_0}. \quad (16.6)$$

Here the circle on the integration sign indicates that the integral is over the *closed* surface S .

While Gauss's law is generally true, one of its important practical uses is that it allows the quick determination of the electric field of a symmetrical distribution of charges. For example, it allows the electric field of a spherical or cylindrical charge to be determined very easily.

As a simple example, suppose we wish to find the electric field E at a distance r from a point charge q . We would imagine drawing an imaginary spherical surface of radius r centered on q , so that the sphere passes through the point at which we wish to calculate E . Then on the left-hand side of Eq. (16.6), the electric flux is the electric field E times the area of the sphere: $\Phi_E = \oint E dA = E \oint dA = E(4\pi r^2)$. Since the total charge inside the sphere is q , the right-hand side becomes q/ϵ_0 . Gauss's law then gives $4\pi r^2 E = q/\epsilon_0$, or $E = q/(4\pi r^2 \epsilon_0)$, in agreement with Coulomb's law.

Both Coulomb's law and Gauss's law allow us to determine the electric field due to an arbitrary distribution of charge. The difference between them is that, for *symmetrical* charge distributions, Gauss's law provides a shortcut that allows us to compute the electric field much more easily than using Coulomb's law. We can always use Coulomb's law—Gauss's law is just much less work when we have a symmetrical charge distribution. For irregular charge distributions, though, we may have no choice but to “do it the hard way” and resort to Coulomb's law.

16.6 Electric Fields of Conductors

If an electrical conductor holds a net charge, then it has a number of important properties. If the conductor is in electrostatic equilibrium (i.e. all charges have stopped moving), then:

- *The electric field inside the conductor is zero.* The conductor has free electrons throughout its interior. If there were an electric field inside the conductor, then there would be a force on those free electrons, causing them to accelerate, in violation of the assumption that the conductor is in equilibrium. Therefore $\mathbf{E} = \mathbf{0}$ inside a conductor.

- *Any excess charge in the conductor must lie on its surface.* Using Gauss's law, draw a Gaussian surface just below the surface of the conductor. Since the electric field inside the conductor is zero, the electric flux through this surface is zero. Then by Gauss's law, the charge inside the surface is zero. Therefore, any excess charge must lie on the surface. (Another way to think of this is that since the charges repel, they will want to get as far away from each other as possible, so they will end up on the surface.)
- *Electric field lines are perpendicular to the surface of the conductor.* If the electric field lines intersected the surface of the conductor at some angle, then there would be a tangential component of the electric field present, which would cause the electrons to accelerate parallel to the surface. Therefore electric field lines must meet the surface of the conductor at right angles.

16.7 Dielectric Breakdown

It is possible for materials that are normally insulators (dielectrics) to become electrically conducting, if they are in the presence of a sufficiently large electric field. For example, air is normally an insulator, but the presence of an electric field of at least 3×10^6 N/C creates channels of ionized gas through which electrons can flow; the result is the familiar *spark*. This phenomenon is called *dielectric breakdown*.

16.8 Lightning

Another example of dielectric breakdown is *lightning*. During a thunderstorm, falling water drops and snow pellets cause the clouds to acquire a negative charge, while the ground becomes positively charged; this creates an electric field pointing upward. Electrons from the thundercloud carve a channel of ionized gas that makes its way to the ground in a series of steps; this channel is called the *stepped leader*. At the same time, a number of shorter ionized leader channels reach from the ground to a short distance upward. At some point the downward-moving stepped leader connects with one of the upward leaders, and forms a complete conducting path of ionized gas from the cloud to the ground. This causes a powerful ionizing wavefront, called the *return stroke*, to move very quickly from the ground back up to the cloud, producing the flash we see. The return stroke heats the surrounding air to a very high temperature, causing it to expand at supersonic speed. This creates a shock wave that produces the sound we hear as thunder. Typically several such strokes carry current between the ground and the cloud during a single flash of lightning.

Chapter 17

Electric Potential

17.1 Potential Energy

There is a potential energy associated with the electric force. Suppose, for example, that you have a positive and negative charge right next to each other. Now separate the two charges by some distance; since the two charges are attracted, they will “want” to come back together. You had to do work against the electric force to separate the two charges, and now the system has a potential energy that will be released if you allow the charges to come back together. The force F and potential energy U are related by

$$F = -\frac{dU}{dx} \approx -\frac{\Delta U}{\Delta x}. \quad (17.1)$$

The same thing happens with the gravitational force. If two masses are separated, the attractive gravitational force will cause the two masses to want to come together again, so the system of separated masses contains potential energy. This potential energy can be released by allowing the masses to come back together. In the case of gravity, the potential energy of two point masses m_1 and m_2 separated by distance r is $U = -Gm_1m_2/r$ (where we choose $U = 0$ at $r = \infty$, so U is always negative). Similarly, with the electric force, the potential energy of two point charges q_1 and q_2 separated by distance r is

$$U = \frac{1}{4\pi\epsilon_0} \frac{q_1q_2}{r}, \quad (17.2)$$

where again $U = 0$ at $r = \infty$, and U is always negative for attracting charges and always positive for repelling charges.

Another common situation is the potential energy in a uniform field. For gravity, the potential energy of a mass m in a uniform gravitational field g is $U = mgh$, where h is the height above some arbitrarily-chosen level for which U is taken to be zero. Similarly, the potential energy of a charge q in a uniform electric field E is

$$U = qEd, \quad (17.3)$$

where d is the distance from some level at which U is chosen to be zero.

17.2 Potential

Recall how the electric field $\mathbf{E}$ was defined: by dividing the force on a small positive test charge by the magnitude of the test charge, we get the electric field, which is a property of space. We can do something

similar with potential energy, and find a similar quantity that is a property of space only. This quantity is called the *potential*.

Let's first look at how this would be done with gravity. As we've seen, the gravitational potential energy of two point masses is $U = -Gm_1m_2/r$. By dividing by one of the masses, we can get the *gravitational potential* $\mathcal{G}$ due to mass m at distance r from the mass: $\mathcal{G} = -Gm/r$. The gravitational potential $\mathcal{G}$ has units of J/kg.

We can do something similar with the electric force. The electric potential energy between two point charges is $U = q_1q_2/(4\pi\epsilon_0r)$; by dividing this by one of the charges, we get an expression for the *electric potential* V due to charge q at distance r from the charge:

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{r}. \quad (17.4)$$

The electric potential is measured in units of *volts* (V), named for the Italian physicist Alessandro Volta. One volt is equal to one joule per coulomb ($1 \text{ V} = 1 \text{ J/C}$). Electric potential is sometimes called *voltage*.

As with potential energy, it is really only *differences* in potential that are physically meaningful. Equivalently, we are free to choose what point in space (or a circuit) is chosen to have a potential of zero volts, and all other potentials are measured with respect to that. In an electric circuit, there is usually a point called *ground* that is connected to the Earth and/or to the negative terminal of a power source, and the ground is taken to be 0 V by convention.

Another common situation is a uniform field. In a uniform gravitational field g , the potential energy is $U = mgh$; dividing by the mass m we find the gravitational potential is $\mathcal{G} = gh$. Similarly, in a uniform electric field E , the potential energy is $U = qEd$; dividing by the charge q we find the electric potential is

$$V = Ed. \quad (17.5)$$

Solving this for E , we can see that the electric field can be expressed in units of V/m as well as N/C. You can check that these are equivalent by breaking everything down into base units (kg, m, s, A) with the help of Table 2-2.

Because of the similarity between electric potential and gravitational potential, it can sometimes be helpful to think of potential as being analogous to height. Positive charge will tend to "fall" from high potential to low potential.

Just as force and potential energy are related by Eq. (17.1), field strength and potential are similarly related. The electric field E is related to the electric potential V by

$$E = -\frac{\Delta V}{\Delta x}. \quad (17.6)$$

(The corresponding relation for gravity is $g = -\Delta\mathcal{G}/\Delta x$.)

17.3 Equipotential Surfaces

Imagine drawing a surface in space such that every point on the surface is at the same potential. Such a surface is called an *equipotential surface*. An important property of equipotential surfaces is that they always intersect electric field lines at right angles. (If this were not so, then there would be a component of $\mathbf{E}$ in the plane of the equipotential surface and thus a component of the net force in the plane of the surface, in violation of the assumption that the surface is one of constant potential.)

17.4 Comparison between Gravity and Electricity

The following table summarizes the formulae for field strength, force, potential, and potential energy, both for a uniform (constant) field and for a field due to a point particle.

Table 17-1. Comparison of quantities in gravity and electricity.

Quantity	Gravity	Electricity
Uniform field		
Field strength	$g = \text{const.}$	$E = \text{const.}$
Force	$F = mg$	$F = qE$
Potential	$\mathcal{G} = gd$	$V = Ed$
Potential energy	$U = m\mathcal{G} = mgd$	$U = qV = qEd$
Point particles		
Field strength	$g = -Gm/r^2$	$E = q/(4\pi\epsilon_0 r^2)$
Force	$F = -Gm_1m_2/r^2$	$F = q_1q_2/(4\pi\epsilon_0 r^2)$
Potential	$\mathcal{G} = -Gm/r$	$V = q/(4\pi\epsilon_0 r)$
Potential energy	$U = -Gm_1m_2/r$	$U = q_1q_2/(4\pi\epsilon_0 r)$

17.5 The Electron Volt

If a particle with an electric charge e (such as an electron or proton) is accelerated through a potential difference of 1 volt, it gains a kinetic energy of 1 *electron volt* (eV). Note that the electron volt is a unit of *energy*, not voltage. One electron volt is equal to $1.6021766208 \times 10^{-19}$ joules.

Notice that it doesn't matter how far the charged particle travels, or how much time it takes to accelerate: it only matters that the particle is accelerated through a potential difference of 1 volt. More generally, if a charge Ne is accelerated through a potential difference V , the particle will gain an energy of NeV electron volts.

The electron volt is a common unit of energy in atomic and particle physics. Common multiples are the kilo-electron volt (keV), mega-electron volt (MeV), and giga-electron volt (GeV).

Chapter 18

The Battery

There are many ways of creating an electrical potential; one of the simplest is the *battery*, in which an electrical potential is created by a chemical reaction. A battery consists of two strips of dissimilar metal (called *electrodes*) placed in solution called an *electrolyte*. The electrolyte will preferentially dissolve one of the electrodes, leaving an electric charge on one of the electrodes and the opposite charge on the other. As an example, consider the common zinc-carbon battery. Two electrodes—one of zinc and one of carbon—are placed in an electrolyte of sulfuric acid. The acid dissolves a little of the zinc electrode, placing Zn^{2+} ions in solution and leaving extra electrons behind on the zinc electrode, so that it becomes negatively charged. If the battery is not connected to anything, then the system reaches an equilibrium condition: as the zinc electrode becomes negatively charged, it will tend to attract the Zn^{2+} ions back to it and restore the zinc again. If the battery is connected to something, the zinc ions will continue to be produced, and will start to pull electrons from the carbon electrode, which will become positively charged. As the battery continues to be used, the electrodes will become more and more dissolved, until one of the electrodes is used up and the battery dies.

The amount of potential difference between the two electrodes (the *terminals* of the battery) depends on the chemistry, and in particular on the two metals present. In the case of a zinc-carbon battery, the potential between the terminals is 1.5 V. Other types of batteries will have other potential differences, as shown in Table 18-1.

Table 18-1. Common battery types.

Battery Type	+ Terminal	– Terminal	Potential
Zn-C	C	Zn	1.5 V
Alkaline	MnO_2	Zn	1.5 V
Silver oxide	Ag_2O	Zn	1.55 V
Lead acid	PbO_2	Pb	2.1 V
Ni-Cd	NiOOH	Cd	1.2 V
Ni-Zn	NiOOH	Zn	1.65 V
NiMH	NiOOH	metal alloy	1.2 V
Lithium ion	Li compound	Li compound	3.6 V

When batteries are connected in *series* (end to end), their voltages add. This is what you're doing when you put several batteries into a device like a calculator or flashlight: the + terminal of one battery is connected to the – terminal of the next. For example, you may put four size AAA alkaline batteries into a calculator, which provides a total potential of $4 \times 1.5 \text{ V} = 6 \text{ V}$. A 9 V battery actually consists of six individual batteries connected in series in a single casing. A car battery consists of six lead-acid batteries connected in series, for

a total potential of 12 V. Sometimes a single electrochemical system is called a *cell*, with the word “battery” being reserved for several cells connected in series.

It's also possible to connect several batteries in *parallel*, so that all + terminals are connected together and all – terminals are connected together. This arrangement will have the same potential as a single battery, but will be able to deliver more electric current. This is not usually done, since one could simply replace the multiple batteries with a single larger-sized battery instead.

In a real battery, the potential delivered by the battery is not constant, but varies with the amount of current delivered by the battery. This is due to the battery's *internal resistance*, and is discussed further in Section 20.5.

A good way to think of a battery is as a kind of electron “pump”: it pumps current around an electrical circuit, much like a water pump would pump water.

Chapter 19

Electric Current

If we place a potential difference V across opposite ends of a conductor (a copper wire, for example), then there will be an electric field $E = -\Delta V/\Delta x$ created inside the conductor. The free electrons will respond to this electric field, moving opposite the direction of the electric field. This motion of electrons is called an *electric current*, and is analogous to the flow of water in a stream.

Current is measured as the amount of current passing a fixed point in the conductor per unit time. A current of 1 coulomb of charge per second is defined to be 1 ampere (A), after the French physicist André-Marie Ampère: $1 \text{ A} = 1 \text{ C/s}$.

By convention, the direction of electric current is taken to be *opposite* the direction of the flow of electrons. Another way to think of this is to imagine electric current to be due to the flow of positive charges through the conductor (even though it's actually the negative electrons that are moving). (This somewhat confusing situation is related to Benjamin Franklin's unfortunate choice of which type of charge to call "positive" and which to call "negative".) Conventional current moves in the direction from high potential to low potential. If a conductor is connected to the terminals of a battery, then conventional current flows from the $+$ terminal, through the conductor, back to the $-$ terminal.

Electric current does not flow smoothly through a conductor. Electrons inside the conductor are moving around at random, bumping into other electrons in their vicinity. Superimposed on this random motion is a gradual drift of the electrons opposite the direction of the electric field. This speed of the electrons through the conductor is called the *drift velocity*. If the density of free electrons (electrons per unit volume) is n , then the total charge per unit volume is ne (where e is the elementary charge). In time t , the volume of electrons that move through the wire is $Av_d t$, where A is the cross-sectional area and v_d is the drift velocity. This means the total charge moving through the wire in time t is $(ne)(Av_d t)$, and so the current is found by dividing this by t :

$$I = neAv_d. \tag{19.1}$$

Here's an important point to keep in mind: one speaks of the potential difference (or voltage) *between two points* in an electrical circuit; but one speaks of the electric current *at one point* in the circuit. For example, you refer to the voltage *across* a resistor, but the current *through* a resistor.

Chapter 20

Resistance

Suppose we apply a potential difference V across the ends of a conductor. If the conductor were to allow the free, unimpeded flow of electrons, then the resulting current in the conductor would be unlimited. But in a real conductor, there is always some electrical *resistance* to the flow of electric current due to the free electrons constantly bumping into their neighbors. This electrical resistance is measured in units of *ohms* (Ω), after German physicist Georg Simon Ohm. One ohm is defined to be that resistance that produces a current of 1 ampere in the presence of a potential difference of 1 volt: $1 \Omega = 1 \text{ V/A}$.

Resistance is often introduced deliberately into electrical devices by electronic components called *resistors*. A resistor is typically a small cylindrical device with metal wires protruding from each end. The cylinder is decorated with color bands, which are a *color code* (Figure 20.1) that indicates the value of the resistance. In a four-band color code, the first two bands are the first two significant digits of the resistance, and the third band is the power of 10 by which the first two bands are to be multiplied. A fourth band indicates the *tolerance*—how far the resistor is allowed to be from its marked value.

20.1 Resistivity

Even a plain conductor—like a copper wire—contain some small amount of resistance. The resistance of a conductor is related to its dimensions and to a quantity called its *resistivity*. If the resistance is R , and the resistivity is ρ , then the two are related by

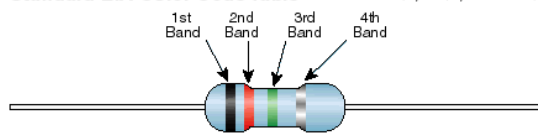
$$R = \rho \frac{L}{A}, \quad (20.1)$$

where R is the resistance (Ω), ρ is the resistivity ($\Omega \text{ m}$), L is the length of the conductor (in the direction of the flow of current), and A is the cross-sectional area of the conductor (perpendicular to the direction of the flow of current). It's important to recognize that the resistivity ρ is an intrinsic property of the material: for example, you can look up the resistivity of copper in a physics handbook. The resistance R , though, depends on the geometry—the length and diameter of the conductor, as well as its resistivity.

It turns out that the resistivity depends on temperature. You can compute the temperature correction using the equation

$$\rho = \rho_0 [1 + \alpha(T - T_0)]. \quad (20.2)$$

Here ρ_0 is the resistivity at temperature T_0 , ρ is the resistivity at temperature T , and α is called the *temperature coefficient of resistivity*. You can find ρ_0 , T_0 and α for a particular conductor in a physics handbook (e.g. Table 20-1); then for any given temperature T , you can find the resistivity ρ at that temperature.

Standard EIA Color Code Table 4 Band: $\pm 2\%$, $\pm 5\%$, and $\pm 10\%$


Color	1st Band (1st figure)	2nd Band (2nd figure)	3rd Band (multiplier)	4th Band (tolerance)
Black	0	0	10^0	
Brown	1	1	10^1	
Red	2	2	10^2	$\pm 2\%$
Orange	3	3	10^3	
Yellow	4	4	10^4	
Green	5	5	10^5	
Blue	6	6	10^6	
Violet	7	7	10^7	
Gray	8	8	10^8	
White	9	9	10^9	
Gold			10^{-1}	$\pm 5\%$
Silver			10^{-2}	$\pm 10\%$

Chart Provided By: 

Figure 20.1: The resistor color code (4 bands).

Table 20-1. Resistivities and Temperature Coefficients (at $T_0 = 20^\circ\text{C}$). [10]

Material	Resistivity ρ ($\Omega \cdot \text{m}$)	Temperature coeff. α ($^\circ\text{C}$) $^{-1}$
<i>Conductors</i>		
Silver	1.59×10^{-8}	0.0061
Copper	1.68×10^{-8}	0.0068
Gold	2.44×10^{-8}	0.0034
Aluminum	2.65×10^{-8}	0.00429
Tungsten	5.6×10^{-8}	0.0045
Iron	9.71×10^{-8}	0.00651
Platinum	10.6×10^{-8}	0.003927
Mercury	98×10^{-8}	0.0009
Nichrome (Ni, Fe, Cr alloy)	100×10^{-8}	0.0004
<i>Semiconductors</i> [†]		
Carbon (graphite)	$(3-60) \times 10^{-5}$	-0.0005
Germanium	$(1-500) \times 10^{-3}$	-0.05
Silicon	0.1-60	-0.07
<i>Insulators</i>		
Glass	$10^9 - 10^{12}$	
Hard rubber	$10^{13} - 10^{15}$	

[†] Values depend strongly on the presence of even slight amounts of impurities.

Multiplying both sides of Eq. (20.2) by L/A , we see the resistance changes with temperature by a similar formula:

$$R = R_0 [1 + \alpha(T - T_0)]. \quad (20.3)$$

Here R_0 is the resistance at temperature T_0 , and R is the resistance at temperature T .

For *copper* wire, a convenient empirical equation is [7]

$$R = R_0 \frac{234.5 + T}{234.5 + T_0}, \quad (\text{copper only}) \quad (20.4)$$

where T and T_0 are in degrees Celsius.

Eq. (20.3) suggests that it would be possible to use a resistor as a thermometer: by accurately measuring the resistance of a resistor, one can infer the temperature. Solving Eq. (20.3) for the temperature T , we find

$$T = T_0 + \frac{1}{\alpha} \left(\frac{R}{R_0} - 1 \right). \quad (20.5)$$

In principle, this equation could be used to measure the temperature of a resistor by measuring its resistance.

A *thermistor* is a type of resistor specifically designed for this type of temperature measurement. However, Eq. (20.5) is not really adequate for accurate temperature measurement with a thermistor. Instead, one uses a more accurate model called the *Steinhart-Hart equation*:

$$T = \frac{1}{a + b \ln R + c (\ln R)^3}, \quad (20.6)$$

where a , b , and c are called the *Steinhart-Hart parameters*, and are provided by the thermistor manufacturer.

20.2 Resistors in Series and Parallel

Several resistors connected end-to-end (*in series*) have an equivalent resistance equal to the sum of the individual resistances:

$$R_s = \sum_i R_i \quad (20.7)$$

$$= R_1 + R_2 + R_3 + \cdots \quad (20.8)$$

If they are connected *in parallel*, the the equivalent resistance is the reciprocal of the sum of the reciprocals of the individual resistances:

$$\frac{1}{R_p} = \sum_i \frac{1}{R_i} \quad (20.9)$$

$$= \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots \quad (20.10)$$

A common error in computing parallel resistances is to compute sum of the reciprocals of the individual resistances, then forget to take the reciprocal of the result at the end. Be careful not to do this!

Note the following points. For resistors connected *in series*:

- The equivalent resistance will be bigger than the largest resistance in the series combination.
- If one resistor in the series combination is much larger than the others, the equivalent resistance will be approximately equal to the largest resistance.

- N equal resistors R connected in series have an equivalent resistance of NR .

For resistors connected *in parallel*:

- The equivalent resistance will be smaller than the smallest resistance in the parallel combination.
- If one resistor in the parallel combination is much smaller than the others, the equivalent resistance will be approximately equal to the smallest resistance.
- N equal resistors R connected in parallel have an equivalent resistance of R/N .
- For the special case of just *two* resistors in parallel, Eq. (20.9) becomes the product of the resistances divided by their sum:

$$R_p = \frac{R_1 R_2}{R_1 + R_2}. \quad (20.11)$$

It may sometimes be handy to use the notation

$$x||y \equiv \frac{xy}{x+y} \quad (20.12)$$

so that the equivalent resistance of two resistors in parallel is $R_1 || R_2 = R_1 R_2 / (R_1 + R_2)$.

20.3 Conductance

Related to the resistance R and resistivity ρ are the conductance G and conductivity σ :

$$G = \frac{1}{R}; \quad \sigma = \frac{1}{\rho}. \quad (20.13)$$

Conductance is measured in units of *siemens* (S), named for German inventor Ernst Werner von Siemens. The siemens is also sometimes called the *mho* ($\oslash$), which is “ohm” spelled backwards. Conductivity is measured in units of S/m.

The relation between conductance and conductivity is found by taking the reciprocal of Eq. (20.1):

$$G = \sigma \frac{A}{L}. \quad (20.14)$$

20.4 Wire

In computing the resistivity or conductivity of wire in Eqs. (20.1) and (20.14), you will need to know the cross-sectional area A of the wire. In the United States, wire is sold in standard diameters that are numbered according to the *American Wire Gauge* (AWG), as shown in Table 20-2. Wire used in laboratory work is typically 20-gauge or 22-gauge copper wire.

By definition, AWG 0000 wire has a diameter of 0.46 inches, and AWG 36 wire has a diameter of 0.005 inches. This implies that AWG n wire has a diameter d of

$$d = 0.005 \times 92^{(36-n)/39} \text{ inches}, \quad (20.15)$$

where $n = -3$ for AWG 0000, $n = -2$ for AWG 000, and $n = -1$ for AWG 00. This formula was used to create Table 20-2. (Note that a *larger* AWG number corresponds to a *smaller* diameter wire.)

20.5 Battery Internal Resistance

A real battery contains an *internal resistance* to the flow of electricity that causes it to have a lower voltage when delivering current to a circuit than when it isn't. The potential difference across a battery's terminals when it's *not* connected to a circuit and doing work is called its *electromotive force* (or "emf"), $\mathcal{E}$. (Note that despite the name, this is *not* a force, but a voltage, measured in volts.)

The actual potential difference across a battery's terminals when it *is* doing work on a circuit is called the *terminal voltage*. The terminal voltage V may be modeled as

$$V = \mathcal{E} - Ir, \quad (20.16)$$

where I is the current being delivered by the battery, and r is the internal resistance of the battery. The more current is drawn from the battery, the smaller its terminal voltage.

Table 20-2. American Wire Gauge (AWG).

Gauge	Diameter (in)	Area (m ²)
0000	0.46000	1.07219×10^{-4}
000	0.40964	8.50288×10^{-5}
00	0.36480	6.74309×10^{-5}
0	0.32486	5.34751×10^{-5}
1	0.28930	4.24077×10^{-5}
2	0.25763	3.36308×10^{-5}
3	0.22942	2.66705×10^{-5}
4	0.20431	2.11506×10^{-5}
5	0.18194	1.67732×10^{-5}
6	0.16202	1.33018×10^{-5}
7	0.14429	1.05488×10^{-5}
8	0.12849	8.36556×10^{-6}
9	0.11442	6.63419×10^{-6}
10	0.10190	5.26115×10^{-6}
11	0.09074	4.17229×10^{-6}
12	0.08081	3.30877×10^{-6}
13	0.07196	2.62398×10^{-6}
14	0.06408	2.08091×10^{-6}
15	0.05707	1.65023×10^{-6}
16	0.05082	1.30870×10^{-6}
17	0.04526	1.03784×10^{-6}
18	0.04030	8.23047×10^{-7}
19	0.03589	6.52706×10^{-7}
20	0.03196	5.17619×10^{-7}
21	0.02846	4.10491×10^{-7}
22	0.02535	3.25534×10^{-7}
23	0.02257	2.58160×10^{-7}
24	0.02010	2.04730×10^{-7}
25	0.01790	1.62359×10^{-7}
26	0.01594	1.28756×10^{-7}
27	0.01420	1.02108×10^{-7}
28	0.01264	8.09755×10^{-8}
29	0.01126	6.42165×10^{-8}
30	0.01003	5.09260×10^{-8}
31	0.00893	4.03862×10^{-8}
32	0.00795	3.20277×10^{-8}
33	0.00708	2.53991×10^{-8}
34	0.00630	2.01424×10^{-8}
35	0.00561	1.59737×10^{-8}
36	0.00500	1.26677×10^{-8}
37	0.00445	1.00459×10^{-8}
38	0.00397	7.96679×10^{-9}
39	0.00353	6.31795×10^{-9}
40	0.00314	5.01036×10^{-9}

Chapter 21

Ohm's Law

For many materials and devices (conductors and resistors, for example), it is found that the greater the potential difference placed across the device, the greater the resulting current. This is called *Ohm's law*, and may be stated as

$$V = IR, \quad (21.1)$$

where V is the potential difference (in volts), I is the current (in amperes), and R is the resistance (in ohms). Ohm's law, like Hooke's law, is an example of what is called an *empirical law*: something that is found to be at least approximately correct in many situations, but is not necessarily always true. This is an important point: Ohm's law is not always true! It is just something that is found to work for many things like conductors and resistors. Ohm's law does *not* apply in some cases: lamp filaments, diodes, and solar cells, for example. Such devices are said to be *non-ohmic*.

Ohm's law may be considered the most important principle in the analysis of electric circuits. You'll use it over and over again as we learn to analyze electric circuits.

21.1 Electric Power

The electric power P consumed by a resistor is given by

$$P = IV, \quad (21.2)$$

where P is in watts. What this specifically refers to is the rate at which electrical energy is converted to heat. Commercially made resistors come in several standard power ratings (e.g. $\frac{1}{8}$ W, $\frac{1}{4}$ W, $\frac{1}{2}$ W, 1 W). When building a circuit, you have to make sure that the product of the current through and voltage across a resistor does not exceed its power rating.

Using Ohm's law (Eq. (21.1)) we can write the electric power (Eq. (21.2)) in several equivalent forms:

$$P = IV = I^2 R = \frac{V^2}{R}. \quad (21.3)$$

You can use any of these to compute the power consumed by a resistor; which one you use depends on which quantities you know: I , V , or R .

Chapter 22

DC Electric Circuits

Electronic components like batteries and resistors may be combined into closed loops called *circuits*. An almost endless variety of such circuits may be used to create useful device: clocks, calculators, radios, etc. In designing an electronic circuit, an engineer or electronics hobbyist will need to perform some simple calculations to figure out how much current is going through each part of the circuit, and how much potential difference there is across each component. We'll look at some of the basic methods of analysis here, using simple circuits consisting only of batteries and resistors.

22.1 Schematic Diagrams

To show how the components of an electronic circuit are connected together, we draw a *schematic diagram*. Such a diagram uses symbols to represent the different components (as shown in Fig. 22.1), along with lines to represent the connecting wires. When two lines in a diagram cross, a dot is used to indicate that the wires are electrically connected at that point; the absence of a dot means that there is no electrical connection.

Note that in the battery symbol, the end with the *long* line is the *positive* $+$ terminal. *Ground* refers to a connection to a large conductor—traditionally to a copper pipe driven into the earth. You will often see several parts of a circuit connected to a common ground, with the $-$ terminal of the battery or power supply serving as the ground.

22.2 Kirchhoff Plots

A good way to visualize what's happening in an electrical circuit is a diagram that has been called a *Kirchhoff plot* (Ref. [17]). This is a three-dimensional plot in which one draws the circuit in the x - y plane; the potential (voltage) at any point in the circuit is then plotted on the z axis. (See Fig. 22.2.) The plot helps you to think of voltage as analogous to elevation: batteries cause an increase in elevation, and resistors cause a drop in elevation. And just as water always flows downhill, you can use the diagram to help visualize electric current flowing from high potential to low potential.

22.3 A Simple Circuit

As an example, let's look at the simple circuit shown in Fig. 22.3, consisting of a battery and three resistors.

Conventional current flows around the circuit clockwise: from the $+$ terminal of the battery, through the resistors, then back into the $-$ terminal of the battery. (Remember that the electrons actually travel in the opposite direction, counterclockwise.) The current flowing from the battery through resistor R_1 we label I_1 ;

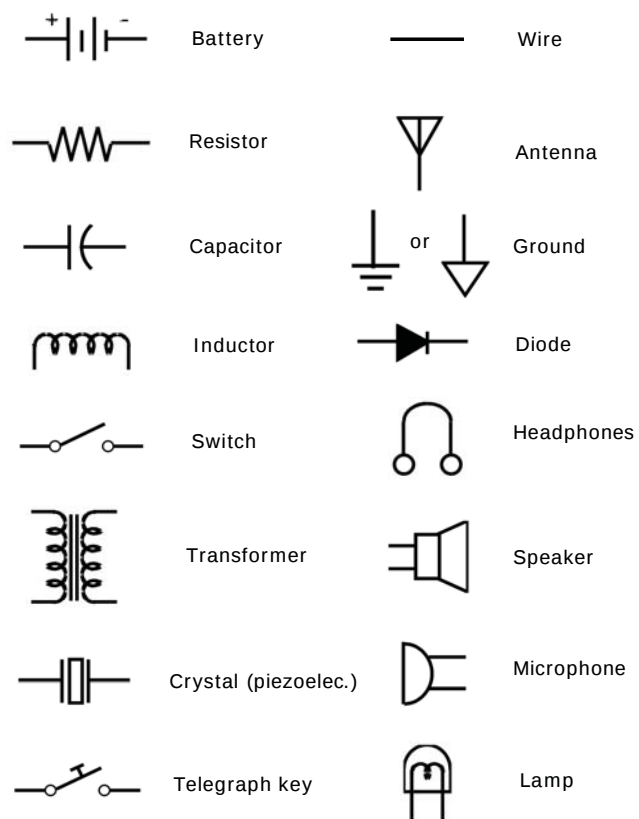


Figure 22.1: Some common schematic symbols.

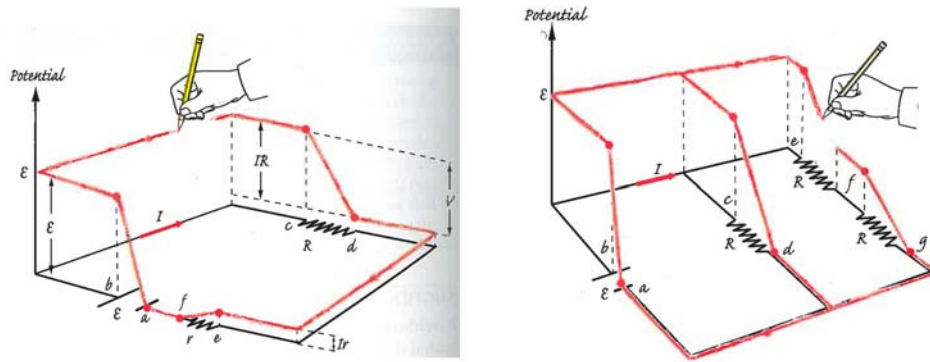


Figure 22.2: Kirchhoff plots of two simple electric circuits. *Left:* A simple circuit with one loop. The resistor r is meant to model the battery internal resistance; $\mathcal{E}$ is the battery electromotive force; and $V = \mathcal{E} - Ir$ is the terminal voltage. *Right:* A more complex circuit. Notice how each battery causes a rise in potential, and each resistor causes a drop in potential. (From Ref. [17].)

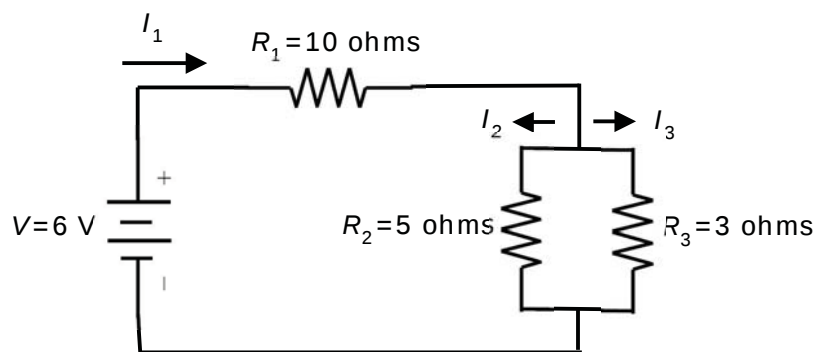


Figure 22.3: A simple circuit.

this current splits at the fork with the two parallel resistors into two currents, $I_2 + I_3$. After passing through these two resistors, the currents recombine, so current I_1 runs through the bottom leg of the circuit before returning to the battery at the $-$ terminal.

Our job is to find the currents through, and voltages across, each of the three resistors. To begin, we'll reduce the three resistors to a single equivalent resistor. The two resistors in parallel are equivalent to a single resistance of

$$R_{23} = \frac{1}{\frac{1}{R_2} + \frac{1}{R_3}} \quad (22.1)$$

$$= \frac{1}{\frac{1}{5\Omega} + \frac{1}{3\Omega}} \quad (22.2)$$

$$= 1.875 \Omega \quad (22.3)$$

Now we've reduced the circuit to two resistors: R_1 in series with R_{23} . The equivalent resistance of these two in series is

$$R_{123} = R_1 + R_{23} \quad (22.4)$$

$$= 10 \Omega + 1.875 \Omega \quad (22.5)$$

$$= 11.875 \Omega \quad (22.6)$$

So now we've reduced all three original resistors to a single equivalent resistance of 11.875Ω connected to the 6 V battery. We can find the current I_1 coming out of the battery using Ohm's law:

$$I_1 = \frac{V}{R_{123}} \quad (22.7)$$

$$= \frac{6 \text{ V}}{11.875 \Omega} \quad (22.8)$$

$$= 0.50526 \text{ A} \quad (22.9)$$

This is also the current through resistor R_1 . Since we know R_1 and the current through R_1 , we can use Ohm's law to find the potential difference across R_1 :

$$V_1 = I_1 R_1 \quad (22.10)$$

$$= (0.50526 \text{ A})(10 \Omega) \quad (22.11)$$

$$= 5.0526 \text{ V}. \quad (22.12)$$

Now we need to find the currents through and voltages across resistors R_2 and R_3 . There are a few ways we could proceed:

1. We can consider the two resistors R_2 and R_3 as equivalent to a single resistor $R_{23} = 1.875 \Omega$, as we've already worked out. The current through this equivalent resistor is $I_1 = 0.50526 \text{ A}$. Knowing the resistance and current, we can find the voltage across the two parallel resistors. Knowing the voltage across each resistor, you can now work out the individual currents in each resistor using Ohm's law.
2. Alternatively, we can note that the sum of the voltage drops for all the resistors must equal the voltage rise due to the battery. Therefore, the voltage drop across the two parallel resistors must be $6 \text{ V} - 5.0526 \text{ V} = 0.9474 \text{ V}$. Knowing this voltage and the resistances R_2 and R_3 , we can use Ohm's law to solve for the currents.

3. A third approach would be to use proportions to figure out how the current splits going through the two parallel resistors. Since the resistors are $R_2 = 5\Omega$ and $R_3 = 3\Omega$, we know that $\frac{3}{8}$ of current I_1 goes through the 5Ω resistor, and $\frac{5}{8}$ goes through the 3Ω resistor. (Proportionally more current goes through the smaller resistor.) In general, for two resistors R_1 and R_2 in parallel, the currents will be

$$I_1 = \frac{R_2}{R_1 + R_2} I \quad (22.13)$$

$$I_2 = \frac{R_1}{R_1 + R_2} I, \quad (22.14)$$

where I is the current going in to the parallel combination, before it splits into I_1 going through R_1 and I_2 going through R_2 . Knowing the currents through each resistor and the two resistances, we can use Ohm's law to solve for the voltages across each resistor. This proportion method is really only useful for two resistors in parallel; for three or more in parallel, the formulæ become too complicated to be practical.

Any of these three approaches will give the same results. The currents through and voltages across each of the resistors of Fig. 22.3 is shown in Table 22-1.

Table 22-1. Results of circuit analysis of the simple circuit of Fig. 22.3.

Resistor	$R (\Omega)$	$I (\text{A})$	$V (\text{V})$
R_1	10	0.5053	5.0526
R_2	5	0.1895	0.9474
R_3	3	0.3158	0.9474

Note the following from this table:

- $V_1 + V_2 = V_1 + V_3 = 6 \text{ V}$. Looping once around the circuit—taking either the path through R_2 or the one through R_3 —gives a total potential drop equal to the battery voltage.
- $V_2 = V_3$. When resistors are connected in parallel, they all have the same potential drop across them.
- $I_2 + I_3 = I_1$. When the current splits at a junction, the sum of the currents leaving the junction equals the current going into the junction.
- $I_2 = \frac{3}{8}I_1$; $I_3 = \frac{5}{8}I_1$. When current splits at a junction, it divides in proportion to the resistance in each branch.

22.4 Circuit Analysis Principles

We can summarize here a few basic principles to keep in mind:

- When making a complete loop around the circuit, the sum of the voltage rises (due to batteries) equals the sum of the voltage drops (due to resistors). This will be true of any loop you take around the circuit.
- When the current splits at a junction, the sum of the currents leaving the junction equals the sum of the currents entering the junction.
- Current will split at a junction in proportion to the resistance in each branch, with more current going through the branch of least resistance.

- For resistors in series, all resistors have the same current, but there will generally be a different voltage across each resistor. A series combination is called a *voltage divider*.
- For resistors in parallel, all resistors have the same voltage, but there will generally be a different current through each resistor. A parallel combination is called a *current divider*.

Chapter 23

Kirchhoff's Rules

Examine the circuit shown in Fig. 23.1. You'll note that the techniques we used in the previous chapter are not suited to analyze this circuit, due to the presence of the 3V battery in the middle of the circuit.

Instead, we must may use of another technique called *Kirchhoff's rules*. There are two of these rules:

1. *Kirchhoff's voltage rule* states that the sum of the voltage rises and drops around any complete loop in the circuit equals zero.
2. *Kirchhoff's current rule* states that at each junction in the circuit, the sum of the currents entering the junction equals the sum of the currents leaving the junction.

23.1 Example Circuit

We'll use the circuit shown in Fig. 23.1 as an example to illustrate how to apply Kirchhoff's rules.

1. Begin by identifying loops in the circuit. The circuit in Fig. 23.1 consists of three loops: the upper loop, the lower loop, and the outer loop. We'll need to choose any two of these three loops to work with—let's choose the upper and lower loops.
2. Next, we choose a direction in which to “evaluate” each loop. This can be either clockwise or counterclockwise; the choice is completely arbitrary, and will not affect the final results. Let's choose to evaluate both the upper and lower loop in the clockwise direction, as indicated by the arrows in the center of each loop.
3. Now identify the currents in the circuit. By inspection of the circuit, we can see that there are three distinct currents: one in the upper branch, one in the middle branch, and one in the lower branch. We'll label these three currents I_1 , I_2 , and I_3 (respectively), and choose a direction for each current, as shown in Fig. 23.1. It doesn't matter whether we choose the directions for the currents correctly—we just guess at each direction. If we guess the wrong direction for a current, then that current will come out negative when we finish the analysis, so the real current flows opposite the direction we guessed.
4. The next step is to apply Kirchhoff's voltage rule to the upper and lower loops of the circuit. Beginning at any point in the loop, we move in the direction chosen in step 2, and write down the terms shown in Fig. 23.2; the sum of these terms is then set to zero.

For the upper loop, beginning in the upper-left corner, we find:

$$-I_1 R_1 - I_1 R_2 + 3 \text{ V} + I_2 R_3 + 6 \text{ V} = 0 \quad (23.1)$$

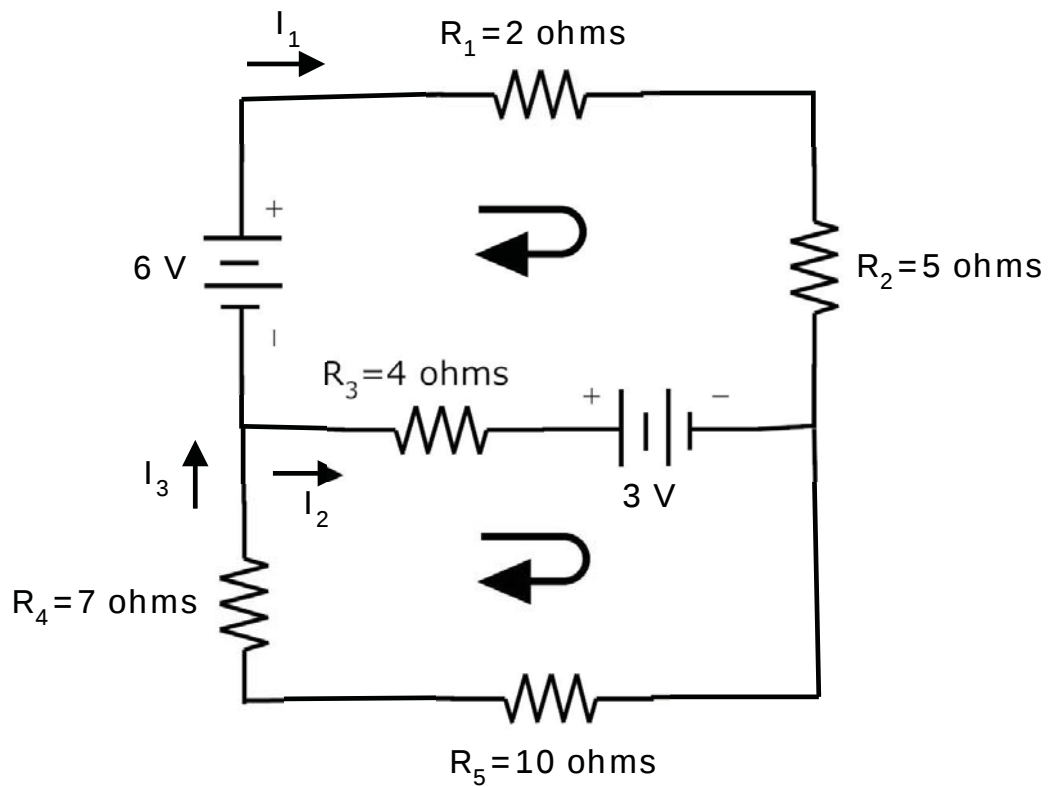


Figure 23.1: Example circuit for analysis using Kirchhoff's rules. As shown by the analysis, the actual direction of current I_2 will turn out to be opposite the direction shown here.

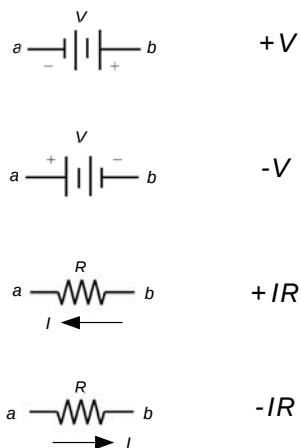


Figure 23.2: Terms for use in Kirchhoff's voltage rule. The evaluation direction is always from a to b .

or, substituting the resistance values,

$$-I_1(2\Omega) - I_1(5\Omega) + 3\text{ V} + I_2(4\Omega) + 6\text{ V} = 0 \quad (23.2)$$

or, simplifying,

$$-I_1(7\Omega) + I_2(4\Omega) + 9\text{ V} = 0 \quad (23.3)$$

Similarly, for the lower loop, beginning in the upper-left corner, we find:

$$-I_2R_3 - 3\text{ V} - I_3R_5 - I_3R_4 = 0 \quad (23.4)$$

Again substituting specific resistance values,

$$-I_2(4\Omega) - 3\text{ V} - I_3(10\Omega) - I_3(7\Omega) = 0 \quad (23.5)$$

or, simplifying,

$$-I_2(4\Omega) - 3\text{ V} - I_3(17\Omega) = 0 \quad (23.6)$$

5. We next apply Kirchhoff's current rule to the junction on the left:

$$I_3 = I_1 + I_2 \quad (23.7)$$

6. Now Eqs. (23.3), (23.6), and (23.7) form a system of three simultaneous linear equations in the three unknown currents, I_1 , I_2 , and I_3 . Writing these three equations in matrix form (and ignoring units for convenience of notation),

$$\begin{pmatrix} -7 & 4 & 0 \\ 0 & -4 & -17 \\ 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} -9 \\ 3 \\ 0 \end{pmatrix}. \quad (23.8)$$

Solving for the currents, we find

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} -7 & 4 & 0 \\ 0 & -4 & -17 \\ 1 & 1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} -9 \\ 3 \\ 0 \end{pmatrix}. \quad (23.9)$$

Evaluating the matrix inverse as the transposed matrix of cofactors divided by the determinant, we find

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = -\frac{1}{215} \begin{pmatrix} 21 & 4 & -68 \\ -17 & 7 & -119 \\ 4 & 11 & 28 \end{pmatrix} \begin{pmatrix} -9 \\ 3 \\ 0 \end{pmatrix}. \quad (23.10)$$

Performing the indicated multiplications, we have

$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 177/215 \\ -174/215 \\ 3/215 \end{pmatrix} = \begin{pmatrix} 0.82326 \\ -0.80930 \\ 0.01395 \end{pmatrix}. \quad (23.11)$$

This tells us the three unknown currents: $I_1 = 823.26\text{ mA}$, $I_2 = 809.30\text{ mA}$, and $I_3 = 13.95\text{ mA}$. The signs of the currents tell us that we guessed the directions of I_1 and I_3 correctly, but we guessed the direction of I_2 incorrectly (since I_2 came out negative). The correct direction of I_2 is *opposite* the direction shown in Fig. 23.1.

Chapter 24

Electronic Instruments

In this chapter we'll examine a few common instruments used to analyze electronic circuits.

24.1 Ammeter

An *ammeter* is used to measure electric current. To use an ammeter, you must *break* the circuit at the point at which you measure the current, then insert the ammeter into the circuit (in series).

24.2 Voltmeter

A *voltmeter* is used to measure the electric potential difference between two points in the circuit. To connect a voltmeter properly, you connect the voltmeter across the two points in the circuit whose potential difference you wish to measure (i.e. in parallel).

24.3 Ohmmeter

An *ohmmeter* is used to measure electrical resistance. You should not connect an ohmmeter to a live circuit; instead, you should completely remove the component in question, and then connect it to the leads of the ohmmeter. The ohmmeter will connect the component to its own internal power supply, and use the resulting current to measure the resistance.

24.4 Multimeter

A common electronic measuring device is the *multimeter*, which is an ammeter, voltmeter, and ohmmeter combined into a single device. A multimeter may also include capacitance meter, inductance meter, and/or a frequency meter.

24.5 Oscilloscope

An *oscilloscope* is an complicated-looking device, consisting of a screen, a probe, and an impressive array of knobs and controls. The oscilloscope is essentially a device for plotting voltage vs. time. The ground wire on the probe is connected to the circuit ground (generally the — of the power supply), and the probe is then

connected to the point in the circuit whose voltage with respect to ground you wish to measure. If the current in the circuit is periodic, the oscilloscope can be made to synchronize itself to this signal so that the plot of voltage vs. time appears “frozen” on the screen.

The oscilloscope often has two or more independent measurement “channels” can also be made to plot one voltage in the circuit vs. another, by using two probes and two different channels.

24.6 Logic Probe

A *logic probe* is used in digital circuits to indicate whether a point in a circuit is at a logic “high” or logic “low” value.

Chapter 25

Capacitance

Suppose we have a conductor carrying a net charge $+Q$ and a second conductor carrying a net charge $-Q$; and suppose the two charges are separated by a fixed distance. Such a device is called a *capacitor* (or *condenser*, an old-fashioned term). The more charge is put on the two conductors of the capacitor, the greater the potential difference between them. In fact, we find that the potential difference is *proportional to* the amount of charge: $Q = CV$, where C is called the *capacitance*:

$$C = \frac{Q}{V}. \quad (25.1)$$

Capacitance is measured in units of *farads* (F), named for English physicist Michael Faraday. One farad is equal to one coulomb per volt ($1 \text{ F} = 1 \text{ C/V}$), and is a very large unit of capacitance. For most laboratory applications, we will be working with units of microfarads (μF), nanofarads (nF), and picofarads (pF).

The reciprocal of capacitance is called the *elastance* S :

$$S = \frac{1}{C}. \quad (25.2)$$

Elastance has units of F^{-1} , sometimes called a *daraf* (“farad” spelled backwards).

25.1 Parallel-Plate Capacitor

One common capacitor configuration consists of two parallel plates (each with area A), separated by a distance d (Fig. 25.1). As you can see in the figure, the electric field between the plates of the capacitor is nearly uniform, except near the edges where there are some edge effects.

To find an expression for the capacitance of the parallel-plate capacitor, we apply Gauss’s law to an imaginary pillbox-shaped Gaussian surface that has one flat end of area A in the region between the plates, and the other in the region to the left of the left plate. The electric flux through all faces of the surface except the face between the plates will be zero; for that face the electric flux will be $\Phi_e = EA$; then by Gauss’s law,

$$\Phi_E = EA = \frac{Q}{\epsilon_0}. \quad (25.3)$$

Since the potential difference between the plates is $V = Ed$, we have

$$V = \frac{Qd}{\epsilon_0 A}. \quad (25.4)$$

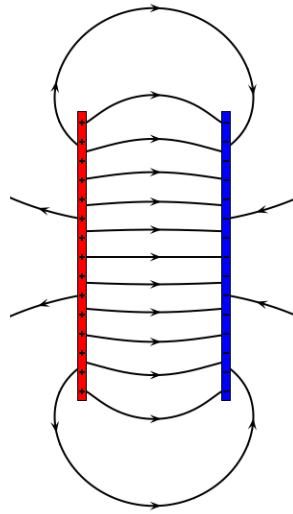


Figure 25.1: Electric field between the plates of a parallel-plate capacitor. The electric field between the plates is uniform, except near the edges. (©GNU-FDL, Wikimedia Commons [11].)

To find the capacitance, we divide this into the charge on each plate, Q :

$$C = \frac{Q}{V} = \frac{\epsilon_0 A}{d}. \quad (25.5)$$

Note that the capacitance depends *only* on the geometry of the capacitor (plate area and spacing), and not on the charge on capacitor or the voltage between the plates. This is true of other capacitor configurations as well: C depends only on the geometrical properties of the capacitor. Note also that the parallel-plate capacitor has a larger capacitance if the plates are larger, or if the plates are closer together.

25.2 Capacitors in Series and Parallel

If several capacitors are connected end-to-end (*in series*), the equivalent resistance is equal to the reciprocal of the sum of the reciprocals of the individual resistances:

$$\frac{1}{C_s} = \sum_i \frac{1}{C_i} \quad (25.6)$$

$$= \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots \quad (25.7)$$

A common error in computing series capacitances is to compute sum of the reciprocals of the individual capacitances, then forget to take the reciprocal of the result at the end. Be careful not to do this!

If the capacitors are connected *in parallel*, the the equivalent capacitance is the sum of the individual capacitances:

$$C_p = \sum_i C_i \quad (25.8)$$

$$= C_1 + C_2 + C_3 + \cdots \quad (25.9)$$

Notice that the formula for capacitors in *series* looks similar to the formula for resistors in *parallel*, and vice versa.

Note the following points. For capacitors connected *in series*:

- The equivalent capacitance will be smaller than the smallest capacitance in the series combination.
- If one capacitor in the series combination is much smaller than the others, the equivalent capacitance will be approximately equal to the smallest capacitance.
- N equal capacitors C connected in series have an equivalent capacitance of C/N .

For capacitors connected *in parallel*:

- The equivalent capacitance will be bigger than the largest capacitance in the parallel combination.
- If one capacitor in the parallel combination is much larger than the others, the equivalent capacitance will be approximately equal to the largest capacitance.
- N equal capacitors C connected in parallel have an equivalent capacitance of NC .

25.3 Dielectric Materials in Capacitors

As shown by Eq. (25.5), the capacitance of a flat-plate capacitor can be increased by increasing the area of the plates, or by decreasing the distance between them. Another way to increase the capacitance is to insert a dielectric material between the plates; this will cause the capacitance to increase by a factor of K :

$$C = K \frac{\epsilon_0 A}{d}, \quad (25.10)$$

where K is called the *dielectric constant* of the material. Inserting a dielectric material between the plates of a capacitor does triple duty: it increases the capacitance by a factor of K ; it serves to keep the two plates *physically* separated by a small fixed distance; and it keeps the plates electrically insulated from each other so that they don't short out.

The combination

$$\epsilon = K \epsilon_0 \quad (25.11)$$

is called the *permittivity* of the material.

25.4 Energy Stored in a Capacitor

A capacitor can be thought of as a device that stores energy in the electric field between the plates of the capacitor. Using the calculus, it can be shown that the potential energy U stored in the electric field of a capacitor of capacitance C , voltage V , and charge Q (on each plate) is given by

$$U = \frac{1}{2} QV = \frac{1}{2} CV^2 = \frac{1}{2} \frac{Q^2}{C}. \quad (25.12)$$

The *energy density* (energy per unit volume) of a capacitor can be found by using the parallel-plate capacitor as an example. The total potential energy stored in a parallel-plate capacitor (of plate area A and separation

d) is

$$U = \frac{1}{2} C V^2 \tag{25.13}$$

$$= \frac{1}{2} \frac{\epsilon_0 A}{d} (Ed)^2 \tag{25.14}$$

$$= \frac{1}{2} \epsilon_0 E^2 Ad. \tag{25.15}$$

Since the volume of the space between the plates is Ad , the energy density $u = U/(Ad)$, or

$$u = \frac{1}{2} \epsilon_0 E^2. \tag{25.16}$$

Chapter 26

RC Circuits

By connecting a resistor and capacitor together in series, we create an *RC circuit*. In an RC circuit, energy is stored in the electric field of the capacitor, and the resistor controls the rate at which charge reaches the capacitor. The characteristic time scale required to charge the capacitor is called the *time constant* τ , and is given by

$$\tau = RC. \quad (26.1)$$

If the resistance R is in ohms and the capacitance C is in farads, then the time constant τ will have units of seconds.

There are two basic types of RC circuits: *charging* and *discharging*.

26.1 Charging RC Circuit

Figure 26.1 shows a charging RC circuit. The circuit includes a battery, so that when the switch S is closed, current flows through the resistor and charges the capacitor. As charge builds up on the plates of the capacitor, it becomes more difficult for the battery to add even more charge to the capacitor, so the current begins to drop. Once an amount of time has gone by that is large compared to the time constant $\tau = RC$, the capacitor will be essentially full charged, and the current will be negligible.

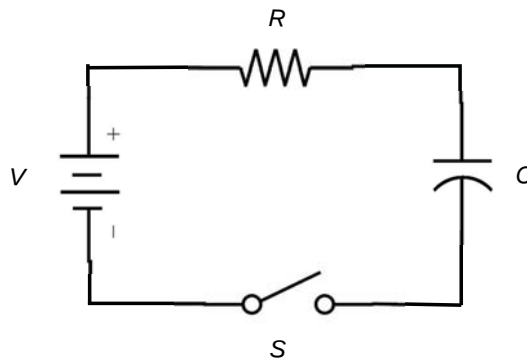


Figure 26.1: A charging RC circuit. The capacitor C begins charging once switch S is closed.

Figure 26.2 shows the resistor voltage, capacitor voltage, circuit current, and capacitor charge in the charging RC circuit as a function of time. The capacitor is initially uncharged, and switch S is closed at time $t = 0$. Shortly after the switch S is closed, a large current flows through the circuit, the voltage across the resistor R is equal to the battery voltage V , and the voltage across the capacitor is zero. At time $\tau = RC$ after the switch is closed, the voltage across the resistor has decreased to $1/e = 0.368$ of the battery voltage; the voltage across the capacitor has increased to $1 - 1/e = 0.632$ of the battery voltage; the current has decreased to $1/e$ of its initial value; and the charge on each of the plates of the capacitor has increased to $1 - 1/e$ of its maximum capacity.

Mathematically, the voltage across the resistor V_R , the voltage across the capacitor V_C , the current in the circuit I , and the charge on each capacitor plate Q can be shown to be

$$V_R(t) = Ve^{-t/\tau} \quad (26.2)$$

$$V_C(t) = V(1 - e^{-t/\tau}) \quad (26.3)$$

$$I(t) = (V/R)e^{-t/\tau} \quad (26.4)$$

$$Q(t) = CV(1 - e^{-t/\tau}) \quad (26.5)$$

As time $t \rightarrow \infty$, current will stop flowing in the circuit, the capacitor will have reached its maximum charge, the voltage across the resistor will be zero, and the voltage across the capacitor will equal the battery voltage.

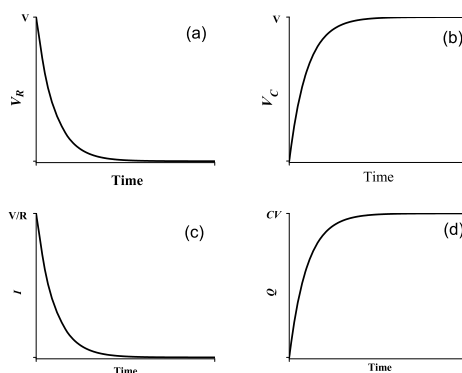


Figure 26.2: Plots vs. time for a charging RC circuit. (a) Resistor voltage vs. time; (b) capacitor voltage vs. time; (c) circuit current vs. time; and (d) charge on the capacitor vs. time. The capacitor is initially uncharged, and the switch S is closed at time $t = 0$.

26.2 Discharging RC Circuit

Figure 26.3 shows a discharging RC circuit. There is no battery in this circuit; instead, we have a capacitor C that is initially fully charged to potential V that is connected in series with a resistor R . When switch S is closed at time $t = 0$, the voltage across the resistor and capacitor, the circuit current, and the capacitor charge *all* decrease exponentially, and reach $1/e$ of their initial value in time $\tau = RC$. As time $t \rightarrow \infty$, the current, all voltages, and the capacitor charge will all dwindle to zero. Mathematically, the voltage across the resistor V_R , the voltage across the capacitor V_C , the current in the circuit I , and the charge on each capacitor plate Q

can be shown in this case to be

$$V_R(t) = Ve^{-t/\tau} \quad (26.6)$$

$$V_C(t) = Ve^{-t/\tau} \quad (26.7)$$

$$I(t) = (V/R)e^{-t/\tau} \quad (26.8)$$

$$Q(t) = C Ve^{-t/\tau} \quad (26.9)$$

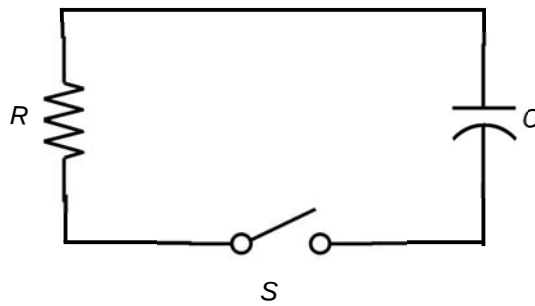


Figure 26.3: A discharging RC circuit. The initially charged capacitor C begins discharging once switch S is closed.

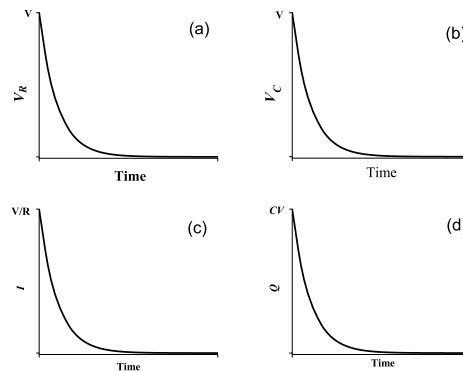


Figure 26.4: Plots vs. time for a discharging RC circuit. (a) Resistor voltage vs. time; (b) capacitor voltage vs. time; (c) circuit current vs. time; and (d) charge on the capacitor vs. time. The capacitor is initially fully charged, and the switch S is closed at time $t = 0$.

Chapter 27

Other Electronic Components

27.1 The Diode

27.2 The Transistor

27.3 Integrated Circuits

Chapter 28

The Incandescent Light Bulb

Chapter 29

Electronics as a Hobby

Many people enjoy electronics as a hobby. They enjoy the creative outlet that electronics offers—you imagine some kind of electronic device or gadget, you design it yourself, and build it. Designing and building something yourself gives you a sense of satisfaction that you simply don't get from buying something ready-made — plus you can design it to work just the way you want. Sometimes hobbyists skip the design stage, and just enjoy building devices from kits.

An extensive electronics textbook, *Lessons in Electric Circuits*, is available on-line at:

<http://www.allaboutcircuits.com/textbook/>.

Here we'll look at a few of the kinds of projects that electronics hobbyists get involved in. Maybe you'll decide you'd like to try some of these things yourself.

29.1 Analog Electronics

Analog electronics involves building things from parts like resistors, capacitors, inductor, transistors, etc. You can design analog electronic circuits to do any number of things: build your own light-activated burglar alarm, a radio receiver, remote weather station, metal detector, electronic organ, computer light pen, electronic measuring equipment, devices for your car, etc. — you're limited only by your imagination.

One place to start with analog electronics might be to build a simple crystal radio receiver; see:

<http://www.midnightscience.com/>.

29.2 Digital Electronics

Digital electronics typically involves components like microprocessors, microcomputer chips, and field-programmable gate arrays (FPGAs). These components are available as integrated circuits that are connected to other digital and analog components to make useful devices. Microprocessor and FPGA training kits are available to help you learn microprocessor and FPGA programming, and how to interface these devices to external displays or other devices. You might even like to try something like building your own calculator or computer completely from scratch.

Microcontrollers are very popular nowadays. These are small computers, typically designed to interact with hardware like motors, sensors, robotic arms, etc. They're surprisingly low cost (typically less than \$50), and can also be configured to work like a small desktop computer. Some popular microcontrollers are:

- Arduino: <http://www.arduino.cc>
- Raspberry Pi: <http://www.raspberrypi.org>

NerdKits (<http://www.nerdkits.com>) sells a good simple microcontroller kit with an informative instruction manual, along with ideas for a few projects to get started.

Maker Shed (<http://www.makershed.com>) is a popular site that contains a *lot* of information about hobby electronics, kits, microcontrollers, etc. Maker Media also publishes a number of books on books on hobby electronics, such as *Make: Electronics*, *Make: More Electronics*, and introductory books on the Arduino and Raspberry Pi microcontrollers.

SparkFun (<http://www.sparkfun.com>) is another site dedicated to electronics hobbyists. They sell electronics and microcontroller kits and parts.

HackerBoxes (<http://www.hackerboxes.com>) offers a subscription service, in which they send out a different box of electronics hobbyist components to subscribers every month.

A field-programmable gate array (FPGA) is a kind of general-purpose logic device that you can design with any logic circuits you wish. For example, you could program it to be a digital clock circuit or a micro-processor. An excellent way to learn FPGA programming is with the Papilio FPGA board, available from the Gadget Factory (<http://www.papilio.cc>). You can use this FPGA board alongside the tutorial *Introducing the Spartan 3E FPGA and VHDL* by Mike Field, available as a free e-book on the Internet. FPGAs are programmed in one of two languages: Verilog or VHDL. This tutorial uses VHDL, which is the more common language in the United States.

29.3 Amateur Radio

How would you like to try transmitting over the air with your own radio station? That's possible, but you'll need to earn an amateur radio license first. You'll study radio theory, electronics, and regulations, then take an exam. If you pass, you'll be assigned your own radio call sign by the FCC, and you can go on the air and talk to people around the country or around the world by voice or by code.

Radio amateurs are involved in lots of activities today:

- *Morse code.* Many amateurs enjoy traditional radiotelegraphy, where you “talk” to people around the world using Morse code and telegraph key.
- *Radioteletype.* This usually involves a computer, rather than a real teletype these days. You can send messages via radioteletype at faster speeds than sending telegraphy by hand.
- *Packet radio.* This is a kind of amateur radio version of the Internet, including a type of amateur radio e-mail.
- *Amateur radio satellites.* You might like to get involved in working with a number of satellites that are in orbit around the Earth, that are especially for use by radio amateurs.
- *Amateur television.* You can go on the air with your own amateur television station.
- *Volunteer work.* Amateur radio operators are needed to help coordinate events like parades, marathons, and long-distance bicycle rides.
- *Emergency response.* During an emergency, all normal lines of communications — including cell phones — may be knocked out. Amateur radio operators are often the only way to get communications in and out of the emergency area. You can train to be prepared to help in case of an emergency.
- *Military work.* Some amateurs work with the military to help coordinate radio communications.
- *Experimental work.* Amateurs are often involved in cutting-edge radio research, including spread-spectrum transmission, very high- or low-frequency transmissions, or bouncing radio signals from auroræ, meteor trails, satellites, or even the Moon. Amateurs may get interested in radiowave propagation in the ionosphere, and conduct their own research.

- *Build your own equipment.* The Amateur Radio Service is the only radio service that allows you to design and build your own transmitting equipment.
- *Low-power operations.* Some amateurs enjoy the challenge of working with simple low-power (< 5 watt) transmitters that they build themselves, just to see how much can be done with low power. This is called *QRP* operation.
- *Contests.* Many amateurs enjoy contests and winning awards, such as those you can win by contacting another amateur radio operator in each state, or in as many different countries as possible. Some organizations hold “contest nights”, where you conduct as many (very brief) contacts as possible in one evening.

For more information on amateur radio, see the American Radio Relay League: <http://www.arrl.org/>. Exams may be taken in this area from local examiners for free or for a small fee.

29.4 Robotics

Combining electronics with sensors and motorized parts involves the popular field of *robotics*. You might want to build a robot that wheels itself around your house while avoiding obstacles, or you might want to build a device that cooks your breakfast for you before you wake up in the morning. The possibilities with robotics are almost endless. Many robotics kits are available to build specific kinds of robots, or you may want to try designing and building your own robots.

To get started in robotics, try using Google to search the Internet for “hobby robotics”. You’ll find quite a bit of information and a number of books and kits available. Also, HackerBoxes (<http://www.hackerboxes.com>) offers a Robotics Workshop for beginners.

29.5 Amateur Rocketry

Model rocketry is another hobby that has become popular in the past few years, and amateur rocketeers have begun building very powerful rockets that approach the power of professional sounding rockets. If you’re interested in this, you can combine this hobby with electronics to build electronic payloads for model rockets, allowing you to telemeter back to Earth information about the Earth’s atmosphere.

More information on rocketry is available from the National Association for Rocketry: <http://www.nar.org/>.

29.6 Amateur Satellites

One very new hobby is the field of *amateur satellites*. It is now actually possible to build your own spacecraft and have it launched into orbit on a commercial rocket. You design and build the satellite from scratch, including sensors, science experiments, electric power systems, attitude determination and control systems, telemetry systems, and radio receivers on the ground. You do the design, building, and testing, then arrange to have it flown “piggyback” on the same rocket along with a large commercial payload.

One popular amateur satellite configuration is called the *CubeSat*, which is constructed of cubical “modules” of size $10\text{ cm} \times 10\text{ cm} \times 10\text{ cm}$, which is called “1 unit”, or 1U. CubeSat satellites can be made of several modules connected together in 1U, 2U, 3U, or 6U configurations. One company, Pumpkin Inc., even sells CubeSat kits to help you get started.

Amateur satellite work can be an expensive hobby. At the time of this writing, building a new satellite and getting it launched will cost roughly as much as buying a new car.

For more information on amateur satellites, see the series of books by Sandy Antunes.

29.7 Sample Electronics Projects

Here are a few fun electronics projects that hobbyists have built:

- Radio receivers of all kinds: AM, FM, shortwave, longwave, maritime, TV, police, fire, etc.
- Radio transmitters (requires an amateur radio license).
- An “alarm clock” that automatically opens your curtains in the morning.
- Home weather station.
- Robots to walk or roll around your house, automatically fix your breakfast, etc.
- Home monitoring system.
- Home planetarium.
- Parabolic microphone for amplifying very faint or distant sounds.
- Electronic circuits to disable and auto-locate your car or motorcycle if it is stolen.
- Lie detector.
- Metal detector.
- Circuits powered by fruit.
- Electronic musical instruments.

The Web site <http://www.instructables.com> is a good source of ideas for many more electronics projects.

Chapter 30

Magnetism

Magnetism, like electricity, has been known since ancient times. The word *magnet* derives from the Greek *μαγνητις λίθος*, or “Magnesian stone”; Magnesia was a region of ancient Greece where one could find *lodestone*, a naturally occurring permanent magnet.¹ Ancient mariners were able to construct primitive magnetic compasses by placing these lodestones on cork and floating them in water. You’re undoubtedly familiar with magnets yourself, from having seen modern compasses and manufactured permanent magnets.

From the perspective of physics, the phenomena of electricity and magnetism are very closely related, and are described by a single theory of *electromagnetism*. Classical electromagnetism, which we’ll study in this course, has at its heart four coupled equations called *Maxwell’s equations*, named for the Scottish physicist James Clerk Maxwell. (The more modern theory, called *quantum electrodynamics*, requires mathematics that is beyond the scope of this course.)

We’ll begin by examining both the similarities and differences between electricity and magnetism.

30.1 Magnetic Poles

Just as electricity consists of two kinds of electric charge, magnetism consists of two kinds of *magnetic pole*. But while the electric charges are called $+$ and $-$, the magnetic poles are called (for historical reasons) N and S . The two kinds of magnetic pole behave similarly to electric charges: like poles (two N poles or two S poles) will repel each other, but unlike poles (an N and an S pole) will attract each other.

The strength of a magnetic pole (analogous to charge q) is called the *pole strength* q^* . Pole strength in SI units is measured in units of ampere-meters (A m).

If two magnetic poles q_1^* and q_2^* are separated by a distance r , then the force F between the two poles is given by a magnetic counterpart of Coulomb’s law:

$$F = \frac{\mu_0}{4\pi} \frac{q_1^* q_2^*}{r^2}, \quad (30.1)$$

where μ_0 is called the *permeability of free space*,² and is equal to exactly $4\pi \times 10^{-7}$ N/A². (Mathematically, in Eq. (30.1), we write an N pole as a positive q^* , and an S pole as negative.)

Although electricity and magnetism are similar in many ways, there is one important difference: while individual electric charges can occur in isolation, *magnetic poles only occur in pairs*. In other words, we *never* see an isolated N pole or S pole by itself: whenever we have an N pole, there will always be an S pole

¹Recent research suggests that lodestone is created when the mineral *magnetite* is struck by a bolt of lightning. See P. Wasilewski and G. Kletetschka, “Lodestone: Nature’s only permanent magnet — What it is and how it gets charged”; *Geophys. Res. Lett.*, **26**, 15, 2275-78 (1999).

² μ_0 is pronounced “mu-nought.”

to go with it. For example, if you take a bar magnet with an N pole and an S pole and break it in half, you will get two smaller bar magnets, each of which has its own N pole and S pole.

(There some theories that predict the existence of isolated magnetic poles, which are called *magnetic monopoles*. These magnetic monopoles, if they exist, would take the form of subatomic particles. However, no magnetic monopoles have yet been detected.)

30.2 Atomic View of Magnetism

Fundamentally, *all magnetism is due to electric currents*. On a macroscopic scale, we can construct an *electromagnet* in the laboratory by running an electric current through a coil of wire. But even permanent magnets are due to electric currents: the motion of an electron around an atomic nucleus creates an electric current, and this electric current creates a magnetic field that ultimately manifests itself as the magnetic field of the permanent magnet. This is described in detail in the discussion of ferromagnetism (Section 34.3). A quantitative treatment of the magnetic field produced by an electric current is given in the next chapter.

Chapter 31

The Magnetic Field

31.1 Magnetic Field

Recall how we defined the electric field $\mathbf{E}$ in Chapter 16: we place a small positive test charge q at a point in space, measure the force $\mathbf{F}$ on it, and then compute the electric field as the force per unit charge: $\mathbf{E} = \mathbf{F}/q$. We can similarly define a *magnetic field* $\mathbf{B}$ by measuring the force $\mathbf{F}$ on a small N magnetic pole q^* ; then the magnetic field is defined as the force per unit pole strength:

$$\mathbf{B} = \frac{\mathbf{F}}{q^*}. \quad (31.1)$$

In SI units, the magnetic field $\mathbf{B}$ is measured in units of *teslas* (T), named for the Serbian physicist Nikola Tesla. One tesla is equal to $1 \text{ N A}^{-1} \text{ m}^{-1}$. A tesla is a very large unit; the largest magnetic fields that can be produced in the laboratory are on the order of a few teslas. A common unit for working with terrestrial magnetic fields is the nanotesla (nT). Another common unit of $\mathbf{B}$ is the *gauss* (G), named for the German mathematician Carl Friedrich Gauss. One gauss is equal to 10^{-4} tesla.

31.2 Magnetic Field due to a Single Magnetic Pole

The magnetic field due to a single magnetic pole q^* can be found by using magnetic version of Coulomb's law. Let's put a small N pole q_0^* at some distance r from the pole q^* ; then by the magnetic Coulomb's law, the force on q_0^* is $F = (\mu_0/4\pi)(q^*q_0^*/r^2)$. Dividing by q_0^* gives us the magnetic field due to a single magnetic pole q^* :

$$B = \frac{\mu_0 q^*}{4\pi r^2}. \quad (31.2)$$

Remember, though, that magnetic pole *never* occur in isolation—they only occur in N - S pole *pairs*.

31.3 Magnetic Field Lines

To help visualize the shape of the magnetic field, it can be helpful to draw diagrams of *magnetic field lines*, similar to the electric field lines we drew earlier. These lines have the following properties:

- The magnetic field lines are directed lines (with arrows) that point *from* the N pole *to* the S pole.
- At any point along a field line, the magnetic field vector $\mathbf{B}$ is tangent to the field line.

- We cannot possibly draw *all* field lines (because they fill all space), so we draw only a few. The number of field lines you draw is somewhat arbitrary — we just draw enough to visualize the field without making the diagram too crowded.
- The closer together the field lines are, the stronger the magnetic field.
- Unlike electric field lines (which terminate on electric charges), magnetic field lines *never* terminate. They form closed loops, or sometimes may form a pattern that continues indefinitely without repeating or terminating.

31.4 The Magnetic Dipole

As an example, consider Fig. 31.1, which shows the magnetic field due to a bar magnet; such an arrangement of two magnetic poles separated by a fixed distance is called a *magnetic dipole*.

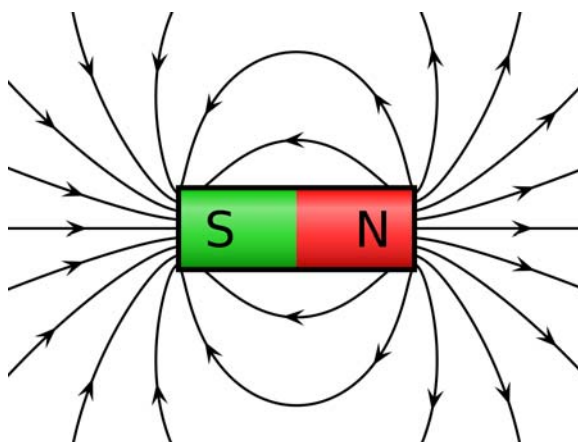


Figure 31.1: Dipole magnetic field due to a bar magnet. (©GNU-FDL, Wikimedia Commons [11].)

A magnetic dipole may be characterized by a quantity called the *magnetic (dipole) moment*. The magnetic moment $\mathbf{m}$ of a magnetic dipole is defined as

$$\mathbf{m} = q^* \mathbf{d}, \quad (31.3)$$

where q^* is pole strength of either end of the dipole, and $\mathbf{d}$ is a vector whose length is equal to the distance between the poles, and which points from the *S* pole to the *N* pole (opposite the direction of the magnetic field line between the poles). The dipole moment essentially measures how magnetically “polarized” a dipole is, with larger values when more pole strength is separated by a greater distance. Magnetic dipole moment is measured in units of A m^2 .

31.5 Magnetic Flux

Magnetic flux may be thought of as being proportional to the total number of magnetic field lines passing through a given area. Given an area A embedded in a magnetic field $\mathbf{B}$, the electric flux Φ_B passing through A is equal to the product of $\mathbf{B}$ and the component of A perpendicular to the field:

$$\Phi_B = \mathbf{B} \cdot \mathbf{A} = BA \cos \theta. \quad (31.4)$$

Here $\mathbf{A}$ is an area vector whose magnitude is equal to the area of the surface, and whose direction is perpendicular to the surface. Magnetic flux is measured in units of *webers* (Wb), where $1 \text{ Wb} = 1 \text{ T m}^2$. The weber is named after the German physicist Wilhelm Eduard Weber.

31.6 Gauss's Law for Magnetism

We've already seen one of the four Maxwell's equations, Gauss's law. Another of Maxwell's equations is the analogous equation for magnetism. It has no proper name, but we may call it *Gauss's law for magnetism*. It states:

$$\Phi_B = 0. \quad (31.5)$$

In other words, the magnetic flux through any closed surface is always equal to zero. This is due to the fact that there are no magnetic monopoles, so magnetic field lines never terminate.

31.7 Biot-Savart Law

As mentioned in the previous chapter, magnetic fields are produced by electric currents. The *Biot-Savart law*¹ gives the magnetic field $\Delta\mathbf{B}$ produced by an electric current I running through a short length of wire $\Delta\mathbf{l}$, where $\Delta\mathbf{l}$ is a vector whose length is equal to the length of the wire, and which points in the direction of the conventional current. The Biot-Savart law states:

$$\Delta\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{I \Delta\mathbf{l} \times \hat{\mathbf{r}}}{r^2} \quad (31.6)$$

Here $\mathbf{r}$ is a vector pointing from the current element $I \Delta\mathbf{l}$ to the field point (the point at which the magnetic field is being observed). Note the presence of the vector cross product operator $\times$ in this equation; the cross product is described in Appendix O.

The Biot-Savart law is a magnetic counterpart of Coulomb's law: just as Coulomb's law gives the electric field due to a point charge q , the Biot-Savart law gives the magnetic field due to a current I flowing through a short wire Δl .

Comparing Eq. (31.2) with Eq. (31.6), we can find the pole strength q^* due to a current I through a short wire of length Δl :

$$q^* = I \Delta l. \quad (31.7)$$

31.8 Magnetic Field due to a Long Wire

By making use of the calculus, one may use of the Biot-Savart law (Eq. (31.6)) to find the magnetic field B due to a very long wire carrying an electric current I , at a perpendicular distance r from the wire. The result is:

$$B(r) = \frac{\mu_0}{2\pi} \frac{I}{r}. \quad (31.8)$$

The *direction* of the magnetic field is given by the right-hand rule: if you point the thumb of your right hand in the direction of the conventional current I , then the fingers of your right hand curl in the direction of the magnetic field lines.

¹Pronounced *BEE-oh sav-AR*, and named for the French physicists Jean-Baptiste Biot and Félix Savart.

31.9 Magnetic Field of a Solenoid

A *solenoid* is a long coil of wire. Just as a parallel-plate capacitor gives a nearly uniform electric field between the plates of the capacitor, a solenoid gives a nearly uniform magnetic field inside the coils (Fig. 31.2). Using the Biot-Savart law, the magnetic field in the region inside the solenoid is given by

$$B = \mu_0 n I, \quad (31.9)$$

where n is the number of turns per unit length in the solenoid, and I is the current in the wire.

The direction of the magnetic field inside the solenoid may be given by another right-hand rule: if you curl the fingers of your right hand in the direction of the current, then the thumb of your right hand points in the direction of the magnetic field inside the solenoid.

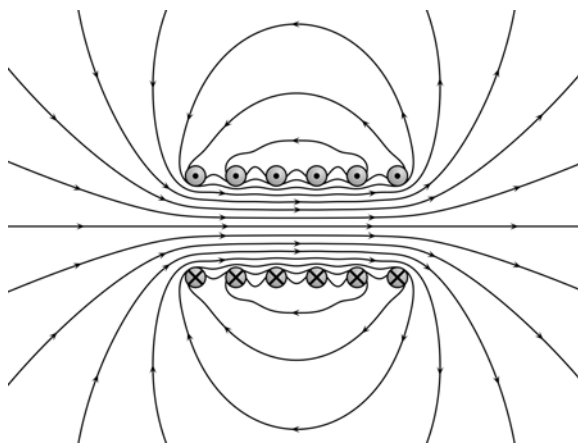


Figure 31.2: Magnetic field due to a solenoid. The solenoid is seen in cross section; current flows out of the page for the wires at the top of the figure, and into the page for wires at the bottom. (©GNU-FDL, Wikimedia Commons [11].)

31.10 Magnetic Field of a Loop or Coil of Wire

As discussed earlier, a magnetic dipole can be created by a bar magnet—but it can also be created by a coil of wire. Given a coil of N turns of wire carrying a current I , the magnetic dipole moment of the coil can be shown to be

$$\mathbf{m} = N I A \hat{\mathbf{n}}, \quad (31.10)$$

where A is the cross-sectional area of the coil, and $\hat{\mathbf{n}}$ is a unit normal vector, pointing perpendicular to the plane of the coil. The direction of $\hat{\mathbf{n}}$ is given by yet another right-hand rule: if the fingers of your right hand curl in the direction of the current, then the thumb of your right hand points in the direction of $\hat{\mathbf{n}}$ (and therefore also in the direction of the magnetic moment $\mathbf{m}$).

31.11 Torque on a Magnetic Dipole in a Magnetic Field

Suppose we put a magnetic dipole $\mathbf{m}$ in a magnetic field $\mathbf{B}$. (The magnetic dipole could be due to a bar magnet, coil of wire, etc.) Then the magnetic field will exert a torque $\boldsymbol{\tau}$ on the dipole, equal to

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}. \quad (31.11)$$

Note again the presence of the vector cross product $\times$ in this equation: the direction of τ will be perpendicular to the plane containing $\mathbf{m}$ and $\mathbf{B}$, in a right-hand sense.

Now suppose we put a magnetic dipole (e.g. a bar magnet or wire coil) of magnetic moment $\mathbf{m}$ in a magnetic field $\mathbf{B}$. What will happen? If $\mathbf{m}$ is parallel or anti-parallel to $\mathbf{B}$ (so the bar magnet is aligned with $\mathbf{B}$, or the plane of the wire coil is perpendicular to $\mathbf{B}$), then the torque on the dipole will be zero, and nothing will happen—the dipole will remain stationary. But if we displace the dipole from this position, then there will be a non-zero torque on the dipole, in a direction that will rotate the dipole back toward the direction of $\mathbf{B}$. But once the dipole moment $\mathbf{m}$ is aligned with $\mathbf{B}$, the dipole's inertia will make it overshoot and rotate past $\mathbf{B}$, where it will experience a torque that will make it rotate back toward $\mathbf{B}$ again, etc. The resulting motion will be that magnetic dipole will oscillate back and forth about the $\mathbf{B}$ direction, with simple harmonic motion. The period of this oscillating motion will depend, in part, on the strength of the magnetic field; in fact, this method was once used to measure magnetic field strength. One would measure the period of oscillation of a well-calibrated dipole in a magnetic field, and use the resulting period to find B .

31.12 Magnetic Pressure

The magnetic field can be thought of as producing a *pressure*, given by

$$P = \frac{B^2}{2\mu_0}, \quad (31.12)$$

where P is the pressure in pascals (Pa; $1 \text{ Pa} = 1 \text{ N/m}^2$). This magnetic pressure can be used to relate the “force” rating of a permanent magnet (which is the maximum weight it is supposed to be able to lift) to the magnetic field strength B at the pole face. Suppose F is the magnet's force rating, and the pole face has area A . Then the magnetic pressure is $P = F/A$, so

$$P = \frac{F}{A} = \frac{B^2}{2\mu_0}, \quad (31.13)$$

So the force rating F is related to the magnetic field strength at the pole face B by

$$F = \frac{AB^2}{2\mu_0}. \quad (31.14)$$

Example. Suppose we have a 100-lb magnet whose pole face is $15 \text{ in} \times 4.5 \text{ in}$. (The 100-lb rating means that the magnet is capable of lifting loads that weigh up to 100 pounds.) What is the magnetic field strength B at the pole face?

Solution. First, convert everything to SI units: the pole face is $38.1 \text{ cm} \times 11.43 \text{ cm}$, and $F = 444.8222 \text{ N}$. Then the area A of the pole face is $A = (0.381 \text{ m}) \times (0.1143 \text{ m}) = 0.043548 \text{ m}^2$. By Eq. (31.14), the magnetic field at the pole face is given by

$$B = \sqrt{\frac{2\mu_0 F}{A}}, \quad (31.15)$$

or

$$B = \sqrt{\frac{2(4\pi \times 10^{-7} \text{ N/A}^2)(444.8222 \text{ N})}{0.043548 \text{ m}^2}} = 0.160 \text{ T} \quad (31.16)$$

Chapter 32

The Lorentz Force

When we place an electric charge in an electric field, or a magnetic pole in a magnetic field, the resulting motion is pretty simple: the charge or pole simply accelerates along the direction of the field. But some more interesting physics goes on when we place an *electric* charge in a *magnetic* field.

Suppose we have an electric charge q moving with velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$. Then it turns out that the charge will experience a force $\mathbf{F}$ given by

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}. \quad (32.1)$$

Note once again the presence of the cross product operator $\times$. This means that the force acting on charge q is perpendicular to both its direction of motion $\mathbf{v}$ and to the magnetic field $\mathbf{B}$.

Note also that since the force is always perpendicular to the direction of motion, the work done by a magnetic field on an electric charge is always zero.

If both an electric field $\mathbf{E}$ and a magnetic field $\mathbf{B}$ are both present, then the net force on the charge q is found by combining Eqs. (16.1) and (32.1), and is called the *Lorentz force*:¹

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}). \quad (32.2)$$

32.1 Plasmas

A plasma is essentially an ionized gas. We can gain some understanding of the behavior of plasmas by examining the motion of charged particles in the presence of electric and magnetic fields.

Suppose, for example, that we have a (negatively charged) electron moving with velocity $\mathbf{v}$ perpendicular to a magnetic field $\mathbf{B}$, and that there is no electric field present. Then there will be Lorentz force acting on the electron that will eventually cause it to move perpendicular to its original direction. By that time, the Lorentz force will be in the direction opposite the direction of the direction of motion of the electron, and so on. The net motion will be that the electron will move in a circle. The direction of motion of a negative charge in a magnetic field will be given by still another right-hand rule: if you point the thumb of your right hand in the direction of $\mathbf{B}$, then the fingers of your right hand will curl in the direction of motion of the electron. (If the magnetic field points into the page, for example, then the electron will move clockwise.)

By similar reasoning, a positively charge (such as a proton) initially moving perpendicular to $\mathbf{B}$ will move in a circle given by a *left-hand* rule: point the thumb of your *left* hand in the direction of $\mathbf{B}$, and the fingers of your left will curl in the direction of motion of the positive charge. For example, if the magnetic field $\mathbf{B}$ points into the page, then a proton will move counterclockwise.

¹Hypothetically, if magnetic monopoles exist, then the force $\mathbf{F}$ on a magnetic monopole q^* in an electric field $\mathbf{E}$ and a magnetic field $\mathbf{B}$ would be given by a similar expression: $\mathbf{F} = q^*[\mathbf{B} - (\mathbf{v}/c^2) \times \mathbf{E}]$.

What if the initial velocity of the charged particle is not necessarily perpendicular to the magnetic field $\mathbf{B}$? In that case, the motion will have two components: motion in a circle perpendicular to the magnetic field, combined with uniform motion parallel to the magnetic field. The net motion will be that of a *helix*: the particles will spiral around magnetic field lines in helices.

How big are the circles that a charged particle moves in? We can find that by equating the magnetic force of Eq. (32.1) (with $\mathbf{v} \perp \mathbf{B}$) to the centripetal force mv^2/r . We then have $qvB = mv^2/r$; solving for the radius r of the circle, we have $r = mv/(qB)$. More generally, if the particle is moving in a helix, then the radius of the helix is determined by the component of the particle's velocity $\mathbf{v}$ that is perpendicular to the magnetic field ($v_\perp$). Also, since the radius is always positive, we want to use the absolute value of the charge q . The general result is that the radius of the helix is

$$r = \frac{mv_\perp}{|q|B}. \quad (32.3)$$

This radius is called the *gyroradius*, *cyclotron radius*, or *Larmor radius*. The gyroradius will be larger for a weaker magnetic field, or for a heavier or faster particle.

Another important quantity is the angular frequency with which the particle gyrates in a circle about the magnetic field lines. The time it takes the particle to complete one circle (i.e. the period of the motion) is the total distance divided by the speed: $T = 2\pi r/v_\perp$. Substituting r from Eq. (32.3), we have $T = 2\pi m/(|q|B)$. Since the angular frequency $\omega = 2\pi/T$, we have that angular frequency of the motion as

$$\omega = \frac{|q|B}{m}. \quad (32.4)$$

This is called the *gyrofrequency* or *cyclotron frequency*. A particle will spin around in circles faster for a stronger magnetic field or a lighter particle.

32.2 Force on a Wire in a Magnetic Field

Now suppose we have a wire carrying an electric current I placed within a magnetic field $\mathbf{B}$. Within the wire, the current is being carried by electrons moving with the drift velocity, each of which experiences a Lorentz force. There will then be a force $\mathbf{F}$ on the wire given by

$$\mathbf{F} = I\mathbf{l} \times \mathbf{B}. \quad (32.5)$$

Here I is the current, and $\mathbf{l}$ is a vector whose length is equal to the length of the wire and which points in the direction of the conventional current. Applying this to a current loop, for example, gives the same torque as given by Eq. (31.11).

32.3 Magnetic Force between Two Long Wires

If we put two long wires next to each other so that they are parallel, then each wire generates a magnetic field that envelopes the other wire. By combining Eq. (31.8) (which gives the magnetic field generated by a wire) with Eq. (32.5) (which gives the force on a wire in a magnetic field), we can find the mutual force between the two parallel wires. The result is

$$\frac{F}{\ell} = \frac{\mu_0}{2\pi} \frac{I_1 I_2}{d}, \quad (32.6)$$

where F/ℓ is the force per unit length, I_1 and I_2 are the two currents, and d is the distance between the two wires.

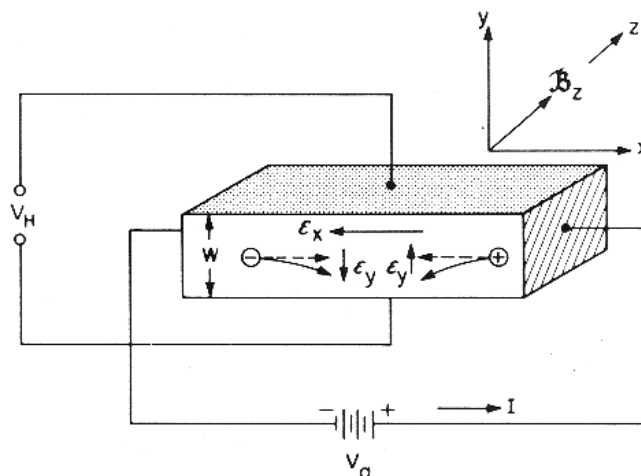


Figure 32.1: The Hall effect. (Ref. [13])

If the two currents I_1 and I_2 are in the *same* direction, the wires will *attract*; if the currents are in the *opposite* direction, the wires will *repel*. (This is in a sense “backwards” from the rule you might expect, based on the rules the force between two electric charges or two magnetic poles—but this is the way it works out.)

Eq. (32.6) is used in the definition of the ampere: 1 ampere is defined to be that current which, when passed through each of two long parallel wires 1 meter apart, gives a force per unit length of 2×10^{-7} newtons per meter, as can be verified by substituting $I_1 = I_2 = 1$ A and $d = 1$ m into Eq. (32.6).

32.4 The Hall Effect

Suppose we run an electric current through a wire—say the current runs from right to left. Such a current could be due to *positive* charges moving from right to left, or to *negative* charges moving left to right. How can we tell the actual charge of the carriers of electric current?

An experiment to determine the correct charge of the carriers of electric current was performed in 1895 at the Johns Hopkins University by Edwin H. Hall. If an electric current is run through a conducting strip in a magnetic field, then opposite sides of the strip will acquire opposite electric charge, and therefore a potential difference will be created across the strip. The *direction* of this potential difference will be different, depending on whether the current is carried by positive or negative charges. This phenomenon is called the *Hall effect*.

The principle of the experiment is shown in Figure 26.1. The positive end of a battery is connected to the right end of the strip, and the negative end to the left end; a magnetic field is directed into the page. If the current is carried by *positive* charges moving right to left, then the Lorentz force on the positive charge will cause the positive charges carrying the current to move downward toward the bottom of the strip, and the electric field due to these charges will point *upward*.

If, on the other hand, the current is carried by *negative* charges moving left to right, then the Lorentz force on the negative charges will cause the negative charges carrying the current to also move downward, toward the bottom of the strip. In this case the electric field due to the charges carrying the current will point *downward*.

When Hall performed his experiment in 1895, he discovered that the latter situation is what actually occurs: the electric field across the strip points downward, so that the carriers of the electric current must be *negative*. This experiment was done in 1895—the year *before* the discovery of the electron by British

physicist J.J. Thompson.

Because of the electric field built up across the conducting strip, there is a potential difference across the strip. It is straightforward to calculate the magnitude of this potential difference: charges will build up across the strip until the magnetic force $qv_d B$ is balanced by the electrostatic force qE , where v_d is the drift velocity, q is the charge on the particles carrying the current, and B and E are the magnetic and electric field strengths, respectively. Since in equilibrium the forces will balance, $qv_d B = qE$, or $E = v_d B$. The potential difference ε_H across the strip is then $\varepsilon_H w$, where w is the width of the strip. Therefore this potential, called the *Hall emf* is given by

$$\varepsilon_H = v_d B w \quad (32.7)$$

Besides its historical interest, the Hall effect can be used today as a means of measuring magnetic field strength. We measure the strip width w , and we can determine the drift velocity v_d by calibration in a known magnetic field. Then Eq. (32.7) can be used to determine the magnetic field strength B by measuring the Hall emf ε_H .

Chapter 33

Geomagnetism

33.1 Earth's Magnetic Dipole

The Earth generates its own internal magnetic field, which is thought to be due to a westward-moving electric current inside the Earth's molten outer core. The resulting field approximates that of a magnetic dipole, with the “poles” of the dipole near (but not *at*) the Earth's geographic poles.

There is a bit of confusing nomenclature to be aware of. If you suspend a bar magnet by a string so that is free to rotate horizontally, it will rotate to align itself with the Earth's magnetic field, with the *N* pole pointing toward geographic north. (That's actually why the poles of a magnet are labeled *N* and *S*: the *N* pole is the “north-seeking” pole and the *S* pole is the “south-seeking pole”.) But since unlike poles attract, the pole near the Earth's geographic *north* pole must be a magnetic *S* pole, and vice versa (Fig. 33.1).

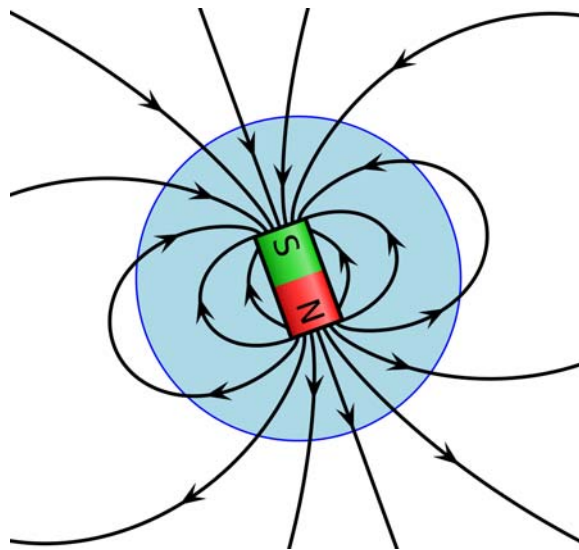


Figure 33.1: Schematic representation of the Earth's internal magnetic field. This should not be taken literally; there is no bar magnet at the Earth's center. This figure is just meant to illustrate that the Earth's geographic *north* pole is a magnetic dipole *S* pole, and vice versa. Note that the Earth's dipole is tilted with respect to the geographic axis, which is vertical in this illustration. (©GNU-FDL, Wikimedia Commons [11].)

The Earth's magnetic poles are not located at the geographic poles, but are some distance away; this is

because the Earth's magnetic dipole is not aligned with the geographic axis, but is tilted at some angle. The magnetic *S* pole is currently located in the Arctic Ocean north of Canada, and the magnetic *N* pole is located just off the coast of Antarctica. For reasons that are not currently understood, the magnetic poles "wander" across the Earth's surface, so the location of the magnetic poles changes from one year to the next.

The strength of the Earth's magnetic field may be characterized by its dipole moment, which has a value of $\mathbf{m} = 7.94 \times 10^{22} \text{ A m}^2$. It can be shown that the magnetic field $\mathbf{B}$ of a magnetic dipole at position $\mathbf{r}$ from the dipole is given by

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \left[\frac{-\mathbf{m} + 3(\mathbf{m} \cdot \hat{\mathbf{r}})\hat{\mathbf{r}}}{r^3} \right]. \quad (33.1)$$

Using this equation, we can find the expected strength of the Earth's magnetic field at the Earth's surface by substituting $r = R_E$, where R_E is the Earth's radius. Assuming the Earth's magnetic field to be a perfect dipole, the magnetic field at the Earth's equator should be roughly $(\mu_0/(4\pi))(m/R_E^3)$, or $B = 30,000 \text{ nT}$. The magnetic field at the poles should be twice this value, or $B = 60,000 \text{ nT}$. The actual values at the equator and poles differ somewhat from these values because the Earth's magnetic field is not a perfect dipole. In fact, it's about 90% dipole, and about 10% higher-order components.

33.2 Magnetic Declination

Because the Earth's magnetic dipole axis is tilted with respect to the geographic axis, a magnetic compass will generally not point toward true geographic north; it will point toward *magnetic* north. The difference (the angle between the two norths) is called the *magnetic declination*. A map showing lines of equal magnetic declination (Fig. 33.2) is called an *isogonic chart*.

As you can see from this chart, there is a 0° line of magnetic declination (the *agonic line*) running near the Mississippi River; along this line, there is no magnetic declination, and a magnetic compass will point to true north. Maryland is at about 11° west declination, meaning that a magnetic compass points about 11° west of true north. To get the compass needle to point to *geographic* north, you would need to adjust the compass dial by 11° .

Since the magnetic poles are wandering with time, the isogonic lines change from one year to the next. If you plan on using a magnetic compass for sailing, hiking, orienteering, or similar activities, you should make sure you have an up-to-date isogonic chart or something similar that shows the current magnetic declination for your location. Of course, if you're traveling large distances, your magnetic declination will be changing as you move, so you will need to re-adjust your compass for declination from time to time.

33.3 Magnetic Inclination

We often think of the Earth's magnetic field as running north-south, but it also has a large *vertical* component: downward in the northern hemisphere, and upward in the southern hemisphere. This vertical component is called *magnetic inclination*.

Because of magnetic inclination, a compass needle will be correctly balanced only for use in a certain part of the world. For example, a magnetic compass made for use in the United States will have a needle that's heavier on the *S* side than the *N* side, to compensate for the downward component of the Earth's magnetic field and allow the needle to balance properly. If you take this compass and try to use it in Australia, the compass needle will not balance properly.

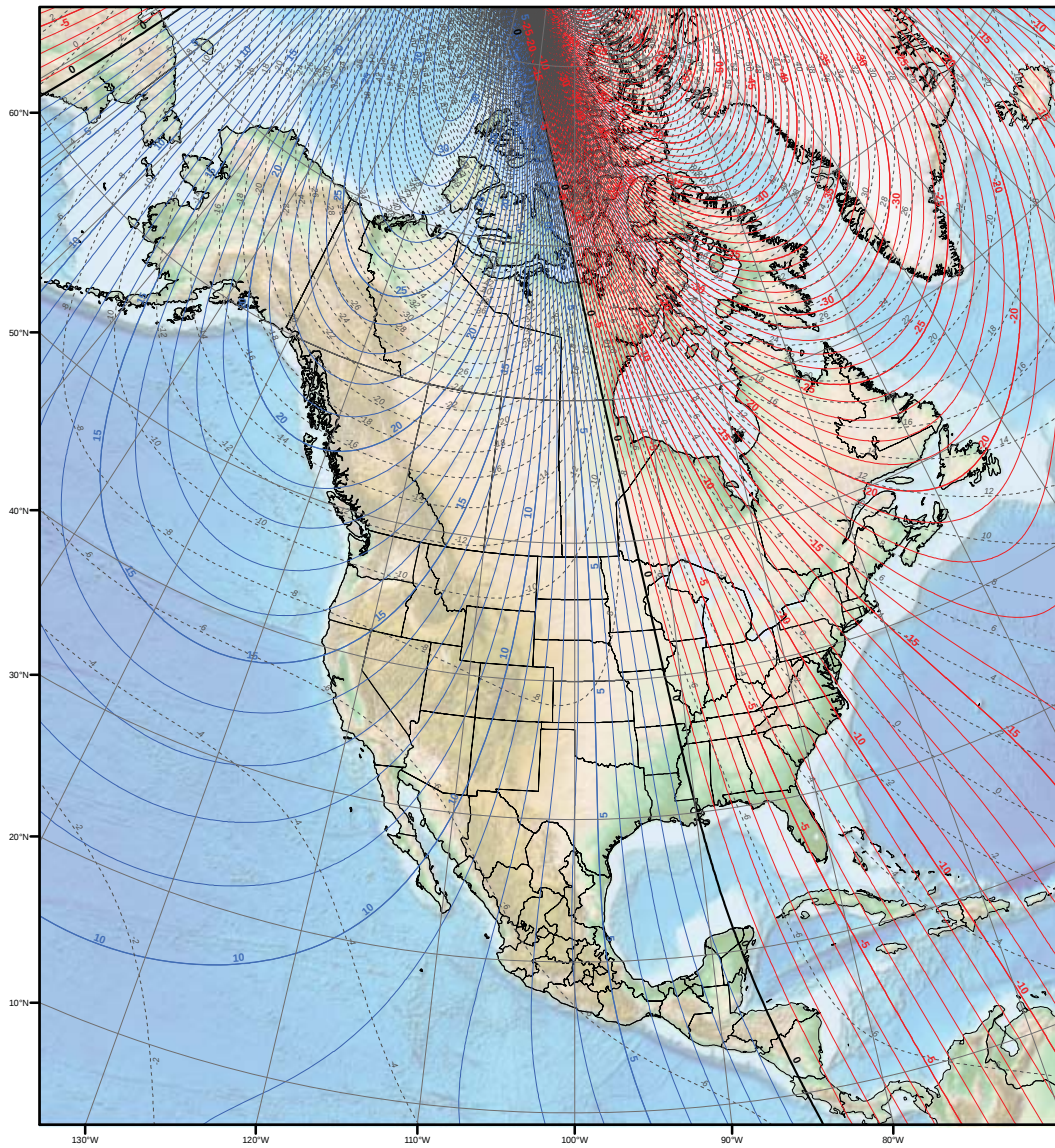


Figure 33.2: Magnetic declination map for North America. *Credit: NOAA.*

33.4 Magnetic Reversals

As mentioned earlier, the Earth's geographic *north* pole is a magnetic *S* pole. This hasn't always been the case, though. The Earth's magnetic field actually reverses direction at irregular intervals; the last such reversal was about 780,000 years ago. Figure 33.3 shows the history of the Earth's magnetic field reversals going back to the late Cretaceous period.

Exactly what causes these reversals of the Earth's magnetic field is unclear, although they have been reproduced in computer simulations. It appears that the reversal process is quite sudden in geological terms — it may take a few decades to a century or so for the magnetic field to reverse, after which it typically stays fairly stable for thousands of years before reversing again. Since these magnetic reversals occur at irregular intervals, we have no way of knowing when the next one will be. There is occasional speculation that the polar wandering may indicate that a magnetic reversal may be going on now, but nobody knows for certain.

How do we know when magnetic reversals have occurred in the past? At the mid-Atlantic ridge in the middle of the Atlantic ocean, the Earth's crust is spreading apart, and new crust is formed as magma seeps up into the crack. As it cools to form rock, this magma “locks in” the direction of the magnetic field at the time it cooled. The result is a set of bands of magnetism on either side of the mid-Atlantic ridge, which records the past magnetic field direction in very much the same way a tape recorder works (Fig. 33.4).

It is not clear what effect, if any, magnetic reversals have on life on Earth. The fossil record doesn't show any correlation between magnetic reversals and mass extinctions, so we can probably infer that any effect on life is relatively minor.

33.5 The Magnetosphere

Although the Earth's magnetic field resembles that of a magnetic dipole near the Earth, further away the dipole becomes distorted due to the presence of the *solar wind*, a “wind” of charged particles (mostly protons and electrons) ejected by the Sun. The solar wind compresses the day side of the Earth's magnetic field, and draws the night side out into a long *magnetotail*. The presence of the solar wind causes the Earth's entire magnetic field to be encapsulated into a structure called the *magnetosphere* (Fig. 33.5).

The Earth's magnetic field serves a very important biological role: it deflects potentially dangerous charged particles from the Sun so that they move harmlessly around the Earth. Without the Earth's magnetic field, we would be bombarded by high-energy solar radiation, which could lead to severe health problems and even death.

The magnetosphere is a fairly complex structure, with various plasmas and electric currents interacting with the Earth's magnetic field; these in turn produce magnetic fields of their own, etc. One of the goals of the field of *space physics* is to investigate this complex structure of the magnetosphere in detail and to understand how it all works.

33.6 The Aurora

In far northern latitudes, one may see the “northern lights”, or *aurora borealis* on some nights, especially during periods of high solar activity (Fig. 33.6). A similar phenomenon is visible in the southern hemisphere, called the *aurora australis*.

Auroræ are produced when charged particles from the Sun reach the Earth's magnetosphere. If the Sun's magnetic field lines are pointing southward at the Earth, they meet the Earth's northward-pointing magnetic field lines in an event called *magnetic reconnection*. When the Earth's magnetic field lines reconnect with the Sun's magnetic field lines, the Earth's lines drape back toward the magnetotail, carrying a load of charged particles with them. A similar reconnection event in the magnetotail causes the magnetic field lines to snap back like rubber bands, and carry a load of charged particles back toward the Earth, where they enter the polar

GEOMAGNETIC POLARITY TIME SCALE LATE CRETACEOUS TO RECENT

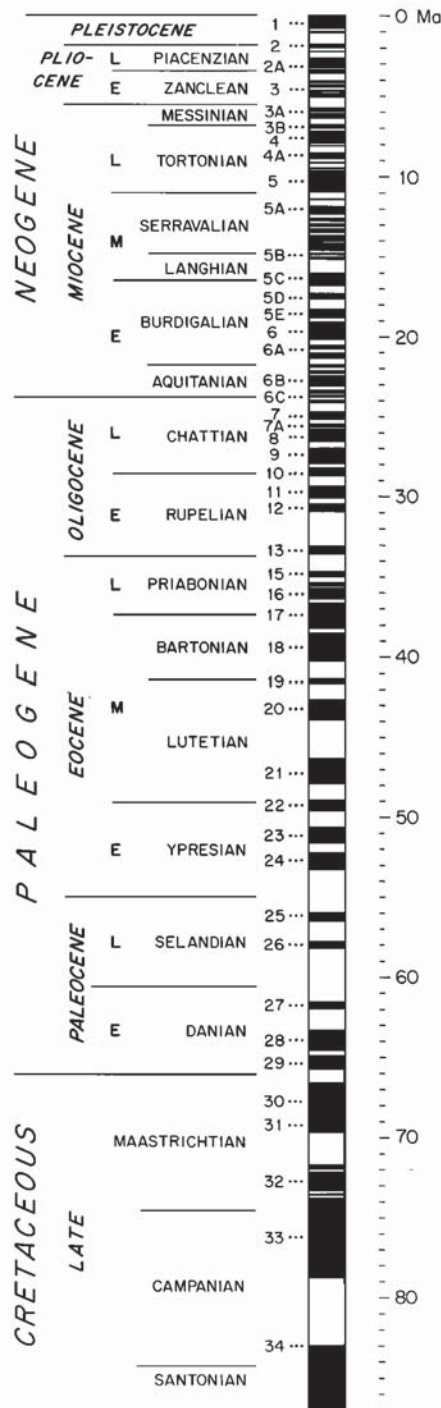


Figure 33.3: History of geomagnetic reversals, going back to the late Cretaceous period. The scale on the right shows time in millions of years ago. Black indicates the same polarity as the current field, and white is a “reversed” field [12].

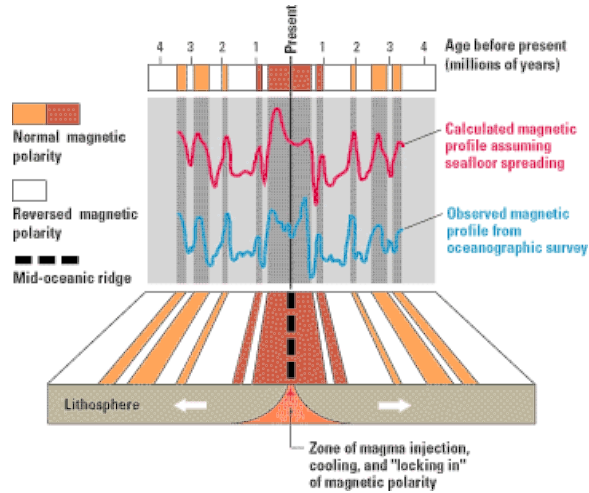


Figure 33.4: The direction of the past geomagnetic field is recorded in the Earth's crust on either side of the mid-Atlantic ridge, in much the same way as information is stored on a magnetic tape by a tape recorder. (Credit: U.S. Geological Survey.)

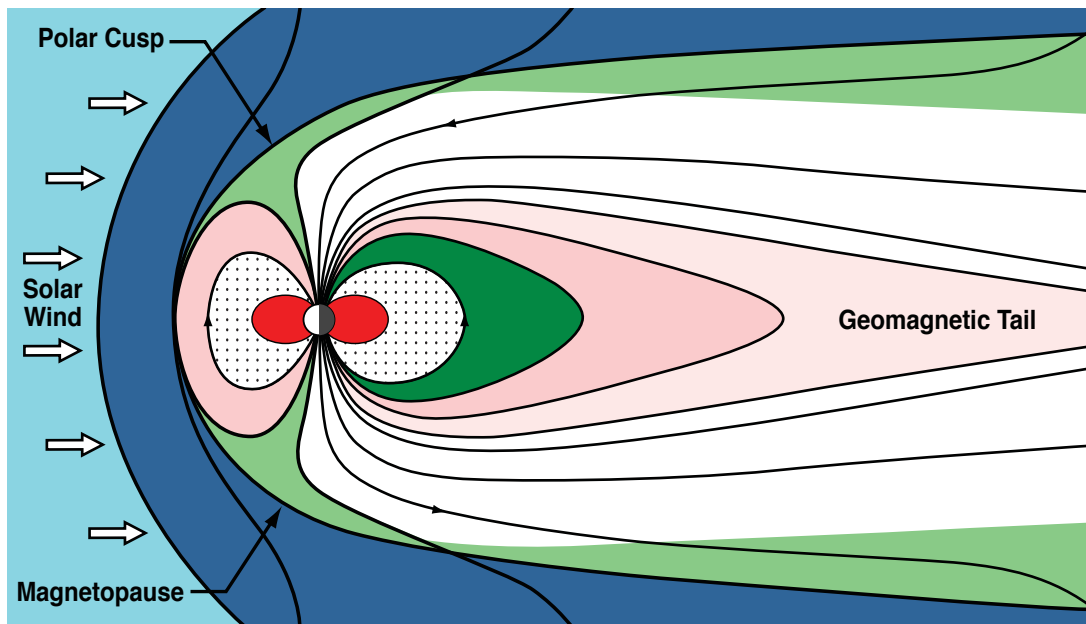


Figure 33.5: The Earth's magnetosphere. The leftmost curve is the *bow shock*, a shock wave in the solar wind. The *magnetopause* shown here is the outer boundary of the Earth's magnetic field. The region between the magnetopause and the bow shock is called the *magnetosheath*. The Sun is outside the figure, to the left. (Credit: NASA.)



Figure 33.6: Aurora borealis over Bear Lake, Eielson Air Force Base, near Fairbanks, Alaska. (*Credit: Joshua Strang, USAF, Wikipedia.*)

regions. These energetic particles excite the oxygen and nitrogen atoms in the atmosphere, producing the green and red lights of the aurora.

Figure 33.7 shows the aurora as seen by NASA's IMAGE spacecraft in ultraviolet light. The images are taken above the Earth, looking down at one of the poles. You can see that the auroræ form an *auroral oval* centered on the pole. The figure shows how the auroral oval grows and then dies out with time.

Similar auroral ovals have been observed on Jupiter and Saturn (Fig. 33.8).

Further information on the Earth's magnetosphere and auroræ is given in Appendix V.

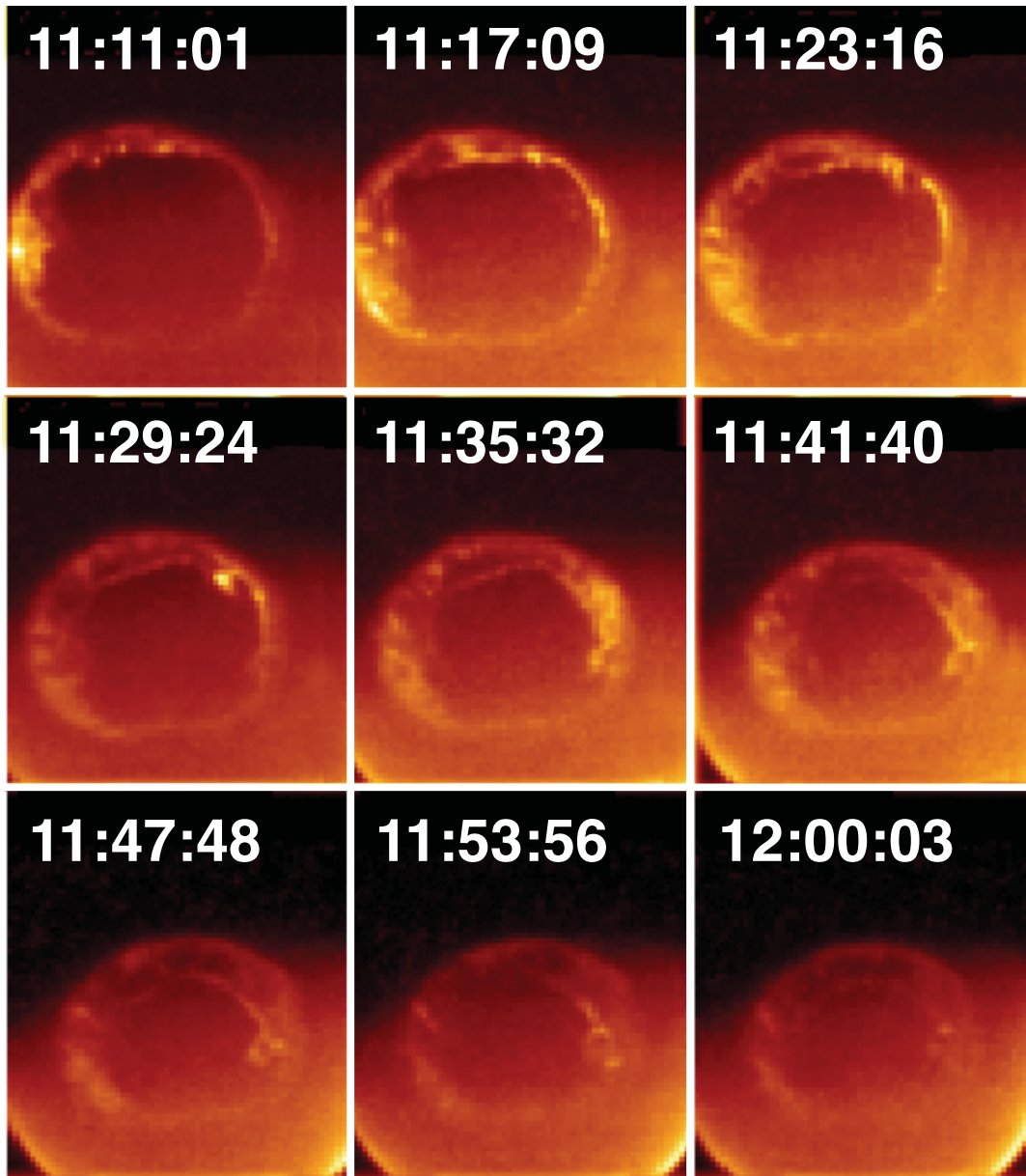


Figure 33.7: The aurora borealis as seen from above, looking down on the Earth. These images were taken by the IMAGE spacecraft's Far Ultraviolet Imaging System. (*Credit: NASA.*)

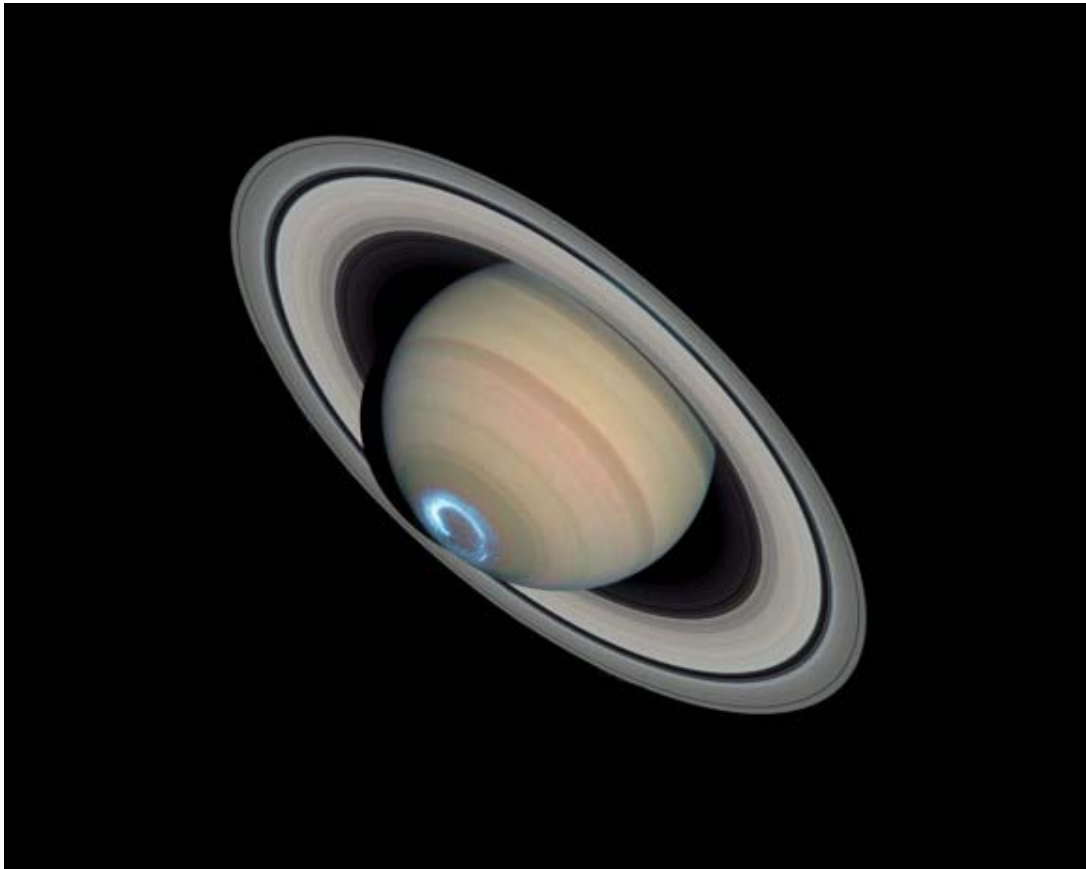


Figure 33.8: Auroral oval at Saturn's pole, taken in ultraviolet light by the Hubble Space Telescope. (*Credit: NASA.*)

Chapter 34

Magnetic Materials

Generally every material responds to a magnetic field in one way or another: materials may be weakly repelled by a magnet (*diamagnetism*), weakly attracted to a magnet (*paramagnetism*), or strongly attracted to a magnet (*ferromagnetism*). Each of these phenomena is described below.

34.1 Diamagnetism

In a *diamagnetic* material, the atoms of the material have no magnetic dipole moment. The external magnetic field alters the speed of electrons in their orbits around the atomic nucleus, which induces an internal magnetic field that repels the external field.

34.2 Paramagnetism

In a *paramagnetic* material, the atoms of the material do have a magnetic dipole moment. The dipole moments of the atoms align themselves with the external magnetic field to create an internal field that is weakly attracted to the external magnetic field.

34.3 Ferromagnetism

Ferromagnetic materials are strongly attracted to magnets, and can be made into permanent magnets. The ferromagnetic elements are iron, cobalt, and nickel, along with the rare earth elements gadolinium and dysprosium.

In ferromagnetic materials, the material is divided into a number of *magnetic domains*, each of which has dimensions on the order of 1 mm or so. Within each domain, the atomic dipole moments are aligned in the same direction. In an unmagnetized ferromagnetic material, each domain has its magnetic moment oriented in a different (random) direction, so that the dipole moments of the material as a whole tend to cancel out. But if a ferromagnetic material is exposed to an external magnetic field, the domains will tend to align themselves with the external field, so that the material as a whole takes on a net dipole moment.

Unlike diamagnetic and paramagnetic materials, ferromagnetic materials respond nonlinearly when placed in an external magnetic field. When a ferromagnetic material is placed within an external magnetic field, and the external field is then removed, the field in the ferromagnetic material will *not* disappear; a remanent magnetic field will remain, turning the material into a permanent magnet. The nonlinear response of a ferromagnetic material to an external magnetic field is called *hysteresis* (*hiss-tuh-REE-sus*) (Fig. 34.1).

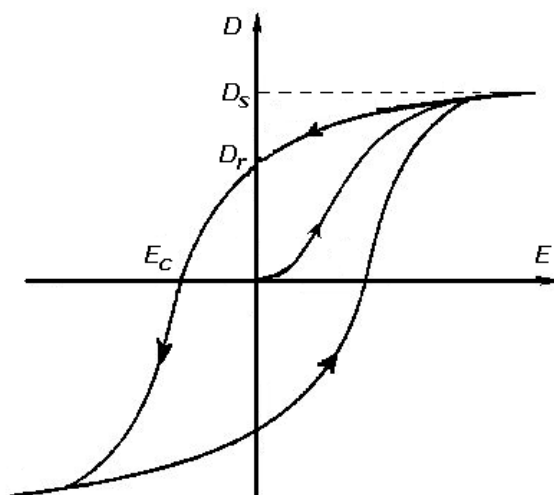


Figure 34.1: Hysteresis in a ferromagnetic material. The vertical axis (labeled D) is the internal magnetic field induced in the material, and the horizontal axis (labeled E) is the external magnetic field. Starting at the origin (an unmagnetized material in no magnetic field) and following the path with E increasing, we find the induced field in the material increases until it reaches a saturation level, labeled D_s . When the external field E is removed, though, the path does not return to the origin; instead a remanent field (D_r , the *remanence*) remains in the material. The point labeled E_c is called the *coercivity* of the material, and is the external field needed to de-magnetize the material. (Credit: Wikipedia, ©GNU-FDL, Wikimedia Commons.)

34.4 Permanent Magnets

Permanent magnets are ferromagnetic materials that have a permanent magnetic field. They are manufactured in a variety of materials; one of the most common is *alnico*, which is an alloy of aluminum, nickel, and cobalt (hence Al-Ni-Co, or “alnico”), and has a shiny metallic appearance like steel. Many horseshoe and bar magnets are made of this material, as well as heavy-duty handle magnets.

Ferrite or *ceramic* magnets are made of a brownish, brittle ceramic material mixed with ferric oxide (Fe_2O_3). They are sometimes used as components in electronic circuits.

Rare-earth magnets are made of alloys that include rare-earth elements (basically the “lanthanides” row of the periodic table). Two kinds of rare-earth magnets are made: *samarium-cobalt* (often used in stereo headphones and speakers), and *neodymium*. Neodymium magnets are made of an alloy of neodymium with iron and boron ($\text{Nd}_2\text{Fe}_{14}\text{B}$), and are the most powerful permanent magnets made. Even very small neodymium magnets are surprisingly powerful, and must be handled with care: two such magnets will attract each other with a very strong force, and can easily shatter. Once stuck together, two neodymium magnets can be very difficult to separate.

34.5 Curie Temperature

Once a ferromagnetic material has been magnetized by exposure to an external magnetic field, it may be de-magnetized by heating it above a temperature called the *Curie temperature*. Above this temperature, the thermal motion of the atoms is sufficient to re-scramble the magnetic dipole moments of the magnetic domains, and the material becomes de-magnetized. The Curie temperatures of ferromagnetic elements are shown in Table 34-1.

Table 34-1. Curie temperatures for ferromagnetic elements.

Element	Curie temperature (°C)
Iron	770
Cobalt	1115
Nickel	354
Gadolinium	20
Dysprosium	−188

34.6 Eddy Currents

A metal like aluminum is not ferromagnetic, so a sample of it cannot be picked up with a magnet the way iron can. However, it can still be influenced by a magnetic field.

Suppose we put a piece of aluminum at the end of a light rod, so that it forms the bob of a pendulum. If this pendulum is allowed to swing back and forth in the presence of an external magnetic field, something surprising happens: the motion will be strongly damped and the pendulum will quickly stop swinging.

What's happening is that as the aluminum metal moves through the magnetic field, electric currents called *eddy currents* are induced in the aluminum; those electric currents in turn produce a magnetic field of their own, in a direction that opposes the external magnetic field. The interaction of the external and induced magnetic fields produces the observed damping motion.

Chapter 35

Ampère's Law

In Chapter 31 we introduced the Biot-Savart law, which gives the magnetic field produced by a short current element, and allows us to find the magnetic field due to any arbitrary geometry of electric current. Another equation that gives the magnetic field produced by an electric current is *Ampère's law*, named, like the SI unit of electric current, for the French physicist André-Marie Ampère. Given an electric current I , imagine drawing a closed curve C around the current, so that the current passes through a surface bounded by C . Now divide the curve C into small segments Δl , and at each segment, measure the component magnetic field that is parallel to Δl ; we'll call that magnetic field $B_{\parallel}$. Then Ampère's law states that

$$\sum B_{\parallel} \Delta l = \mu_0 I. \quad (35.1)$$

In other words, when we add together the products $B_{\parallel} \Delta l$ for all the segments Δl that make up curve C , we get μ_0 times the current passing through the surface bounded by C .

So what? The Biot-Savart law tells us the magnetic field produced by an arbitrary arrangement of electric current; why do we need another law that tells us the same thing? Recall Gauss's law from Chapter 16: it allows us to compute the electric field due to an arbitrary distribution of charge, although we could do the same thing with Coulomb's law. The difference is that Gauss's law allows us to compute the electric field for *symmetrical* charge distributions very easily—much more easily than using Coulomb's law. In these cases, Gauss's law can save a great deal of work. But if we have an irregular distribution of charge, we may have no choice but to rely on Coulomb's law and compute the electric field “the hard way.”

The relationship between the Biot-Savart law and Ampère's law is similar. Although the Biot-Savart law will always work, it can be difficult to use. In some cases where the distribution of current is highly symmetrical, Ampère's law gives us a shortcut for finding the magnetic field that is much less work than using the Biot-Savart law. For irregular arrangements of electric current, though, we may have no choice but to “do it the hard way” and resort to the Biot-Savart law.

For example, let's find an expression for the magnetic field due to a current I in an infinitely long, straight wire, at a perpendicular distance r from the wire. To use Ampère's law, we imagine drawing a circle of radius r around the wire, so that the plane of the circle is perpendicular to the wire and the wire passes through the center of the circle. We already know that the magnetic field due to the wire is in the shape of concentric circles around the wire, so when we divide the circle into a number of small segments Δl , we know the magnetic field B will already be parallel to Δl for each segment. Therefore for an infinitely long, straight wire,

$$\sum B_{\parallel} \Delta l = B \sum \Delta l = 2\pi r B. \quad (35.2)$$

Then by Ampère's law,

$$2\pi r B = \mu_0 I, \quad (35.3)$$

or

$$B = \frac{\mu_0 I}{2\pi r}, \quad (35.4)$$

in agreement with Eq. (31.8). We could have arrived at the same result using the Biot-Savart law, but it would be much more work.

Chapter 36

Faraday's Law

The fundamental equations of classical electricity and magnetism are four equations called *Maxwell's equations*. The four equations are:

1. *Gauss's law* (Chapter 16) describes the electric field created by electric charges.
2. *Gauss's law for magnetism* (Chapter 31) states that there are no magnetic monopoles.
3. *Ampère's law* (Chapter 35) describes how a time-varying electric field creates a magnetic field.
4. *Faraday's law* describes how a time-varying magnetic field creates an electric field.

In this chapter we'll look at the fourth of these equations, Faraday's law. Mathematically, Faraday's law states that if the magnetic flux Φ_B through a closed loop of wire changes with time, then there will be an electromotive force $\mathcal{E}$ (i.e. a voltage) induced in the wire given by

$$\mathcal{E} = -N \frac{\Delta\Phi_B}{\Delta t} \quad (36.1)$$

Here $\mathcal{E}$ is the induced electromotive force, N is the number of turns of wire in the loop, and $\Delta\Phi_B$ is the change in magnetic flux in time Δt . (Recall that the magnetic flux is given by $\Phi_B = \mathbf{B} \cdot \mathbf{A}$.)

The magnetic flux through the loop(s) of wire can be changed in several ways: the magnetic field B can change in magnitude with time; the magnetic field can change direction with time; the loop of wire can change its orientation with time; the area of the loop can change with time; or some combination of these.

Faraday's law forms the basis of the *electric generator*, which is responsible for producing most of the electricity we use every day (except for electricity produced by batteries or solar arrays). Loops of wires are turned inside a stationary magnetic field (or magnets may be turned inside stationary wires); this causes the magnetic flux through the wires to change with time, creating an electric current. The turning motion may be created by a water wheel, by geothermal steam, by steam created from burning coal or oil, or by steam created by heat from a nuclear reaction. In effect, an electric generator converts mechanical motion into electrical energy.

Faraday's law may also be used in reverse: electrical energy may be converted into mechanical motion. This creates an *electric motor*, a device that is familiar in such household appliances as vacuum cleaners and electric dryers.

36.1 Lenz's Law

Faraday's law gives the magnitude of the induced electromagnetic force (emf) created by a changing magnetic field. The *direction* of the induced electromotive force and current may be found from a statement known as

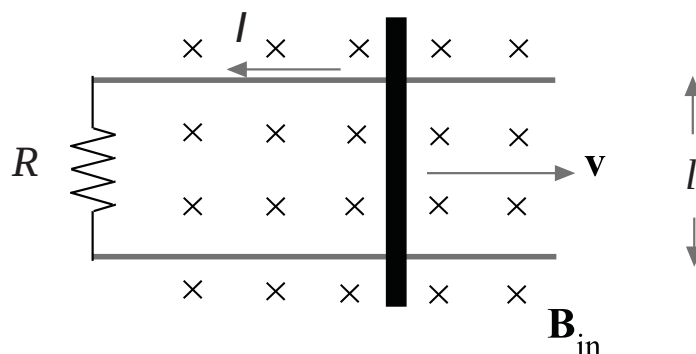


Figure 36.1: Motional emf.

Lenz's law (named for the 19th century Russian physicist Heinrich Lenz):

The emf and induced current are in such a direction as to tend to oppose the change which produced them.

36.2 Motional EMF

As an example to illustrate both Faraday's law and Lenz's law, consider the situation shown in Figure 36.1. Two parallel conducting rails separated by a distance l are connected on their left end by a resistor. A conducting bar is placed across the rails, and the entire apparatus is placed in a uniform magnetic field B pointing into the page. Now move the conducting rail to the right with velocity v ; this will increase the area enclosed by the circuit, which will increase the magnetic flux inside the circuit. By Faraday's law, this will induce an electromotive force (voltage) in the circuit. An emf induced in this way is called *motional emf*.

We can find the magnitude of the induced emf using Faraday's law. At a given instant, there is an area A enclosed by the circuit, formed by the rails, resistor, and conducting bar. When the bar is moving at velocity v to the right, then in a time interval Δt the area increases by an amount $lv\Delta t$. The rate at which the area changes is then

$$\frac{\Delta A}{\Delta t} = \frac{lv\Delta t}{\Delta t} = lv. \quad (36.2)$$

Since the magnetic flux $\Phi_B = BA$, we have

$$\frac{\Delta \Phi_B}{\Delta t} = B \frac{\Delta A}{\Delta t} = Blv. \quad (36.3)$$

Therefore by Faraday's law, the magnitude of the induced electromotive force is

$$|\mathcal{E}| = \frac{\Delta \Phi_B}{\Delta t} = Blv. \quad (36.4)$$

We can deduce the *direction* of the induced current by using Lenz's law, which says that the induced current must be in such a direction that the magnetic field it produces will tend to oppose the change in magnetic flux. If the conducting bar moves to the right, then the magnetic flux Φ_B is *increasing* with time. Therefore the induced current must be *counterclockwise*, because, by the right-hand rule, a counterclockwise

current will produce a magnetic field inside that circuit that points *out of* the page, which will tend to decrease the magnetic flux. In other words, the magnetic flux “wants” to remain relatively constant; if the moving bar increases the magnetic flux through the circuit, then the induced current will be in a direction to decrease it, so that it tries to stay as constant as possible.

Another way to determine the direction of the induced current is via the Lorentz force. The conducting bar is full of free (negatively charged) electrons. As the bar moves across the magnetic field, the Lorentz force on each electron will be $\mathbf{F} = -e\mathbf{v} \times \mathbf{B}$; since $\mathbf{v}$ is to the right and $\mathbf{B}$ is into the page, this means $\mathbf{F}$ will be *downward*, so the electrons will move downward. The conventional current moves opposite the direction of the electrons, so the current in the bar is *upward*, and the current in the circuit is therefore counterclockwise.

Chapter 37

Inductance

The space surrounding an electric circuit contains a magnetic field due to the electric currents running in the circuit. By the Biot-Savart law, the magnetic flux Φ_B is proportional to the current I :

$$\Phi_B = LI. \quad (37.1)$$

Here the proportionality constant L is called the *inductance*. Of course, the magnetic flux at a given point in space depends on a number of factors besides the current: it also depends on the distance from the current, the permeability of free space, etc. The inductance L can be thought of as all these other factors lumped together.

The SI unit of inductance is the *henry* (H), named for the American physicist Joseph Henry. Since magnetic flux is measured in webers and current in amperes, Eq. (37.1) indicates that one henry is equal to one weber per ampere: $1 \text{ H} = 1 \text{ Wb/A}$.

37.1 Solenoid Inductor

One very common device for introducing inductance into an electric circuit is the *solenoid*, which is a coil of wire wrapped on an insulating cylinder. As discussed earlier in Section 31.9, the magnetic field B inside a solenoid carrying a current I is given by

$$B = \mu_0 n I, \quad (37.2)$$

where $n = N/\ell$ is the total number of turns of wire N divided by the length ℓ of the solenoid. We can find an expression for the inductance L of a solenoid by starting with this equation for B . Let A be the cross-sectional area of the solenoid, and let N be the total number of turns of wire. Then the magnetic field passes through N turns, each of which has area A . The total area through which the magnetic field passes is therefore NA , and so the magnetic flux is

$$\Phi_B = B(NA) \quad (37.3)$$

$$= \mu_0 N n I A \quad (37.4)$$

$$= \mu_0 n^2 I A \ell. \quad (37.5)$$

The inductance L is then found to be

$$L = \frac{\Phi_B}{I}, \quad (37.6)$$

so the inductance of a solenoid is

$$L = \mu_0 n^2 A \ell = \mu_0 N^2 \frac{A}{\ell}. \quad (37.7)$$

Note that the inductance, like the capacitance, depends only on factors involving the geometry of the inductor: its length ℓ , cross-sectional area A , and number of turns of wire N .

37.2 Inductors in Series and Parallel

Inductors connected in series and parallel follow the same equations as resistors.

Several inductors connected end-to-end (*in series*) have an equivalent inductance equal to the sum of the individual inductances:

$$L_s = \sum_i L_i \quad (37.8)$$

$$= L_1 + L_2 + L_3 + \cdots \quad (37.9)$$

If they are connected *in parallel*, the the equivalent inductance is the reciprocal of the sum of the reciprocals of the individual inductances:

$$\frac{1}{L_p} = \sum_i \frac{1}{L_i} \quad (37.10)$$

$$= \frac{1}{L_1} + \frac{1}{L_2} + \frac{1}{L_3} + \cdots \quad (37.11)$$

Note the following points. For inductors connected *in series*:

- The equivalent inductance will be bigger than the largest inductance in the series combination.
- If one inductor in the series combination is much larger than the others, the equivalent inductance will be approximately equal to the largest inductance.
- M equal inductors L connected in series have an equivalent inductance of ML .

For inductors connected *in parallel*:

- The equivalent inductance will be smaller than the smallest inductance in the parallel combination.
- If one inductor in the parallel combination is much smaller than the others, the equivalent inductance will be approximately equal to the smallest inductance.
- M equal inductors L connected in parallel have an equivalent inductance of L/M .

37.3 Magnetic Materials in Inductors

As shown by Eq. (37.7), the inductance of a solenoid can be increased by increasing the cross-sectional area of the plates, or by increasing the number of turns of wire. Another way to increase the inductance is to insert a magnetic material inside the solenoid; this will cause the inductance to increase by a factor of K_m :

$$L = K_m \mu_0 N^2 \frac{A}{\ell}, \quad (37.12)$$

where K_m is called the *relative permeability* of the material. The combination

$$\mu = K_m \mu_0 \quad (37.13)$$

is called the *permeability* of the material. Since the relative permeability K_m is typically a number very close to 1, it is convenient to introduce the *magnetic susceptibility* χ_m , defined by

$$K_m = 1 + \chi_m. \quad (37.14)$$

37.4 Energy Stored in an Inductor

An inductor can be thought of as a device that stores energy in the magnetic field between inside the coils of the inductor. Using the calculus, it can be shown that the potential energy U stored in the magnetic field of an inductor of inductance L carrying current I , is given by

$$U = \frac{1}{2} L I^2. \quad (37.15)$$

The *energy density* (energy per unit volume) of an inductor can be found by using the solenoid as an example. From Eq. (37.7), the total potential energy stored in a solenoid (of plate area A and separation d) is

$$U = \frac{1}{2} L I^2 \quad (37.16)$$

$$= \frac{1}{2} \mu_0 n^2 A \ell I^2 \quad (37.17)$$

Since the magnetic field inside the solenoid is $B = \mu_0 n I$, this gives

$$U = \frac{1}{2} (\mu_0 n I)^2 (A \ell) \frac{1}{\mu_0} \quad (37.18)$$

$$= \frac{1}{2} \frac{1}{\mu_0} B^2 A \ell \quad (37.19)$$

Since the volume inside the solenoid is $A \ell$, the energy density $u = U/(A \ell)$, or

$$u = \frac{1}{2 \mu_0} B^2. \quad (37.20)$$

Compare this result with the analogous equation for a capacitor, Eq. (25.16):

$$u = \frac{1}{2} \epsilon_0 E^2. \quad (37.21)$$

Chapter 38

LR Circuits

By connecting an inductor and a resistor together in series, we create an *LR circuit*. In an LR circuit, energy is stored in the magnetic field of the inductor, and the resistor controls the rate at which current reaches the inductor. The characteristic time scale required to create a full-strength magnetic field in the inductor is called the *time constant* τ , and is given by

$$\tau = \frac{L}{R}. \quad (38.1)$$

If the inductance L is in henries and the resistance R is in ohms, then the time constant τ will have units of seconds.

Figure 38.1 shows an LR circuit. The circuit includes a battery, so that when the switch S is closed, current flows through the resistor and inductor, and begins building up a magnetic field inside the coils of the inductor. The resulting magnetic field will be in a direction that, by Lenz's law, will tend to oppose changes in the direction of the current, so that it becomes harder to increase the current. Once an amount of time has gone by that is large compared to the time constant $\tau = L/R$, the magnetic field in the inductor will have essentially reached its maximum value, and the current will be constant.

Figure 38.2 shows the resistor voltage, inductor voltage, circuit current, and inductor magnetic flux in the LR circuit as a function of time. The switch S is closed at time $t = 0$. Shortly afterwards, a small current flows through the circuit, the voltage across the resistor R is equal zero, and the voltage across the inductor is equal to the battery voltage V . At time $\tau = L/R$ after the switch is closed, the voltage across the resistor

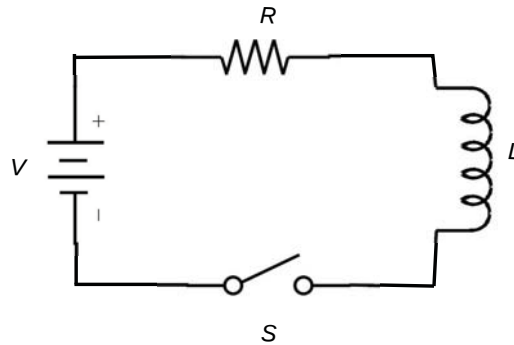


Figure 38.1: An LR circuit.

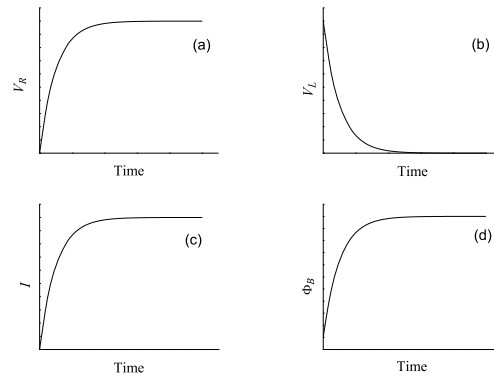


Figure 38.2: Plots vs. time for an LR circuit. (a) Resistor voltage vs. time; (b) inductor voltage vs. time; (c) circuit current vs. time; and (d) magnetic flux in the inductor vs. time. The switch S is closed at time $t = 0$.

has increased to $1 - 1/e = 0.632$ of the battery voltage; the voltage across the inductor has decreased to $1/e = 0.368$ of the battery voltage; the current has increased to $1 - 1/e$ of its maximum value; and the magnetic flux in the inductor has increased to $1 - 1/e$ of its maximum value.

Mathematically, the voltage across the resistor V_R , the voltage across the capacitor V_C , the current in the circuit I , and the magnetic flux Φ_B in the inductor can be shown to be

$$V_R(t) = V(1 - e^{-t/\tau}) \quad (38.2)$$

$$V_L(t) = Ve^{-t/\tau} \quad (38.3)$$

$$I(t) = (V/R)(1 - e^{-t/\tau}) \quad (38.4)$$

$$\Phi_B(t) = (LV/R)(1 - e^{-t/\tau}) \quad (38.5)$$

As time $t \rightarrow \infty$, current will reach a maximum value $I = V/R$, the magnetic flux in the inductor will have reached its maximum value LV/R , the voltage across the resistor will equal the battery voltage, and the voltage across the inductor will be zero.

Chapter 39

LC and LCR Circuits

We've previously looked at the RC circuit (a resistor and capacitor connected to a battery) and the RL circuit (a resistor and inductor connected to a battery). The two circuits have complementary behavior: in the RC circuit, the current starts out with a maximum value at the instant the switch is closed and decreases exponentially toward zero. In the RL circuit, the current starts out small the instant the switch is closed, increases with time, and eventually levels off to its maximum value.

39.1 LC Circuits

By connecting a charged capacitor and an inductor together, we create something called an *LC circuit* (Figure 39.1). In an LC circuit, the complementary behavior of the capacitor and the inductor give some interesting results. As the capacitor discharges, a current is created in the circuit, which starts to build a magnetic field in the inductor. As time goes on, the current will increase, but start to level off because the inhibiting effect of the inductor: the inductor will create magnetic field an induced current in a direction that will oppose the increase in the magnetic field in the inductor. By the time the capacitor has fully discharged, the magnetic field in the inductor will have reached its maximum value.

At this point the current would stop, were it not for the presence of the magnetic field in the inductor. Once the capacitor has fully discharged, it can no longer provide current to the inductor, and the magnetic field in the inductor begins to collapse. But this change in the magnetic field induces a current in a direction that opposes the collapse in the magnetic field—in other words, in a direction that will continue the current in its original direction, so that the capacitor will begin to charge with the opposite polarity that it originally had. By the time the magnetic field in the inductor has completely collapsed, the capacitor will be fully re-charged (with opposite its original polarity), and the process begins again in reverse. Current will now begin to flow in the opposite direction, creating a magnetic field in the inductor whose polarity is opposite what it was before. The process will continue as before (but in the opposite direction) until the capacitor is fully charged with its original polarity, and the cycle begins again, repeating over and over.

The result is an electrical form of simple harmonic motion, with energy moving back and forth between the electric field stored in the capacitor and the magnetic field stored in the inductor. It can be shown that this oscillation has angular frequency

$$\omega = \frac{1}{\sqrt{LC}}, \quad (39.1)$$

where L is the inductance of the inductor and C is the capacitance of the capacitor. The period of oscillation is therefore $T = 2\pi/\omega$, or

$$T = 2\pi \sqrt{LC}. \quad (39.2)$$

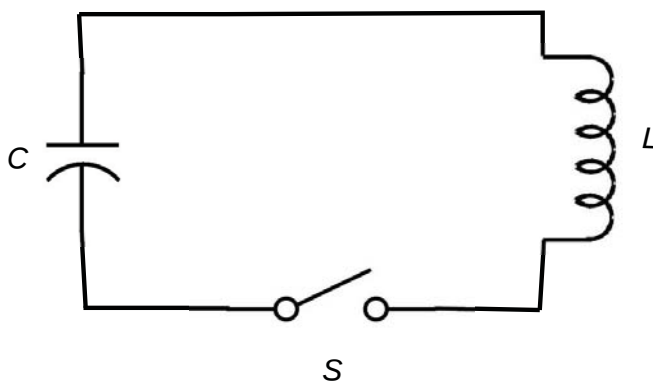


Figure 39.1: An LC circuit.

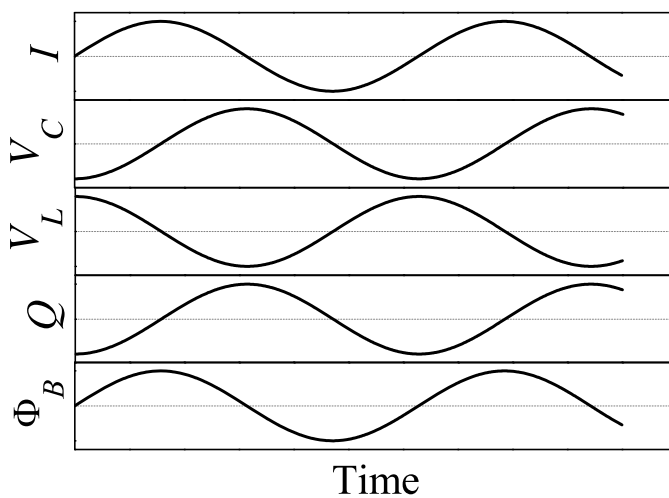


Figure 39.2: Plots vs. time of LC circuit current I , capacitor voltage V_C , inductor voltage V_L , capacitor charge Q , and inductor magnetic flux Φ_B . The current and inductor flux are in phase with each other, as are the voltage and charge on the capacitor. The voltages on the capacitor and inductor are 180° out of phase.

Figure 39.2 shows plots vs. time of the circuit current, voltages across the inductor and capacitor, charge on the capacitor, and magnetic flux in the inductor. For these plots, at $t = 0$ (the instant the switch is closed), the capacitor in Fig. 39.1 is initially fully charged with the top plate positive. Positive current is taken to be clockwise. All quantities vary sinusoidally with the same period, but may be shifted in phase with respect to each other. If the initial charge on the capacitor is Q_0 , then the amplitudes for each quantity will be as shown in the following table:

Quantity	Symbol	Amplitude
Current	I	ωQ_0
Capacitor voltage	V_C	Q_0/C
Inductor voltage	V_L	$L\omega^2 Q_0$
Capacitor charge	Q	Q_0
Inductor mag. flux	Φ_B	$L\omega Q_0$

Energy of an LC Circuit

In a simple harmonic oscillator formed by a mass on a spring, energy is continuously sloshing back and forth between kinetic and potential energy, with the sum of the two (the total energy) being constant. Similarly, in an LC circuit, energy is continuously sloshing back and forth between electric energy in the capacitor and magnetic energy in the inductor. The electric energy in the capacitor is given by Eq. (25.12):

$$U_e = \frac{1}{2} \frac{Q^2}{C}. \quad (39.3)$$

From Figure 39.2 and the above table, we have the charge on the capacitor (lower plate) as a function of time is given by

$$Q(t) = -Q_0 \cos \omega t. \quad (39.4)$$

and so the electric energy U_e as a function of time is

$$U_e(t) = \frac{Q_0^2}{2C} \cos^2 \omega t. \quad (39.5)$$

Similarly, the magnetic energy in the inductor is given by Eq. (37.15):

$$U_m = \frac{1}{2} L I^2, \quad (39.6)$$

where the current at time t is

$$I(t) = \omega Q_0 \sin \omega t. \quad (39.7)$$

Substituting this expression for $I(t)$ into the formula for U_m gives an expression for the magnetic energy of the inductor as a function of time:

$$U_m(t) = \frac{L\omega^2 Q_0^2}{2} \sin^2 \omega t \quad (39.8)$$

$$= \frac{Q_0^2}{2C} \sin^2 \omega t, \quad (39.9)$$

where we have used the fact that $\omega^2 = 1/LC$. The total energy U is then

$$U = U_e + U_m \quad (39.10)$$

$$= \frac{Q_0^2}{2C} \cos^2 \omega t + \frac{Q_0^2}{2C} \sin^2 \omega t = \frac{Q_0^2}{2C}, \quad (39.11)$$

which is a constant, as expected.

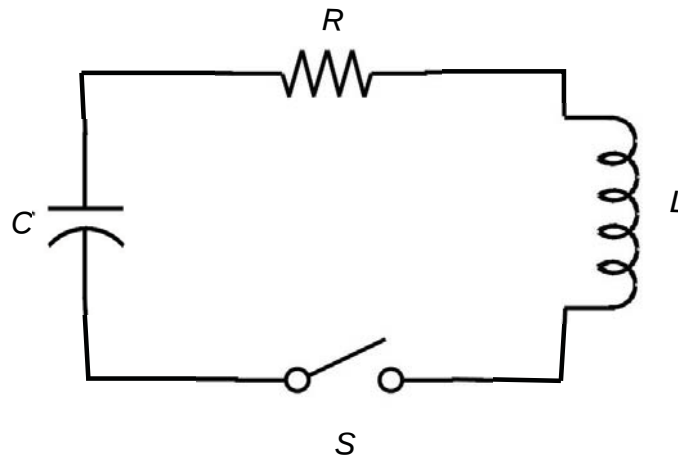


Figure 39.3: An LCR circuit.

39.2 LCR Circuits

If a resistor is placed in an LC circuit, we have an *LCR circuit* (Fig. 39.3). The effect of the resistor is to introduce damping into the oscillation of the circuit: while an LC circuit oscillates like a simple harmonic oscillator, an LCR circuit behaves like a *damped* harmonic oscillator. Depending on the value of the resistance R , the current in the circuit may be underdamped, overdamped, or critically damped, just as with the damped harmonic oscillator.

Chapter 40

AC Circuits

All of the electric circuits we have seen so far are *direct current* (DC) circuits. This means that the current is always traveling in the same direction at each point in the circuit. This is in contrast to *alternating current* (AC) circuit, in which the direction of the current alternates back and forth between one direction and another. Batteries provide direct current; the electric outlets in the walls of your house provide alternating current.

Capacitors and inductors are electrical components that are more typically seen in AC circuits than in DC circuits. An AC circuit containing resistors, capacitors, and inductors can be analyzed using an interesting mathematical trick: we simply treat all three components as if they were *complex-valued resistors*, then use all the methods we used earlier to analyze DC circuits with resistors (but with complex arithmetic). This complex-valued resistance is called *impedance*, and is given the symbol Z . Impedance has the same units as resistance, ohms (Ω).

The value of the complex impedance for a resistor, capacitor, and inductor is shown in the table below. In the table, the symbol j stands for the imaginary unit,¹ i.e. $j = \sqrt{-1}$. The variable f in the table refers to the frequency of the voltage source attached to the component.

Component	Impedance
Resistance R	$Z_R = R$
Capacitance C	$Z_C = \frac{1}{j2\pi fC}$
Inductance L	$Z_L = j2\pi fL$

We interpret the final results of the analysis (complex numbers) as giving information about both the amplitude and phase of the signal at any point in the circuit.

¹In most areas of science, mathematics, and engineering, the symbol i is used for $\sqrt{-1}$. But in electrical engineering, i is used for current; so to avoid confusion, electrical engineers write j instead of i for $\sqrt{-1}$.

Example. Suppose a $15\ \Omega$ resistor, a $300\ \mu\text{F}$ capacitor, and a $4\ \text{mH}$ inductor are connected in series to a sinusoidal voltage source of frequency $60\ \text{Hz}$. Then the equivalent *impedance* of the series combination is

$$\begin{aligned} Z &= Z_R + Z_C + Z_L \\ &= R + \frac{1}{j2\pi fC} + j2\pi fL \\ &= (15\ \Omega) + \frac{1}{j2\pi(60\ \text{Hz})(300 \times 10^{-6}\ \text{F})} + j2\pi(60\ \text{Hz})(4 \times 10^{-3}\ \text{H}) \\ &= (15 - 7.3340j)\ \Omega \end{aligned}$$

where we have used the identity $1/j = -j$.

Chapter 41

Memristance

We have seen the three basic components of analog electronics are the resistor, capacitor, and inductor. Let's arrange the defining equations for resistance R , capacitance C , and inductance L into a 2×2 table:

$R = V/I$	$L = \Phi_B/I$
$C^{-1} = V/Q$	

(We'll use the reciprocal of capacitance (elastance) to make the pattern clear.) Notice the pattern: in the first row the current is in the denominator, and in the first column the voltage is in the numerator. You might guess that there could be another combination, Φ_B/Q , to fill in the lower-right corner. This idea led American electrical engineer Leon Chua to predict the existence of a *fourth* analog electronic component in 1971, the *memristor*.¹ The *memristance* $M = \Phi_B/Q$ completes the table:

$R = V/I$	$L = \Phi_B/I$
$C^{-1} = V/Q$	$M = \Phi_B/Q$

Memristance has the same units as resistance, ohms (Ω).

The memristor was finally discovered during experiments with molecular electronics at the Hewlett-Packard laboratories in 2008. It behaves like a resistor with a "memory" (hence the name): when voltage is removed from a memristor, it still "remembers" how much voltage was last applied to it, and for how long. The resistance increases when the current flows through it in one direction, decreases when current flows in the opposite direction, and remains unchanged when no current flows through it.

Practical applications are still being discussed, but possibilities include applications to non-volatile computer memory, including computers that could remember their previous state when being powered on, thus avoiding the usual lengthy boot-up process.

¹See *IEEE Spectrum*, <http://spectrum.ieee.org/semiconductors/design/the-mysterious-memristor> (May 2008).

Chapter 42

Electromagnetism

42.1 Electromagnetic Waves

As mentioned in Chapter 36, the theory of classical electricity and magnetism is based on four equations called *Maxwell's equations*, named for the 19th-century Scottish physicist James Clerk Maxwell. Around the time of the American Civil War (1865) Maxwell collected the four equations together, and realized that there was a crucial term called the *displacement current* missing from Ampère's law that was required to make it self-consistent. After adding this term to Ampère's law, Maxwell was able to show that the four equations could be combined to derive a *wave equation*, which describes a wave moving with speed

$$\frac{1}{\sqrt{\epsilon_0 \mu_0}} = c = 299,792,458 \text{ m/s}, \quad (42.1)$$

which is the speed of light in vacuum. This is a remarkable result: by combining equations that summarized the results of laboratory experiments on electric and magnetic fields, Maxwell was able to demonstrate that light is an *electromagnetic wave*, thus connecting the fields of electromagnetism and optics.

Specifically, the classical view of electromagnetic waves (including visible light) is that it consists of a transverse electric wave; the electric wave in turn creates a perpendicular magnetic wave, which in turn produces the electric wave, and so on. In other words, light (and other electromagnetic waves) consist of electric and magnetic waves that sustain each other as they propagate through space, so that no material medium is required. (Fig. 42.1.) Light can propagate in a vacuum.

Visible light is just one of many forms of electromagnetic wave. Electromagnetic waves are categorized (somewhat arbitrarily) according to their wavelength, as shown in Table 42-1. It's important to realize, though, that all these waves are really the same thing: they differ only in their wavelength, and the different names we give them are for our own convenience.

Table 42-1. Electromagnetic waves and their wavelengths. Wavelength increases going down the table from top to bottom; frequency and energy increase going up the table from bottom to top.

Wave	Wavelengths
Gamma rays	< 0.1 nm (shortest λ ; highest f, E)
X-rays	0.1 – 10 nm
Ultraviolet	10 – 400 nm
Visible	400 – 700 nm
Infrared	0.7 – 100 μm
Submillimeter	0.1 – 1 mm
Microwaves	1 mm – 1 m
Radio	> 1 m (longest λ ; lowest f, E)

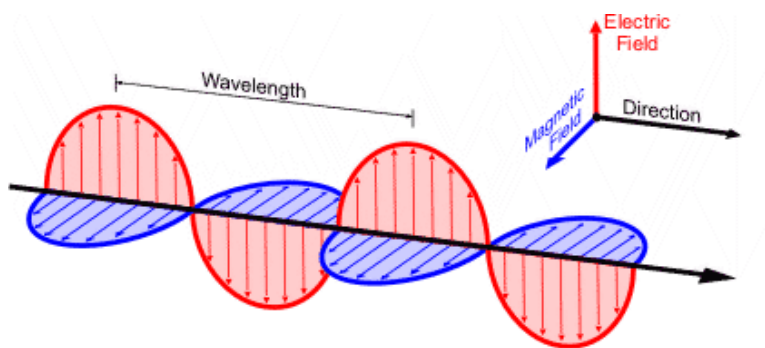


Figure 42.1: Diagram of an electromagnetic wave. The electric and magnetic vectors are perpendicular, and they peak together and go to zero together. The wave travels in the $\mathbf{E} \times \mathbf{B}$ direction (to the right in this figure). (Credit: NOAA.)

Gamma rays are the shortest-wavelength, highest-frequency, highest-energy waves. They are generally associated with nuclear processes (such as nuclear fission and fusion), and with other high-energy reactions such as matter-antimatter annihilation.

X-rays are familiar for their medical uses. Human tissue is transparent in X-rays, but human bone is opaque. By viewing an image of the human body in X-rays, one may create images of the human skeleton. X-rays are generally less energetic than gamma rays, and are generally produced by atomic reactions.

Ultraviolet light is light whose wavelength is shorter than can be seen by the human eye (although some animals can see in ultraviolet light). The Sun emits a significant amount of ultraviolet light, which can cause sun tans and sunburns in humans.

Visible light is light whose wavelengths are visible to the human eye. Violet light is the shortest wavelength (about 400 nm) and highest frequency and energy; red light is the longest wavelength (about 700 nm) and lowest frequency and energy. The order of the colors of visible light (from longest to shortest wavelength) is given by the mnemonic “ROY G. BIV”: Red, Orange, Yellow, Green, Blue, Indigo, Violet.

Infrared light is light whose wavelength is longer than can be seen by the human eye—although some animals like the pit viper can see in infrared light. Bodies that we consider “warm” (say, somewhat above room temperature) emit significant amounts of infrared light. For example, the human body can be seen to be “glowing” in infrared light, although it does not glow significantly in visible light. This is the principle of the night-vision scope, which is a device that converts infrared light to visible wavelengths, so that the user can detect this glowing of warm bodies.

Submillimeter waves are electromagnetic waves whose wavelength is between 0.1 to 1 mm. These waves are of some astronomical interest, and have a few applications in medicine.

Microwaves are electromagnetic waves whose wavelengths are typically measured in centimeters. Microwaves are familiar in their use in microwave ovens: the oven emits microwaves designed to resonate with the water molecules in food, thereby “shaking” the water molecules and heating the food. Microwaves also find uses in communications (high-frequency radio).

Radio waves are the longest-wavelength ($\lambda > 1$ m), lowest-frequency, lowest-energy electromagnetic waves. They are used in radio astronomy and in radio communications. These wavelengths include AM and FM radio and broadcast television. (FM radio lies between TV channels 6 and 7.)

Chapter 43

Radio

Electromagnetic waves can be *modulated* to carry information; this means that the pure sinusoidal electromagnetic wave is modified in some fashion to include information such as voice or music. There are two common methods for modulating a radio wave:

- *Amplitude modulation.* Here the frequency of the wave is constant, but the *amplitude* of the wave is modified to carry information. In a sense, the basic sinusoidal radio frequency (the *carrier wave*) is multiplied by the sound wave, and the superposition of the two is transmitted by a device called the *transmitter*. Another person has a device called a *receiver* that extracts the audio (sound) information and sends it to a speaker or headphones.
- *Frequency modulation.* Here the amplitude of the wave is kept constant, but the *frequency* is varied slightly about the carrier wave frequency in order to carry the audio information. Frequency modulation has the advantage of being less susceptible to noise from phenomena such as lightning discharges, but requires a more complex transmitter and receiver.

Radio is used in a number of ways:

- *AM Radio.* So named because it uses amplitude modulation, AM radio is a commercial service that is used to broadcast music, talk, news, sports, etc. It first appeared around 1920. Stations broadcast on frequencies between 520 kHz and 1700 kHz, separated by 10 kHz. During the day, AM stations may travel a few hundred miles, while at night they may travel across the continent by reflecting from the Earth's ionosphere.

Some AM stations broadcast at low power, or may broadcast only during the day. A few stations (Table 43-2) broadcast at the maximum allowed power (50 kW) day and night, and may be heard around the country at night.

- *FM Radio.* So named because it uses frequency modulation, FM radio is a commercial service whose content is similar to that of AM radio. Stations broadcast on frequencies between 87.9 MHz and 107.9 MHz, separated by 0.2 MHz. FM radio frequencies are in a gap between television channels 6 and 7. FM radio is less susceptible to “static” noise than AM, but the signals don't travel far—typically just a few dozen miles at most.
- *Shortwave.* Many countries broadcast an international service around the world in various languages. These stations broadcast on so-called “shortwave” frequencies between 1800 kHz and 30 MHz, and use amplitude modulation. At these frequencies, radio signals can bounce off of the Earth's ionosphere and travel around the world.

Shortwave broadcast content typically includes news and cultural features from that country, and often propaganda as well. Shortwave stations are not assigned a single fixed frequency the way our AM and FM stations are. Instead, they broadcast in blocks of an hour or so in length, with each block at a specific time and frequency, and in a specific language directed to a particular part of the world. And the broadcast schedules are often changed throughout the year. For this reason, a printed or on-line shortwave broadcast guide is helpful for finding times and frequencies of English-language broadcasts directed to North America.

Some of the best known shortwave stations are:

- *Voice of America* (United States)
- *Radio Canada International* * (Canada)
- *BBC World Service* (United Kingdom)
- *Radio Deutsche Welle* (Germany)
- *Radio Sputnik*, (formerly *Radio Moscow* and *Voice of Russia*; Russia)
- *Radio Australia* * (Australia)
- *China Radio International* (formerly *Radio Peking*; China)

Just very recently, a number of these stations (marked with an asterisk) have stopped broadcasting by radio, in favor of Internet service.

- *Television*. Television signals are sent in much the same way as radio signals, with information about the television picture being sent along with the audio. Television signals were encoded by frequency modulation until the switch to digital television in 2009. Television channels are broadcast in several contiguous “blocks” of frequencies, as shown in Table 43-1. Each television channel is 6 MHz wide. Originally, with analog television this was to allow room for both the video signal (lower part of the band) and audio signal (upper part of the band). Now with digital television, each channel requires less “bandwidth”, and so television stations often divide their 6 MHz channel into several sub-channels, each carrying different programming.

Table 36-1. Television channel frequencies.¹

Channels	Frequencies (MHz)
2 - 4	54 - 72
5 - 6	76 - 88
7 - 13	174 - 216
14 - 36	470 - 608
37 - 61	614 - 764
62 - 64	776 - 794

- *Cellular telephone*. Cellular telephone transmissions occur over a range of frequencies lying between 800 MHz and 2700 MHz. At these high frequencies, radio signals do not travel very far, so cellular telephone relies on a nation-wide system of “repeater” transmitters, which receive signals and re-transmit them until they reach their destination. (The range of each repeater is the “cell” of cellular telephone.)
- *Amateur radio*. Radio amateurs have access to a number of blocks of frequency all over the radio frequency spectrum. They use these for informal hobby chatting (called *ragchewing*), emergency communications, and experimenting with radio technology. Transmitting on the amateur radio bands requires an amateur radio license from the Federal Communications Commission.

¹There is no television channel 1. Channel 1 existed at one time, but was eliminated by the Federal Communications Commission in 1948 as part of negotiating competing interests in the radio frequency spectrum.

- *Citizen's Band radio.* This is a radio service available for use by anyone, with no license required. Citizen's Band (or "CB") radio enjoyed a huge (but brief) boom of popularity during the 1970s, when it was often used as a mobile radio service by drivers of cars and trucks. Today CB radio is much less popular, but is still used by truck drivers.
- *Other uses.* Various other radio services are available for use by police, fire, military, taxicabs, maritime and aircraft communications, etc.

Table 43-2. Some 50 kW clear-channel AM radio stations that can be heard from the east coast of the U.S.

f (kHz)	Call Sign	City
640	KFI	Los Angeles
650	WSM	Nashville
700	WLW	Cincinnati
710	WOR	New York
720	WGN	Chicago
750	WSB	Atlanta
760	WJR	Detroit
770	WABC	New York
780	WBBM	Chicago
830	WCCO	Minneapolis
840	WHAS	Louisville
850	KOA	Denver
870	WWL	New Orleans
880	WCBS	New York
890	WLS	Chicago
1020	KDKA	Pittsburgh
1030	WBZ	Boston
1040	WHO	Des Moines
1060	KYW	Philadelphia
1090	WBAL	Baltimore
1110	WBT	Charlotte
1120	KMOX	St. Louis
1160	KSL	Salt Lake City

43.1 The Ionosphere

The distance that radio waves can travel depends strongly on their frequency. At some frequencies, radio waves are able to reflect off of a layer of ionized gas in the Earth's atmosphere called the *ionosphere*. The ionosphere actually consists of three layers, called *D*, *E*, and *F*.² The lowest layer is the *D* layer, above that is the *E* layer, and the highest layer is *F*.³ During the day, sunlight ionizes these three layers, turning them into a plasma. At night, when the sunlight is gone, the ions in the *D* and *E* layers re-combine with the free electrons, and these layers become neutral. The *F* layer is high enough that the gas is at a very low density—low enough that the gas particles do not have time to collide and re-combine with the electrons, and the *F* layer remains ionized all night.

Radio waves interact with the ionosphere in different ways depending on their frequency. Shortwave radio frequencies, for example, are able to travel through the *D* and *E* layers to reach the *F* layer of the

²Layers of the ionosphere were lettered starting with *D* to allow *A*, *B*, and *C* to be used for possible other layers that might be discovered below the *D* layer. No such layers exist, though.

³The *F* layer splits into two layers (F_1 and F_2) during the day, and merges back into a single layer at night.

ionosphere throughout the day and night. After reaching the F layer, the signals reflect and bounce back to the ground, where they reflect again back toward the ionosphere, and so on. With multiple “hops,” like this, the radio waves can travel around the globe.

The interaction of AM radio waves is a bit different. During the night, AM radio waves can reach the F layer and travel like shortwave radio, making multiple hops across the globe. But during the day, AM radio waves are absorbed by the ionized D layer and cannot reach the F layer. In effect, the D layer acts as kind of a “curtain” that is pulled in front of the F layer during the day. The net effect is that AM radio signals can travel great distances at night, but much shorter distances during the day.

43.2 The Crystal Radio

We'll examine the operation of a simple radio receiver by looking in detail at the design of a simple *crystal radio receiver*. This is one of the first types of radio receiver, and has been in use since the 1920s. A crystal radio can be built from just a few spare parts—in fact, soldiers during World War II would often build a variety of crystal radio called a “foxhole radio” from wire, scrap wood, a razor blade, toilet paper tube, a safety pin or pencil lead, and headphones.

One remarkable feature of a crystal radio is that it requires *no batteries*: it runs entirely on the power provided by the transmitter. Once you build a crystal radio, you can run it forever for free.

Tuning Circuit

The crystal radio circuit begins with a tuned LC circuit (Fig. 43.1(a)). The LC combination is designed to oscillate at the same frequency as the AM radio signal to be received.

Recall from Chapter 39 that an LC circuit with inductance L and capacitance C oscillates with angular frequency

$$\omega = \frac{1}{\sqrt{LC}}. \quad (43.1)$$

Since $f = \omega/2\pi$, the frequency (in hertz) is

$$f = \frac{1}{2\pi\sqrt{LC}}. \quad (43.2)$$

Antenna and Ground

Now let's add an *antenna* (or *aerial*) and ground to the tuned LC circuit (Fig. 43.1(b)). The antenna is typically just a long wire (50–100 ft.) strung outdoors, up into a tree or other tall structure if possible. The ground connection is a connection to a long conductor, typically the Earth itself. A traditional ground connection is a connection to a copper pipe driven into the ground, or a connection to a copper cold-water pipe (which also goes to the ground).

The antenna is a large conductor that picks up radio signals of all frequencies and feeds them to the LC circuit. But the LC circuit only resonates with those input signals that are at the frequency given by Eq. (43.2). The ground connection essentially gives the current someplace to go; without a good ground, the current would get “backed up” in the circuit, and the radio would not operate.

The Crystal

Now that we have a circuit resonating at the frequency of the carrier wave (the frequency at which the radio station is transmitting), we need to extract the audio signal. In a crystal radio, this is done with a *diode*: a

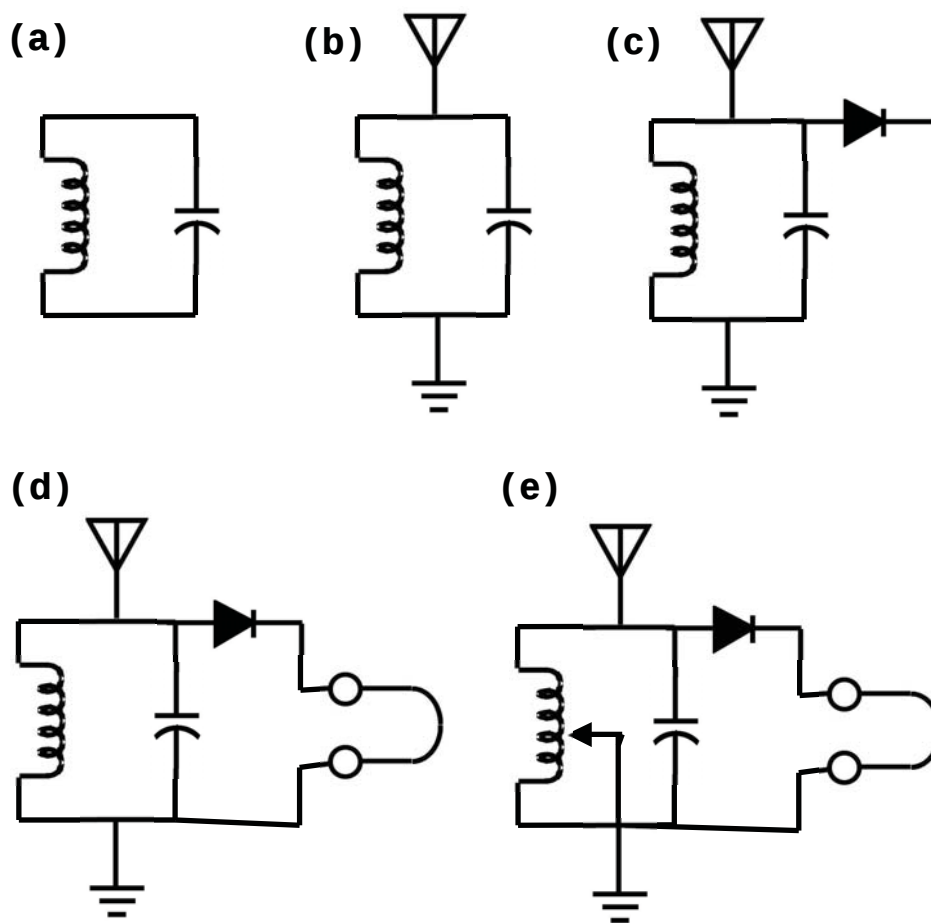


Figure 43.1: Construction of a crystal radio receiver. (a) Tuned LC circuit. (b) Addition of an antenna and ground to drive the LC circuit. (c) Addition of a diode to rectify the signal and extract the audio. (d) Adding headphones completes the radio circuit — the headphones convert the electrical signal from the diode to sound. (e) Making the inductor or capacitor variable allows the radio circuit to tune different stations. Shown here is a variable inductor.



Figure 43.2: A galena crystal with cat's whisker. (©GNU-FDL,Wikimedia Commons.)

kind of one-way valve that allows current to travel in one direction, but not another. By connecting a diode to the tuned LC circuit (Fig. 43.1(c)), we get just one-half of the resonating signal. That is, the incoming signal will cause the LC circuit to oscillate back and forth with an alternating current, where the current sloshes back and forth, going clockwise, then counterclockwise, then clockwise again, etc. The diode allows only the current going in one direction to pass, which allows us to pick up the audio signal that is modulated on the carrier wave.

If we were to connect headphones directly to the LC circuit with no diode, we would hear nothing. The LC current sloshing back and forth would average out to zero, so we would hear no audio. Adding the diode leaves a net non-zero signal coming out of the diode, which has the audio signal in it.

In a traditional crystal radio of the 1920s, a simple diode was constructed from a crystal of the mineral *galena*, which is a heavy silvery metallic mineral consisting of crystalline lead sulfide (PbS). The galena crystal was touched with a fine wire called a *cat's whisker*. The cat's whisker was attached to a movable arm so that it could be placed in contact with different areas of the galena crystal surface (Fig. 43.2). At some point you would find a "sensitive" area of the crystal that would allow the whole assembly to act as a diode, and conduct current in only one direction.

In building a "foxhole radio," soldiers found that galena crystals were very difficult to come by. Instead, they would substitute a razor blade, and used a safety pin or pencil lead as the cat's whisker. This was somewhat less satisfactory than a galena crystal, but was often adequate for picking up a station or two.

In more modern crystal radios, we often replace the galena crystal and cat's whisker with a germanium diode (called a *1N34 germanium diode*). This kind of diode contains a tiny crystal of germanium metal and tiny cat's whisker wire already placed so that the device will always conduct current in just one direction.

Headphones

Finally, we connect a set of headphones or crystal earpiece to the circuit (43.1(d)). This takes the signal coming from the diode and uses it to drive the vibration of a diaphragm that produces sound waves that can be heard by the ear.

Variable Tuning

The radio built so far can tune only one station, whose frequency is at the resonant frequency of the LC circuit. By making either the inductor or the capacitor variable (Fig. 43.1(e)), the circuit can be made to tune different stations. Typically we choose an inductor with $L = 250 \mu\text{H}$ and a capacitor with $C = 365 \text{ pF}$, which gives a resonant frequency (Eq. 43.2) of $f = 527 \text{ kHz}$, which is at the lower end of the AM radio band. If the inductor can vary between 0 and $250 \mu\text{H}$ or the capacitor can vary between 0 and 365 pF , then the

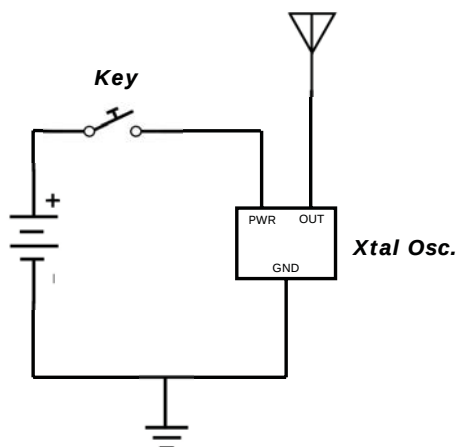


Figure 43.3: Simple radio transmitter for sending Morse code.

resonant frequency of the LC circuit can vary between a lowest frequency of 527 kHz and an upper frequency of, in theory, infinity (if L or C is zero).

Other Issues

The radio receiver described here is a very simple one, but will allow you to pick up strong nearby signals. There are many ways to improve on this circuit. For example, a carefully designed antenna can allow the radio to pick up weaker stations. Adding more sophistication to the circuit can increase its *selectivity*, allowing you to separate stations that are close together in frequency.

43.3 The Radio Transmitter

Suppose you were stranded on a deserted island, and needed to build a simple radio transmitter to signal passing ships or nearby islands so that you could be rescued. How could you do it?

A radio *transmitter* is a device that creates modulated radio waves that can be picked up by a receiver such as the crystal radio receiver described earlier. A very simple radio transmitter can be constructed from a battery, a *crystal oscillator*, and some wire.

A crystal oscillator is a circuit at the heart of which is a small crystal of quartz—a transparent mineral made of silicon dioxide (SiO_2). Quartz is chosen because it exhibits a *piezoelectric effect*, meaning that applying an electric field to the crystal causes it to flex a bit, and flexing the crystal creates an electric field. The crystal oscillator circuit is designed to flex the crystal, then feed any resulting voltage back to the crystal again; this feedback process causes the crystal to oscillate at its natural resonant frequency, and produces an output signal at a well-defined frequency. A crystal oscillator circuit like this is used as the time basis of a quartz watch, for example.

Figure 43.3 shows a very simple radio transmitter for sending Morse code signals. The battery powers the crystal oscillator, whose output is connected to an antenna. The telegraph key is used to turn power to the transmitter on and off. While the telegraph key is held down, the circuit causes the antenna to emit a radio wave at a frequency equal to the crystal oscillator's output frequency. Holding down the key for a short time transmits a “dot”, while holding it down for three times as long as a dot transmits a “dash”. These dots and dashes form the elements of Morse code (Figure 43.4).

International Morse Code

- 1 dash = 3 dots.
- The space between parts of the same letter = 1 dot.
- The space between letters = 3 dots.
- The space between words = 7 dots.

A	• ■■	V	• • • ■■
B	■■ • • •	W	• ■■ ■■
C	■■ • ■■ •	X	■■ • • ■■
D	■■ • •	Y	■■ • ■■ ■■
E	•	Z	■■ ■■ • •
F	• • ■■ •	.	• ■■ • ■■ • ■■
G	■■ ■■ •	,	■■ ■■ • • ■■ ■■
H	• • • •	?	• • ■■ ■■ • •
I	• •	/	■■ • • ■■ •
J	• ■■ ■■ ■■	@	• ■■ ■■ • ■■ •
K	■■ • ■■	1	• ■■ ■■ ■■ ■■
L	• ■■ • •	2	• • ■■ ■■ ■■
M	■■ ■■	3	• • • ■■ ■■
N	■■ •	4	• • • • ■■
O	■■ ■■ ■■	5	• • • • •
P	• ■■ ■■ •	6	■■ • • • •
Q	■■ ■■ • ■■	7	■■ ■■ • • •
R	• ■■ •	8	■■ ■■ ■■ • •
S	• • •	9	■■ ■■ ■■ ■■ •
T	■■	0	■■ ■■ ■■ ■■ ■■
U	• • ■■		

Figure 43.4: The International Morse Code.

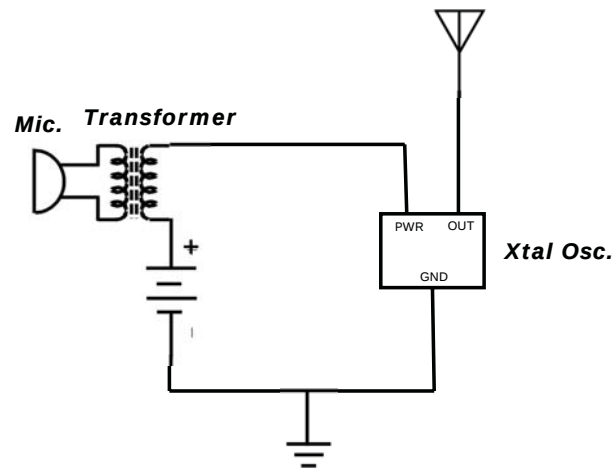


Figure 43.5: Simple transmitter with a microphone.

A more sophisticated transmitter can be built by replacing the telegraph key with a microphone (Figure 43.5). The microphone is coupled to the transmitter circuit via a *transformer*, which consists of two separate coils of wire wrapped around a common core. The transformer also serves to amplify the microphone's signal as it is sent to the the rest of the transmitter circuit.

Part V

Optics

Chapter 44

Geometrical Optics

Optics is the branch of physics concerned with the study of light. It may be divided into three main areas:

1. *Geometrical optics* is the study of mirrors, lenses, and the images formed by these devices. Geometrical optics generally ignores the wave nature of light.
2. *Physical optics* studies phenomena related to the wave nature of light: interference, diffraction, polarization, and so on.
3. *Photometry* is the study of the brightness of light.

We'll begin our study of optics in this chapter with geometrical optics, studying mirrors first, then lenses.

44.1 Law of Reflection

We begin with the simplest of the laws of optics, the *law of reflection*. The law of reflection states that when a light ray strikes a reflective surface (e.g. a mirror), it will reflect off of that surface at an angle equal to its incident angle:

$$\theta_i = \theta_r \tag{44.1}$$

Here θ_i is the *angle of incidence*, and θ_r is the *angle of reflection*. In optics, all angles are by convention measured with respect to the *normal* (perpendicular) to the surface.

Chapter 45

Mirrors

A *mirror* is a reflective surface. By using curved mirrors, it is possible to form an optical *image* of a real object. The simplest curved mirror is called a *spherical mirror*, so called because it can be thought of as being a circle punched out of a hollow sphere that is silvered on one side. If we punch a circle out of a hollow sphere that is silvered on the *inside*, we get a *concave mirror*. If the sphere is instead silvered on the *outside*, we get a *convex mirror*. (Figs. 45.1 and 45.2.) The radius of the (imaginary) sphere that the mirror is “punched out of” is called the *radius of curvature* of the mirror. The point that would be at the center of this sphere is called the *center of curvature* of the mirror.

Ideally, to form a perfect image, the mirror should be in the shape of a *paraboloid*. However, spherical mirrors are easier to manufacture, and can be almost as good, although the deviation from the ideal paraboloidal shape does give rise to an optical defect called a *spherical aberration*, to be described later.

A concave mirror causes light to reflect in towards the axis of the mirror, and is called a *converging* mirror. A convex mirror causes light to reflect away from the axis, and is called a *diverging* mirror.

Light coming from an object infinitely far away will come together at a single point in a concave (converging) mirror; this point is called the *focus* of the mirror, and the distance between the mirror and the focus is called the *focal length* of the mirror. It turns out that the focus is located half-way between the lens and the center of curvature, so that we have

$$f = \frac{R}{2}, \quad (45.1)$$

where f is the focal length and R is the radius of curvature.

The typical problem in mirror optics is this: we are typically given:

- The distance between the object and the mirror, called the *object distance*, d_o .
- The “height” (size) of the object, called the *object height*, h_o .
- The focal length of the mirror, f . (If f is not known, it can be determined from the radius of curvature using Eq. (45.1).

We typically wish to find:

- The distance between the image and the mirror, called the *image distance*, d_i
- The “height” (size) of the image, called the *image height*, h_i
- The *magnification* of the image, m . This is a dimensionless number that indicates how much bigger the image is than the original object.

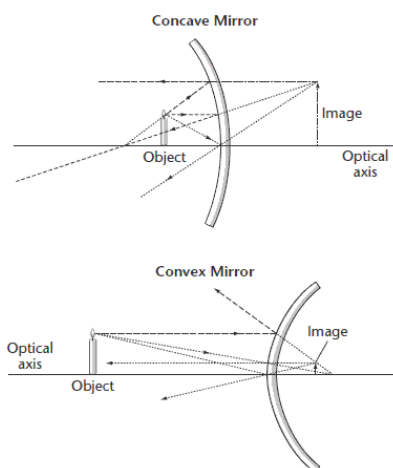


Figure 45.1: Types of mirrors. *Top:* Concave mirror. *Bottom:* convex mirror. (Credit: education.com)

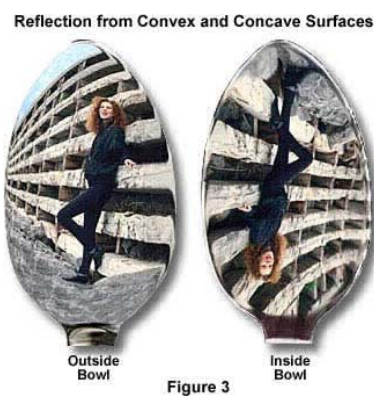


Figure 45.2: Types of mirrors, as illustrated in a table spoon. *Left:* The bottom of a spoon forms a convex mirror. *Right:* The top surface of a spoon forms a concave mirror. (Credit: Florida State University.)

- Whether the image is *real* or *virtual*. (In a real image, light is present at the image location, and the image can be projected onto a screen. In a virtual image, there is no light present; a virtual image cannot be projected onto a screen.)
- Whether the image is *upright* (rightside-up) or *inverted* (upside-down).

There are two methods that can be used to solve this type of problem:

- The *ray diagram method* is a graphical method. It gives a good intuitive picture of what's going on, but it can be a bit time-consuming, and is not particularly accurate.
- The *algebraic method* uses only algebra. It doesn't give a good picture of what's happening, but it's faster and more accurate. However, the algebraic method requires that you are very careful with the equations, particularly with regard to getting the signs correct.

We'll cover both methods here.

45.1 Ray Diagrams

A *ray diagram* is used to locate the image produced by a mirror. To create such a diagram we draw the mirror, its axis, the object, and three light rays, as shown in Fig. 45.3. We also need to locate the focus F and center of curvature C along the mirror's axis. The three rays we draw are:

1. In parallel to axis, out through the focus.
2. In through the focus, out parallel to the axis.
3. In through the center of curvature, and back out through the center of curvature.

(Only two rays are really needed; the third acts as a check.) The image will be located at the point where the three outgoing rays meet, as shown in the figure. If the outgoing rays do *not* meet (i.e. they diverge), then trace the outgoing rays back behind the mirror; in this case you will have a virtual image.

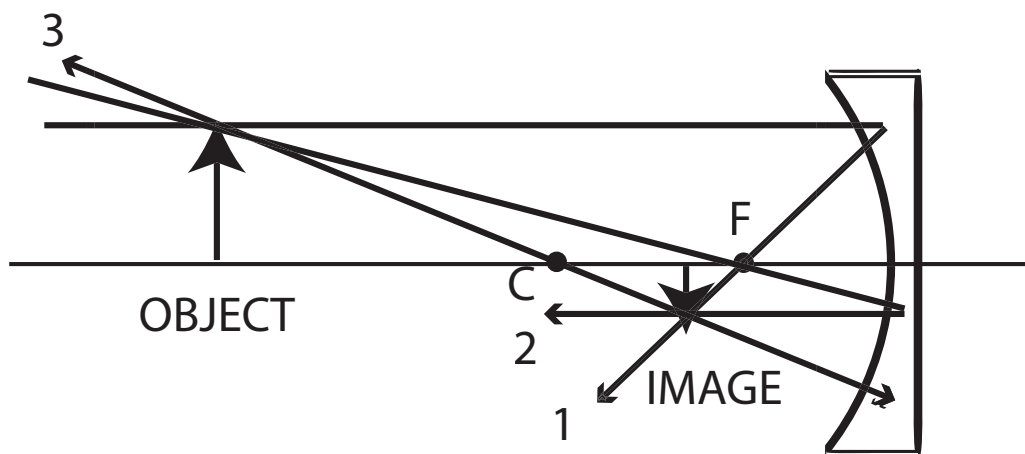


Figure 45.3: Ray diagram for a converging (concave) mirror.

If a ray diagram is drawn with great care and correctly to scale, it may be used to measure (with a ruler) the image distance d_i and image height h_i . You can tell whether the image is real or virtual, or whether it is upright or inverted, simply by inspecting the diagram.

45.2 Algebraic Method

An alternative to the ray diagram method is the *algebraic method*. This is simpler, faster, and more accurate than the ray diagram method, but it does not give a good intuitive picture of what's going on. Also, it is *very* easy to make a sign error with the algebraic method and get the wrong answer.

Solving a mirror optics problem algebraically involves three equations:

1. *Focal length equation*. If we aren't given the focal length, we can find it from the radius of curvature using the equation given earlier:

$$f = \frac{R}{2} \quad (45.2)$$

2. *Mirror equation*. This equation relates the image and object distances to the focal length:

$$\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f} \quad (45.3)$$

Typically one is given the object distance and focal length, and solves this for the image distance d_i .

3. *Magnification equation*. This equation lets us find the image height h_i and magnification m :

$$m = \frac{h_i}{h_o} = -\frac{d_i}{d_o} \quad (45.4)$$

Typically, you're given the image object distance d_o and object height h_o , and have found the image distance d_i from the mirror equation. You can then use this equation to find the image height h_i and magnification m .

When using these equations, it is *very* important that you give each quantity the correct *sign*. The sign convention for mirrors is shown in Table 45-1.

Table 45-1. Sign conventions for mirrors.

Variable	+	−
d_o	real object	virtual object
d_i	real image	virtual image
h_o	always	—
h_i, m	upright image	inverted image
f	converging mirror	diverging mirror

By inspecting the sign of d_i (which you find from the mirror equation), you can determine whether the image is real or virtual. Also, when you compute h_i , its sign will tell you whether the image is upright or inverted. So the equations above give you not only the image distance d_i and image height h_i , but their *signs* give you additional information about the image (real/virtual, upright/inverted).

45.3 Segmented Mirrors

For astronomical telescopes, the bigger the mirror, the more light is collected and the better the resolution—so generally bigger is better. But there is a limit to how large one may build a mirror in a reflecting telescope: at some point very large mirrors become too costly impractical to manufacture.

A relatively new trend in large-mirror technology is to create large astronomical mirrors not as a single large mirror, but as a collection of segments (typically hexagons) that are fitted together to form one large mirror. The smaller mirrors are easier to deal with (although they must be formed to complex asymmetrical shapes), and if one breaks, it can be replaced much more easily than replacing an entire large mirror. A disadvantage is that the segments are subjected to various deformations due to temperature changes, mechanical stress, etc. that can easily place the segments out of alignment with each other. Keeping all the segments properly aligned requires that each segment's position be controlled by a computer, in a system called *active optics*.

Several ground-based segmented-mirror telescopes have already been built, and the upcoming James Webb Space Telescope will incorporate segmented mirrors.

Chapter 46

Refraction

Light travels fastest (299,792,458 m/s) when it's traveling through a vacuum. If light is traveling through some other material, it slows down by a factor called the *index of refraction*. The index of refraction n is a dimensionless number defined by

$$n = \frac{c}{v}, \quad (46.1)$$

where $c = 299,792,458$ m/s is the speed of light *in vacuum*, and v is the speed of light *in the medium*. Since $v \leq c$ always, this means $n \geq 1$; typically n is some number between 1 and 3, and will depend on the medium.

46.1 Snell's Law

If a light ray travels through some transparent medium and comes to an interface with another transparent medium, the light ray will be bent as it moves into the new medium. This phenomenon is called *refraction*. The ray will bend *toward* the normal if it moves into a medium of higher index of refraction, and *away* from the normal if it moves into a medium of lower index of refraction. The angle at which the ray is refracted is given by *Snell's law*, sometimes called the *law of refraction*:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (46.2)$$

Here n_1 and n_2 are the indices of refraction of the two media, and θ_1 and θ_2 are the angles of the incident and refracted rays with respect to the normal.

In traveling from one medium to another, light will follow the path that takes the *least time*; this idea is called *Fermat's principle*. Using the calculus, it is possible to derive Snell's law from Fermat's principle, and thus show that Snell's law gives the path light must follow in order to travel through the two media in the least time.

46.2 Total Internal Reflection

It may sometimes happen that when light travels from a high-index medium to a low-index medium at a high angle of incidence, that Snell's law gives the sine of the angle of refraction to be greater than 1, so that the angle of refraction is not defined. In this case, light is not refracted into the next medium at all; instead, the light reflects off of the interface between the two media, and back into the higher-index medium. For example, if a light ray in water ($n = 1.33$) is headed for an interface with air ($n = 1.00$) at an angle of

incidence of 80° , Snell's law (Eq. 46.2) gives the sine of the angle of refraction to be 1.31, so the angle of refraction is undefined. No refraction occurs in this case: instead, the light will reflect off of the water-air interface (following the law of reflection), and go back into the water. This phenomenon is called *total internal reflection*.

The critical angle for total internal reflection is given by

$$\sin \theta_c = \frac{n_2}{n_1}, \quad (46.3)$$

where θ_c is the critical angle, the light is incident from medium 1, and n_1 and n_2 are the indices of refraction of the two media ($n_1 > n_2$). For example, the critical angle for total internal reflection for an air-water interface is $\theta_c = \sin^{-1}(1.00/1.33) = 48.75^\circ$; this means that any light ray in water headed toward an interface with air will be reflected back into the water if its angle of incidence is greater than 48.75° . If its angle of incidence is less than this critical angle, the ray will be refracted out into the air.

Chapter 47

Lenses

A *lens* is a disk of transparent material (such as glass or plastic), of which one or both surfaces is curved. The curved surfaces allow the lens to form an optical image of a real object, similar to the way an image is formed by a curved mirror.

Each side of the lens may be either concave or convex (Fig. 47.1). If both sides of the lens are convex, the lens is called *double convex*; if both sides are concave, the lens is called *double concave*. If one side is convex and the other concave, the lens is called a *meniscus* lens. If one side of the lens is flat, the lens is called *plano-convex* or *plano-concave*. In general, if the lens is thicker in the middle than at the edges, the lens will be *converging*, and light will be bent toward the axis; if it is thinner in the middle than at the edges, it will be *diverging*, and light will be bent away from the axis.

Ideally, to form a perfect image, the lens surfaces should be in the shape of *hyperboloids* (of two sheets). However, spherical surfaces are often easier to manufacture, and can be almost as good, although the deviation from the ideal hyperboloidal shape does give rise to an optical defect called a *spherical aberration*, to be described later.

Light coming from an object infinitely far away will come together at a single point in a converging (e.g. double-convex) lens; this point is called the *focus* of the lens, and the distance between the lens and the focus is called the *focal length* of the lens.

The typical problem in lens optics is the same as in mirror optics: we are given

- The distance between the object and the lens, called the *object distance*, d_o .
- The “height” (size) of the object, called the *object height*, h_o .
- The focal length of the lens, f . (If f is not known, it can be determined using the *lens maker’s equation*, Eq. (47.1).)

We typically wish to find:

- The distance between the image and the lens, called the *image distance*, d_i
- The “height” (size) of the image, called the *image height*, h_i
- The *magnification* of the image, m . This is a dimensionless number that indicates how much bigger the image is than the original object.
- Whether the image is *real* or *virtual*. (In a real image, light is present at the image location, and the image can be projected onto a screen. In a virtual image, there is no light present; a virtual image cannot be projected onto a screen.)

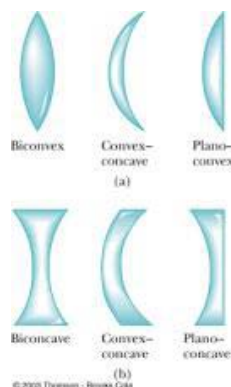


Figure 47.1: Types of lenses. (a) Converging lenses, left to right: biconvex, meniscus, plano-convex. (b) Diverging lenses, left to right: biconcave, meniscus, plano-concave.

- Whether the image is *upright* (rightside-up) or *inverted* (upside-down).

Just as with mirrors, there are two methods that can be used to solve this type of problem:

- The *ray diagram method* is a graphical method. It gives a good intuitive picture of what's going on, but it can be a bit time-consuming, and is not particularly accurate.
- The *algebraic method* uses only algebra. It doesn't give a good picture of what's happening, but it's faster and more accurate. However, the algebraic method requires that you are very careful with the equations, particularly with regard to getting the signs correct.

We'll cover both methods here.

47.1 Ray Diagrams

Ray diagrams for lenses are very similar to ray diagrams for mirrors. To create such a diagram we draw the lens, its axis, the object, and three light rays, as shown in Fig. 47.2. We also need to locate the focus F along the mirror's axis. The three rays we draw are:

1. In parallel to the axis, out through the focus.
2. In through the focus, out parallel to the axis.
3. In through the center *of the lens*, out through the center of lens.

Notice one difference between these rays and the rays used for mirror diagrams: for mirrors, the third ray is through the center *of curvature*; for lenses, the third ray is through the center *of the lens*.

A complication arises with lenses that did not occur with mirrors: while mirrors have a single focus, lenses have *two* foci. So which focus should you use for rays 1 and 2? It depends on whether you have a converging lens or a diverging lens, as shown in the following table. (Here “near” refers to the focus closer to the object, and “far” is the focus farther from the object.)

Ray	Converging	Diverging
1	far	near
2	near	far

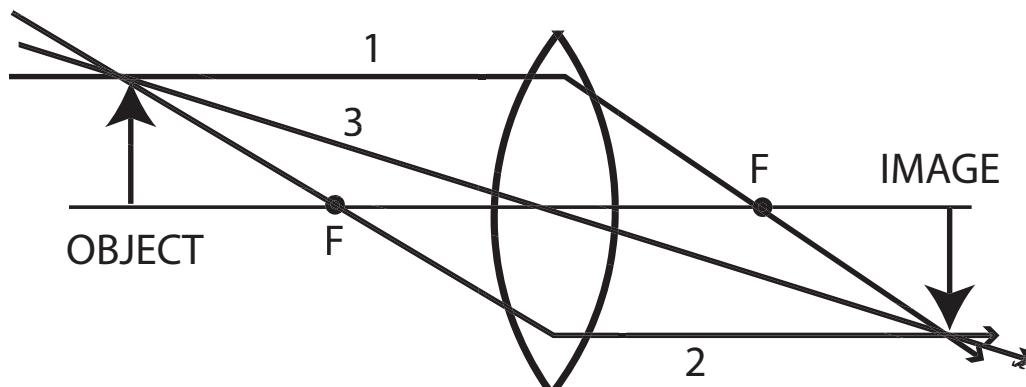


Figure 47.2: Ray diagram for a converging (bi-convex) lens.

47.2 Algebraic Method

An alternative to the ray diagram method is the *algebraic method*. This is simpler, faster, and more accurate than the ray diagram method, but it does not give a good intuitive picture of what's going on. Also, it is very easy to make a sign error with the algebraic method and get the wrong answer.

Solving a lens optics problem algebraically involves three equations:

1. *Lens maker's equation*. If we aren't given the focal length, we can find it from the radii of curvature of the two lens surfaces and the index of refraction of the lens material, using the *lens maker's equation*:

$$\frac{1}{f} = \left(\frac{n_{\text{lens}}}{n_{\text{air}}} - 1 \right) \left(\frac{1}{R_1} + \frac{1}{R_2} \right) \quad (47.1)$$

where R_1 and R_2 are the radii of curvature of the two surfaces, n_{lens} is the index of refraction of the lens material, and $n_{\text{air}} = 1$ is the index of refraction of the air.

2. *Thin lens equation*. This equation relates the image and object distances to the focal length, and is identical in form to the mirror equation:

$$\frac{1}{d_i} + \frac{1}{d_o} = \frac{1}{f} \quad (47.2)$$

Typically one is given the object distance and focal length, and solves this for the image distance d_i .

3. *Magnification equation*. This equation (which is the same as it is for mirrors) lets us find the image height h_i and magnification m : Magnification equation:

$$m = \frac{h_i}{h_o} = -\frac{d_i}{d_o} \quad (47.3)$$

Typically, you're given the image object distance d_o and object height h_o , and have found the image distance d_i from the thin lens equation. You can then use this equation to find the image height h_i and magnification m .

When using these equations, it is *very* important that you give each quantity the correct *sign*. The sign convention for lenses is shown in Table 47-1, and is essentially the same as the sign convention for mirrors.

Sign convention:

Table 47-1. Sign conventions for lenses.

Variable	+	−
d_o	real object	virtual object
d_i	real image	virtual image
h_o	always	—
h_i, m	upright image	inverted image
f	converging lens	diverging lens
R_1, R_2	convex surface	concave surface

47.3 The Fresnel Lens

If you examine the path of a light ray through a lens carefully, you'll notice there is a fair amount of “unused” glass. The incoming light ray is refracted (bent) when it first hits the surface of the lens, then travels in a straight line all the way through the lens, then is refracted again on the way out. The surfaces of the lens do all the work—it seems like all that glass inside the lens is kind of a waste, doing nothing but allowing the light to travel in a straight line. For a large lens, it might be nice to eliminate all that unused glass; is that possible?

Yes, we can eliminate all that unused glass, as shown in Fig. 47.3. The result is called a *Fresnel lens*. Its advantage is that a very large lens can be made very flat—for example, a reading lens can be made the size of a sheet of paper, with roughly the thickness of a credit card. The disadvantage, as seen from the figure, is that the process of eliminating the “unused” parts of the lens leaves behind a series of “steps” or ridges that appear as rings in the lens. A Fresnel lens is therefore not suitable where high-quality optics are needed, but it can be useful as a reading lens, for an overhead projector, or for making something like a solar furnace that focuses sunlight to produce heat.

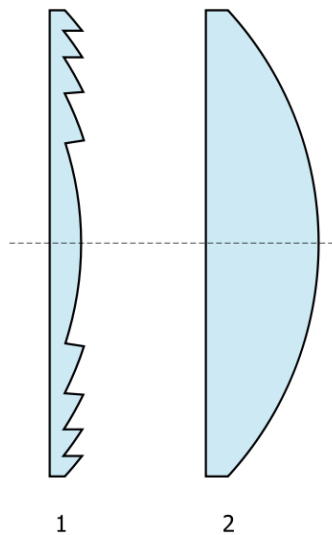


Figure 47.3: Cross section of (1) a Fresnel lens, and (2) the equivalent normal lens. (©GNU-FDL, Wikimedia Commons.)

Chapter 48

Optical Defects

A number of defects, or *aberrations* can occur with mirrors and lenses that prevent them from forming an ideal image. A few of these defects are described here.

48.1 Spherical Aberration

If the surface of a mirror deviates from its ideal paraboloidal shape, or the surface of a lens deviates from its ideal hyperboloidal shape, (for example, if the optical surfaces are sections of spheres), then the mirror or lens is said to have a *spherical aberration*. If a lens or mirror has a spherical aberration, then light rays far from the axis focus at a different point than light rays near the axis, causing a blurring of the image. (See Fig. 48.1, top.)

48.2 Chromatic Aberration

In lenses, light of different wavelengths will generally focus at different points. This phenomenon (to be described later) is called *dispersion*, and is the variation of index of refraction with wavelength. This effect in lenses causes a defect called *chromatic aberration*, which causes the image to be surrounded by a rainbow-like halo. It can be corrected by using combinations of several lenses, each made of a material of a different index of refraction. Chromatic aberration does not occur in mirrors. (See Fig. 48.1, bottom.)

48.3 Astigmatism

Astigmatism is caused by an asymmetrical lens or mirror, and causes light along different axes to be focused at different points.

48.4 Coma

Another type of optical aberration is called *coma*. Coma does not affect light rays parallel to the optical axis, but light rays from objects *off-axis* tend to be smeared into a comet-like shape.

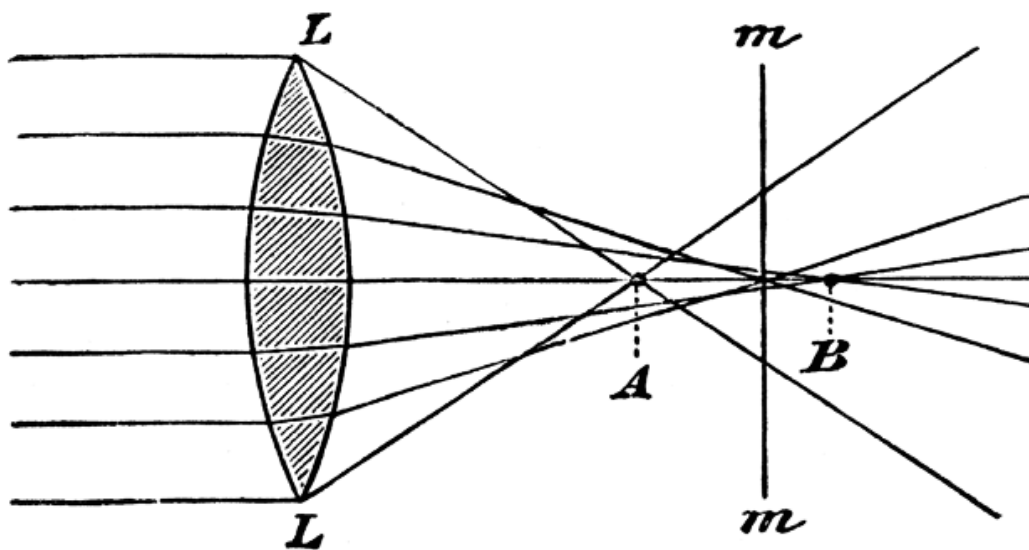


Fig. 1.

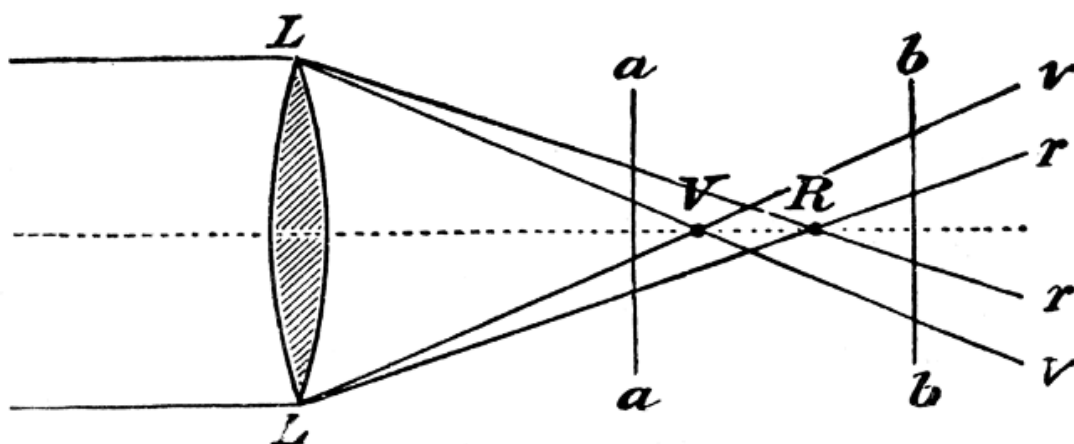


Figure 48.1: Optical defects in a lens. *Top*: Spherical aberration. Rays incident near the edge of the lens focus closer to the lens (A), while rays near the optical axis focus farther from the lens (B). *Bottom*: Chromatic aberration. Red light comes to a focus at point R , while violet light comes to a focus at a different point, V . (Ref. [15])

Chapter 49

Optical Instruments

In this chapter, we'll examine a number of common optical instruments, both natural and man-made. These instruments are designed to form and record simple images (the eye and the camera), to make enlarged images of very small, close objects (the magnifying glass and microscope) or of very distant objects (the telescope), or to simply shift an image through some distance (the periscope).

49.1 The Magnifying Glass

In its simplest form, a *magnifying glass* or *magnifier* consists of a single converging (convex) lens. The human eye can normally get as close as about 25 cm from an object and still have it comfortable in focus; this is called the *near point*. By placing a magnifying glass near the eye (so the eye is closer to the lens than the focus), an enlarged virtual image of the object is created. (See Figure 49.1.)

High-power magnifying glasses often contain a compound lens, consisting of two or more single lenses cemented together. The combination of lenses can help correct unwanted optical defects.

49.2 The Human Eye

The *human eye* is a naturally occurring optical instrument that gives humans their sense of sight. The active optical components are the *cornea* and the *lens*. Most of the focusing of the image is done by the cornea, while the lens acts as a secondary optical element. The image produced by the cornea and lens is focused onto the *retina* on the back of the eye. (Figure 49.2.) The image projected onto the retina is actually upside-down; our brains invert the image so that we seem to see the image rightside-up.

The retina is covered with a grid of two kinds of light detectors: *rods* can detect very faint light, but produces only black-and-white images. *Cones* require a somewhat brighter light level before they activate, but they can see in color. There are three types of cones: one type is most sensitive to red light, another most sensitive to green light, and another most sensitive to blue light. The brain receives signals of different strengths from each type of cone at each location in the image, and from that is able to infer the color of that part of the image. (Any color can be formed from combinations of the three primary colors red, green, and blue; see chapter 56 on color.) From the retina, signals are transmitted to the brain via the *optic nerve*.

Things are a bit more complicated than this, though. The eye and brain are able to perceive objects as having a constant color, even when viewed under widely varying lighting conditions. It is believed that the eye views an entire image (by means of cones) in each of the three primary colors, then sends this information to the brain; the brain then determines the correct color both from the strength of the signals from the different colors of cones, and also by comparing the perceived brightness of each part of the image with those of

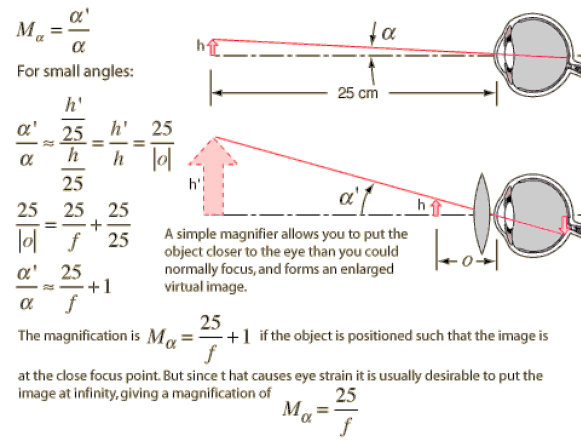


Figure 49.1: The principle of the magnifying glass. (Credit: "Hyperphysics," Georgia State University, <http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html>)

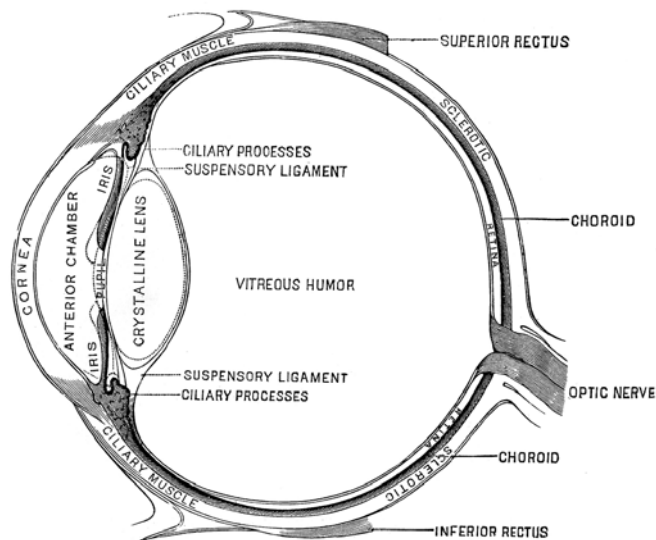


Figure 49.2: The human eye. (Ref. [2])

adjacent parts. Somehow, in a way that is not fully understood, the brain processes this information to deduce the color of the object, largely independent of the color of the light source. This phenomenon is called the *Land effect*, and you can see it for yourself: check the color of some object under fluorescent light indoors, and compare it with the color when seen outdoors in sunlight. Its color will not appear to have changed, even though fluorescent lights are tinted somewhat blue and sunlight is somewhat redder.

The human eye is capable of *accommodation*, meaning that it is able to use muscles to automatically focus images. Sometimes this doesn't work properly, though. If the image is formed *before* reaching the retina, the person has a condition known as *myopia*, or *nearsightedness*. For people with this condition, closeup objects appear in focus, but distant images are fuzzy and out of focus. This condition may be corrected by placing diverging (concave) lenses in front of the eye, either with eyeglasses or contact lenses.

If the image has not yet formed when light reaches the retina, then the person has a condition called *hyperopia*, or *farsightedness*. For people with this condition, distance objects are clear, but closeup objects are out of focus. This condition may be corrected by placing converging (convex) lenses in front of the eye.

A condition of perfect vision (neither myopia nor hyperopia) is called *emmetropia*.

Many people upon reaching their 40s have difficulty focusing on closeup objects because the eye's accommodation abilities are not as robust as they were during youth — a condition called *presbyopia*. Older people often require converging lenses (reading glasses) to see close objects. People who have presbyopia *and* either myopia or hyperopia often wear either reading glasses with contact lenses, or eyeglasses with *bifocal* lenses, which are shaped so that looking through the top half of the lens corrects for distance vision, while looking through the lower half corrects for close-up vision. Somewhat less common are *trifocal* lenses, which correct for distant, mid-range, and close-up vision when looking through the top, middle, and bottom of the lenses, respectively. A recent innovation being offered by ophthalmologists and optometrists is *computer glasses*, which are similar to reading glasses, but designed to help the wearer focus clearly at the typical distance of a computer monitor.

49.3 The Trilobite Eye

Trilobites are an extinct class of arthropods that were among the first living organisms on Earth. Trilobites pre-dated the dinosaurs; they lived from the early Cambrian period (about 550 million years ago) until the great Permian extinction of 250 million years ago, which almost wiped out all life on Earth.^{1,2} (See Figure 49.3.) Trilobite fossils can be found in great numbers, and range in size from 1 millimeter to as much as 2 feet long.

Most species of trilobites had a pair of *compound eyes*, similar to those found on many species of insects today. A compound eye consists of a grid of a large number of very small lenses, all spaced very closely together. Unlike the flexible lens of the human eye, though, trilobite eyes had rigid lenses composed of the crystalline mineral calcite, and thus lacked the ability of accommodation that human eyes have.

To minimize the effect of optical aberrations, trilobites developed an eye lens in a shape that tended to minimize spherical aberrations. These shapes bear a remarkable resemblance to minimum-aberration lens designs developed by French mathematician and philosopher René Descartes (Fig. 49.4) and by Dutch mathematician and physicist Christiaan Huygens (Fig. 49.5).³

¹The cause of the Permian extinction is not known.

²Some paleontologists believe there may be some small chance that trilobites may still be alive even today, in some unexplored depths of the oceans.

³The Descartes and Huygens minimum-aberration lens designs are based on a mathematical curve now called the *oval of Descartes*. See E.N.C. Clarkson and R. Levi-Setti, "Trilobite eyes and the optics of Des Cartes and Huygens"; *Nature*, **254**, 663–667 (1975).



Figure 49.3: A typical trilobite. This is a fossil specimen of the species *Elrathia kingii*, and is 4.2 cm long. (Credit: www.fossilmuseum.net.)

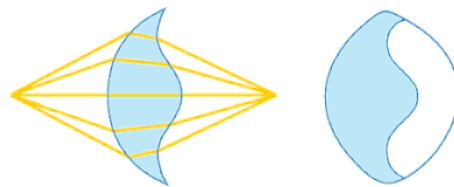


Figure 49.4: Descartes' lens design for minimal aberration (above left) is found in the lens of the trilobite *Crozonaspis* (right). Light ray paths entering the lens from the left come into focus a short distance to the right of the lens (blue). In the eye of *Crozonaspis*, an intralensar body (white) further corrects focus after passing through the outer lens layer (blue). (Credit: "A Guide to the Orders of Trilobites," www.trilobites.info. Image copyright ©1999, 2000 by S.M. Gon III, modified from Clarkson and Levi-Setti, 1975.)

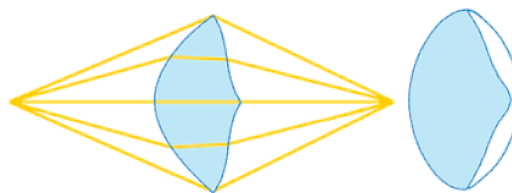


Figure 49.5: Huygens' lens design for minimal aberration (above left) is found in the lens of the trilobite *Dalmanitina* (right). (Credit: "A Guide to the Orders of Trilobites," www.trilobites.info. Image copyright ©1999, 2000 by S.M. Gon III, modified from Clarkson and Levi-Setti, 1975.)

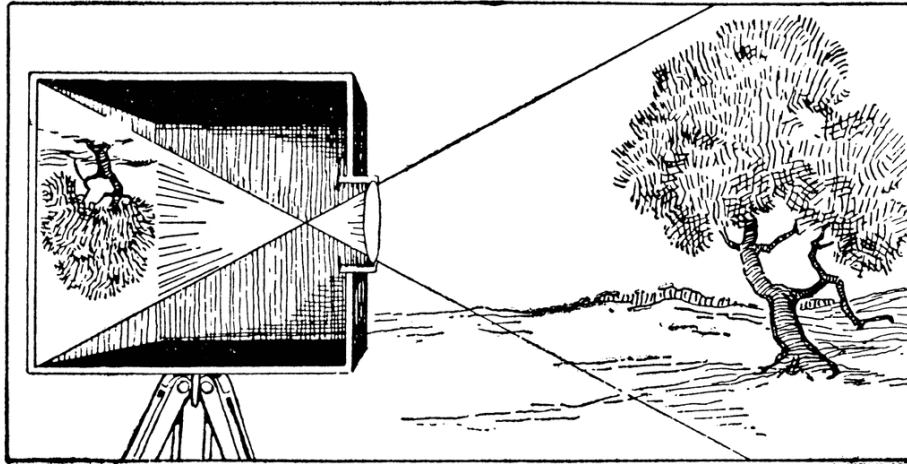


Figure 49.6: A simple camera. (Ref. [4])

49.4 The Camera

A *camera* is an instrument used to record an optical image. It is similar in design to the human eye: the image of an object is focused by a lens onto a plane, where the image is recorded (Fig. 49.6). At one time, images were recorded on chemically-coated glass plates; later, a flexible plastic chemical-coated *film* was used. Since the 1990s, it has become very common to replace the film with a CCD (charge-coupled device) detector that can record a digital image.

A very simple type of camera is called a *pinhole camera*, in which the lens is replaced by a very small hole. One could build a very simple, inexpensive camera by placing (in a darkened room!) photographic film at one end of a lightproof box (a shoebox, for example) that has a covered pinhole at the other end. To take a picture, the pinhole is uncovered for several seconds; the film is then removed in a darkened room and *developed* (chemically processed to bring out the image).

If the distance from the pinhole to the film is L and the wavelength of light is λ , then it can be shown that the optimum diameter d of the pinhole is given by

$$d = \sqrt{2L\lambda}. \quad (49.1)$$

Most modern cameras use a lens instead of a pinhole, and many have a variety of settings to control the focus and aperture size. Focusing is accomplished by changing the distance between the lens and the film plane: the closer the object, the farther the lens must be from the film. This is because of the relation $1/d_i + 1/d_o = 1/f$: decreasing the object distance d_o causes $1/d_o$ to increase; but since f is constant, that means $1/d_i$ must decrease to compensate, which means the image distance d_i must increase. To take extreme closeups, some cameras can be equipped with a set of *extension rings* that allow the lens to be placed very far away from the film. This allows one to take photographs of objects like grains of salt, as if they were being seen under a microscope.

49.5 The Microscope

A *microscope* (from the Greek *μικρος*, “small”, and *σκοπεω*, “see”) is an instrument that allows one to see very small objects—generally much smaller than can be seen with a magnifying glass. (See Fig. 49.7.) The optics consist of a short-focal length *objective lens* that is placed near the object, and an *eyepiece* that

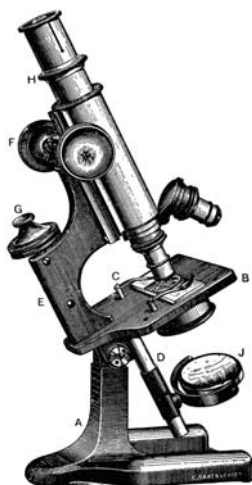


Figure 49.7: A compound microscope. (Ref. [3])

essentially enlarges the image produced by the objective lens. In most modern microscopes, there are several objective lenses mounted on a rotating platform, as well as interchangeable eyepieces, so that the user can select an appropriate combination for the desired magnification.

The microscope is often used in biology to observe cells and protozoa.

49.6 The Telescope

A *telescope* (from the Greek $\tau\eta\lambda\epsilon$, “far”, and $\sigma\kappa\omicron\pi\epsilon\omega$, “see”) is an instrument designed to observe far-away objects. A small hand-held telescope is called a *spyglass* or *monocular*; a pair of such small telescopes mounted side-by-side provide stereo vision and are called *binoculars*.

Larger telescopes are used for astronomical observations. Astronomical telescopes are of one of two types:

- A *refracting telescope* is made of lenses: a large *objective lens*, and a smaller *eyepiece*.
- A *reflecting telescope* consists of one or more curved lenses in place of the objective lens; an eyepiece lens creates the final image.

Refracting astronomical telescopes were built until around 1900; since then, all large astronomical telescopes have been of the reflecting type. This is because there are a number of problems with refracting telescopes that are avoided in reflecting telescope designs. First, there is a limit on how large it is practical to make the objective lens. The lens is made of glass; it is therefore fluid, and will tend to flow. There's not much that can prevent this, since the lens can only be supported by the edges. Second, since light must pass through the lens, it is subject to being scattered by any imperfections (bubbles, etc.) that may be in the glass, which will cause imperfections in the image. Third, since light has to travel through the lens, there is a tendency for light to be lost as it travels through the lens, and so very faint objects are difficult to observe. Fourth, a refracting telescope is subject to chromatic aberration. All of these problems are avoided by reflecting telescopes.

The largest refracting astronomical telescope still in use is at the Yerkes Observatory in Wisconsin; this telescope has an objective lens with a 40-inch diameter. In contrast, reflecting telescopes of over 400 inches diameter have been constructed.

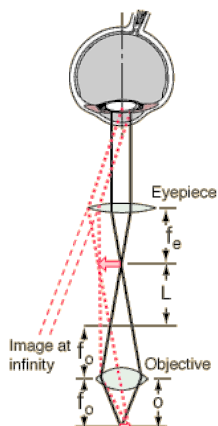


Figure 49.8: Optical principle of the compound microscope. (Credit: “Hyperphysics,” Georgia State University, <http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html>)

Reflecting telescopes are made in a number of different designs (Fig. 49.9). The simplest is the *Newtonian* reflector. In this design, light enters the telescope tube, reflects from a large parabolic *primary mirror*, up to a flat *secondary mirror*, and from there out through the side of the tube to an eyepiece.

In a *Cassegrain*, light reflects from a large parabolic primary mirror to a *hyperbolic* secondary mirror; from there it travels back through a hole in the center of the primary mirror to the eyepiece.

An interesting property of reflecting telescopes is that they produce an *inverted* (upside-down) image. This is not a problem for astronomical observations, but this means that using a reflecting telescope to make terrestrial observations requires the use of an *image erector* — an optical device placed at the eyepiece to make the image rightside-up.

A number of reflecting telescopes have been placed in space, either in Earth orbit or elsewhere. There are a number of reasons for placing a telescope in space:

1. The Earth’s atmosphere is a fluid with turbulent air currents that tend to blur images in ground-based telescopes. By placing the telescope above the Earth’s atmosphere, the telescope no longer need “look” through the atmosphere, so the images are much sharper and more detailed.
2. The Earth’s atmosphere absorbs many wavelengths of light. A telescope in space can observe at wavelengths that are impossible for a ground-based telescope.
3. Since the sky is always dark in space, a space-based telescope can make observations at any time — unlike a ground-based telescope, which can only make observations at night.

49.7 The Periscope

A *periscope* (from the Greek $\pi\epsilon\rho\iota$, “around”, and $\sigma\kappa\omicron\pi\epsilon\omega$, “see”) is an instrument designed to allow the user to observe above or around a barrier. Most famously, periscopes are used in submarines to observe above the water while the submarine remains submerged. In its simplest form, a periscope consists of two mirrors mounted at 45° angles: one at the top, and one at the bottom, where the observer is located (Fig. 49.10).

If you have seen a submarine periscope in use (in real life or in a movie), you will notice that in order to look around, the periscope operator rotates the entire instrument by walking around in a circle. Why not

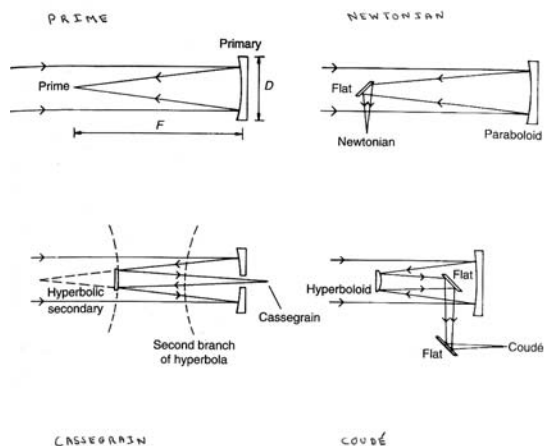


Figure 49.9: Several different designs of reflecting telescopes. Credit: University of New South Wales, Australia.

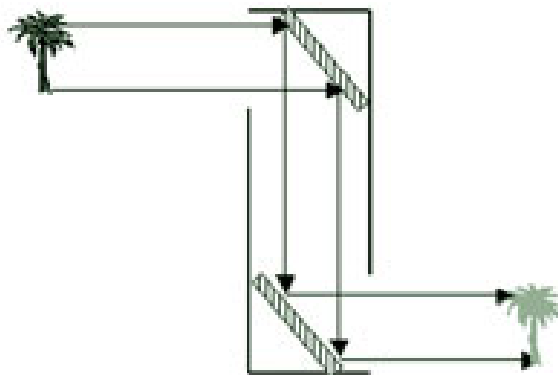


Figure 49.10: Optics of the periscope. The tree is at the upper left, and produces a displaced image at the lower right. (Credit: Arbor Scientific Co., www.arborsci.com.)



Figure 49.11: Image observed in a kaleidoscope. (Credit: “Kaleidoscope Optics” at http://www.4physics.com/phy_demo/kaleidoscope/kaleidoscope-0.html. Image copyright © 4physics.com.)

just rotate the top mirror? Because if the top mirror were rotated, objects behind the observer would appear upside-down. The entire instrument must be rotated to keep the image rightside-up.

49.8 The Kaleidoscope

A *kaleidoscope* (from the Greek *καλός*—“beautiful,” *εἶδος*—“form,” and *σκοπεω*—“see”) is an optical instrument invented by Scottish physicist David Brewster, whose purpose is to produce varying beautiful, colorful patterns for the enjoyment of the user. One end of the kaleidoscope has a rotating disk containing colorful objects like beads or glass. The tube of the kaleidoscope contains mirrors — typically three long rectangular mirrors fastened together to form a prism whose cross section is an equilateral triangle. The user holds the kaleidoscope up to a light source and looks through the mirrors toward the colored objects, and sees the objects reflected multiple times in the mirrors, producing a pleasing colorful design. The design can be modified by rotating the disk of colored objects, producing an endless variety of patterns.

Different mirror configurations are sometimes used to produce images with different symmetry patterns. In one type of kaleidoscope, sometimes called the *teleidoscope*, the disk of colored objects is replaced by a lens, so that patterns are formed from images of whatever objects are in the direction the instrument is pointed.

Chapter 50

Photometry

A frequently neglected area of optics is the field of *photometry*—the study of the measurement of the brightness of light. A related field is *radiometry*, where one measures the intensity of electromagnetic radiation at all wavelengths. In photometry, though, we take into account the physiology of human vision. The goal of photometry is to measure the brightness of visible light, *as it appears to the human eye*.

We begin with a simple mathematical model of human vision. Fig. 50.1 shows such a model, called the *luminous efficiency curve*; it models how the eye’s sensitivity varies with wavelength. As shown by the figure, the human eye is most sensitive to visible light in the green part of the spectrum, at a wavelength of about 555 nm. The eye is much less sensitive to red and violet light, where the curve has values near zero.

50.1 Luminous Flux

We now introduce definitions of some basic photometric quantities. First, the *luminous flux* Φ is the total amount of visible light emitted by a light source, in all directions. Luminous flux is analogous to the total amount of electromagnetic radiation emitted by the light source, except that it is “weighted” by the luminous efficiency curve. For example, electromagnetic radiation with a wavelength near 555 nm is given more “importance” than radiation with a wavelength near 400 nm. This weighted average is luminous flux.

In SI units, luminous flux is measured in units of *lumens* (lm). If you look closely at Fig. 50.1, you’ll see that the vertical axis has units of lumens per watt (lm/W). When we take the intensity of electromagnetic radiation (in watts) and multiply by this luminous efficiency curve to “weight” different wavelengths according to the sensitivity of human vision, we get units of lumens. Note that the peak of the luminous efficiency curve is at $\lambda = 555$ nm, where the human eye is most sensitive; at this wavelength the luminous efficiency is 683 lm/W.

You may see the lumen used on packages of light bulbs, where it may be listed as the “light output”. For example, a typical 60-watt incandescent light bulb may have a luminous flux of 820 lumens. This means that the bulb consumes electric power at the rate of 60 watts (60 joules of energy per second), while producing 820 lumens of light. High-efficiency light bulbs produce more visible light while using less electric power, at the expense of producing less electromagnetic radiation at non-visible wavelengths. For example, a compact fluorescent light bulb may produce 1200 lumens of light, while consuming only 20 watts of electric power. If you’re trying to replace incandescent light bulbs with compact fluorescent bulbs, you should try to find a compact fluorescent bulb that has a luminous flux (in lumens) similar to that of the bulb you’re replacing. *Don’t* replace it with a bulb that has the same power consumption (in watts). For example, a 60-watt incandescent bulb that emits 820 lumens of light should be replaced by a compact fluorescent bulb that emits about 820 lumens of light.

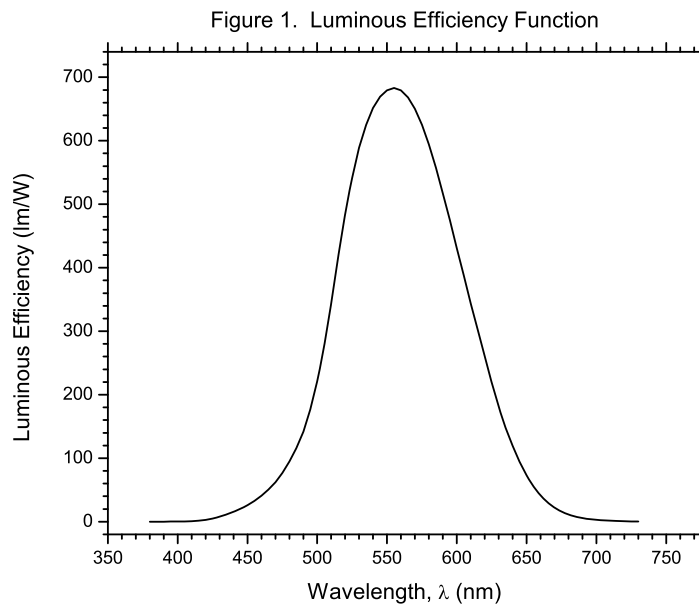


Figure 50.1: The luminous efficiency curve.

50.2 Luminous Intensity

A quantity related to luminous flux is *luminous intensity* I , which is the luminous flux Φ per unit solid angle Ω :

$$I = \frac{\Phi}{\Omega} \quad (50.1)$$

The SI unit of luminous intensity is the *candela* (cd): one candela is equal to one lumen per steradian. A candela is approximately equal an older unit called the *candlepower*, which was the light intensity emitted by the flame of a candle. So a candle flame has a luminous intensity of about 1 candela; by comparison, a 60-watt incandescent light bulb has a luminous intensity of about 65 candelas, while a typical searchlight has a luminous intensity of about 800 million candelas.

If a light source is *isotropic* (so it emits light equally in all directions), then there is a simple relationship between luminous flux Φ and luminous intensity I : $I = \Phi/(4\pi \text{ sr})$.

The candela is the fundamental photometric unit in SI units, and is determined as the result of an experiment. Other photometric units (the lumen and the lux) are defined in terms of the candela.

50.3 Illuminance

The level of illumination seen by an observer is called the *illuminance*. To find the illuminance E , we divide the luminous flux Φ emitted by the light source by the area A over which that luminous flux is spread:

$$E = \frac{\Phi}{A} \quad (50.2)$$

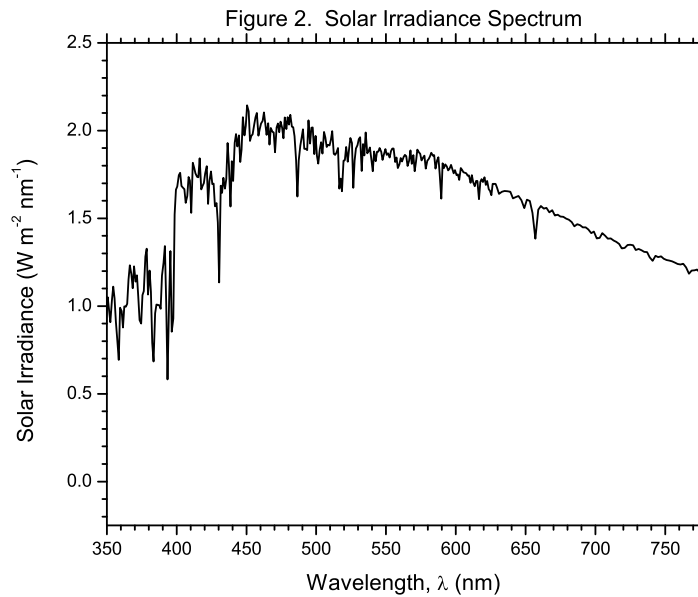


Figure 50.2: Solar irradiance spectrum.

The SI unit of illuminance is the *lux* (lx), where one lux is one lumen per square meter. To give a sense of scale, the level of illumination in a typical office is around 400 lux, while direct sunlight is around 100,000 lux (depending on how high the Sun is in the sky, cloud conditions, etc.).

(An older unit of illuminance, the *foot-candle*, is one lumen per square foot, or about 10.76391 lux.)

50.4 Example: The Sun

As an example of how photometric calculations are done, consider the Sun. To find the illuminance of the Sun at the Earth, we would begin by measuring the intensity of the Sun's radiation at different wavelengths; this gives a plot of the *solar irradiance spectrum*. Fig. 50.2 shows just part of the solar irradiance spectrum—the part that's within visible light wavelengths. We multiply this solar spectrum by the luminous efficiency curve (Fig. 50.1), and find the area under the resulting curve. The result is the illuminance of the Sun's light at the Earth, and works out to be $E = 133,000$ lux.

We can now use Eq. (50.2) to find the luminous flux of the Sun. Here A is the total area over which the luminous flux is spread to give illuminance E , so $A = 4\pi r^2$, where r is the distance of the Earth from the Sun. We find

$$\Phi = EA \quad (50.3)$$

$$= E(4\pi r^2) \quad (50.4)$$

$$= (133,000 \text{ lux})[4\pi(1.4959787 \times 10^{11} \text{ m})^2] \quad (50.5)$$

$$= 3.75 \times 10^{28} \text{ lumens.} \quad (50.6)$$

From this result, we can compute the luminous intensity of the Sun. Since the Sun emits light equally in all directions (is isotropic), the luminous intensity I of the Sun is

$$I = \frac{\Phi}{\Omega} \quad (50.7)$$

$$= \frac{3.75 \times 10^{28} \text{ lm}}{4\pi \text{ sr}} \quad (50.8)$$

$$= 2.98 \times 10^{27} \text{ candelas.} \quad (50.9)$$

In summary, for the Sun, we find

- Luminous flux: $\Phi_s = 3.75 \times 10^{28} \text{ lm}$
- Luminous intensity: $I_s = 2.98 \times 10^{27} \text{ cd}$
- Illuminance at Earth: $E_s = 133 \text{ klx}$

50.5 Example: Incandescent Light Bulb

A 60-watt incandescent light bulb emits a luminous flux of 820 lumens. If this light bulb is isotropic and is the only illumination in a room, then what is the illuminance at a distance of 80 cm from the light bulb?

Solution. From Eq. (50.2), the illuminance E is

$$E = \frac{\Phi}{A} \quad (50.10)$$

$$= \frac{\Phi}{4\pi r^2} \quad (50.11)$$

$$= \frac{820 \text{ lm}}{4\pi (0.80 \text{ m})^2} \quad (50.12)$$

$$= 102 \text{ lx} \quad (50.13)$$

50.6 Astronomical Photometry

In astronomy, the brightness of celestial bodies such as stars and planets is not measured in the photometric units just described; instead, a logarithmic scale of *magnitudes* is employed. The magnitude scale was originally defined so that the brightest stars in the sky are magnitude 0, the dimmest visible to the unaided human eye are magnitude 5, and a magnitude 0 star is 100 times as bright as a magnitude 5 star. (Note that magnitudes measure *dimness*, not brightness. The larger the magnitude, the dimmer the star.) The magnitude scale is logarithmic, so an increase of 1 magnitude corresponds to a decrease in the brightness of the star by a factor of $\sqrt[5]{100} \approx 2.5119$.

There are two types of magnitudes defined: the *apparent magnitude* is the brightness of a star as seen from Earth; a star's apparent magnitude depends both on its intrinsic brightness and on its distance from Earth. The *absolute magnitude* is a measure of intrinsic brightness alone: is the brightness a star would have if it were at a standard distance of 10 parsecs, or 3.0857×10^{17} meters. It is straightforward to show that the apparent magnitude m is related to the absolute magnitude M by

$$M = m - 5(\log_{10} D - 1), \quad (50.14)$$

where D is the distance to the star in parsecs. (Notice that if $D = 10$ parsecs, then this formula gives $M = m$, as expected.)

Some examples: the brightest star in the sky, Sirius (in the constellation Canis Major), has an apparent magnitude of -1.44 . Polaris (the North Star) is a variable star that varies in magnitude, but has an average magnitude of $+1.97$. The Sun has an apparent magnitude of -26.72 , and an absolute magnitude of $+4.85$.

One the brightest stars in the sky is the blue-white supergiant Deneb (in the constellation Cygnus). Most of the bright stars in the sky are around 50–100 light-years from Earth, but Deneb is some 1500 light-years away, so it must be intrinsically very bright. Indeed, Deneb has an apparent magnitude of $+1.25$ and an absolute magnitude of -7.13 .

It is possible to convert between the magnitude scale and conventional photometric units by using the Sun as a calibration point. To convert between apparent magnitude m and illuminance, the formula can be shown to be

$$E = E_s 10^{-\frac{2}{5}(m-m_s)}, \quad (50.15)$$

where E is the illuminance due to the star (in lux), E_s is the illuminance due to the Sun at the Earth ($E_s = 133,000$ lux), m is the apparent magnitude of the star, and m_s is the apparent magnitude of the Sun ($m_s = -26.72$). Similarly, to convert between absolute magnitude M and luminous flux Φ and luminous intensity $I = \Phi/(4\pi \text{ sr})$, the formulæ are found to be (using Eqs. (50.14) and (50.15))

$$\Phi = \Phi_s 10^{-\frac{2}{5}(M-M_s)} \quad (50.16)$$

$$I = I_s 10^{-\frac{2}{5}(M-M_s)} \quad (50.17)$$

where $\Phi_s = 3.75 \times 10^{28}$ lm is the luminous flux of the Sun, $I_s = 2.98 \times 10^{27}$ cd is the luminous intensity of the Sun, M is the absolute magnitude of the star, and $M_s = +4.85$ is the absolute magnitude of the Sun.

Example. The illuminance at Earth due to light from star Sirius (apparent magnitude $m = -1.44$) is

$$\begin{aligned} E &= E_s 10^{-\frac{2}{5}(m-m_s)} \\ &= (133,000 \text{ lux}) 10^{-\frac{2}{5}[-1.44-(-26.72)]} \\ &= 10.28 \mu\text{lx} \end{aligned}$$

Example. The luminous flux Φ of the star Deneb (absolute magnitude $M = -7.13$) is

$$\begin{aligned} \Phi &= \Phi_s 10^{-\frac{2}{5}(M-M_s)} \\ &= (3.75 \times 10^{28} \text{ lm}) 10^{-\frac{2}{5}(-7.13-4.85)} \\ &= 2.32 \times 10^{33} \text{ lm} \end{aligned}$$

which means Deneb is intrinsically $\Phi/\Phi_s = 62,000$ brighter than the Sun. We can similarly find the luminous intensity I of Deneb:

$$\begin{aligned} I &= I_s 10^{-\frac{2}{5}(M-M_s)} \\ &= (2.98 \times 10^{27} \text{ cd}) 10^{-\frac{2}{5}(-7.13-4.85)} \\ &= 1.85 \times 10^{32} \text{ cd} \end{aligned}$$

or 185 *nonillion* candelas.

Chapter 51

Young's Experiment

We now begin a study of *physical optics*, which is the study of the physical properties of light, including its wave nature.

A key experiment in physical optics is *Young's experiment*, first performed by British physicist Thomas Young (1773–1829). In this experiment, one allows a light source to pass through two closely-separated slits and then be projected onto a screen. The light source should be *monochromatic* (that is, of a single wavelength) and *coherent* (each wave train is many wavelengths long). In Young's time such light sources were very faint and difficult to work with, but today we can perform the experiment easily using a *laser* as a coherent monochromatic light source.

The significance of Young's experiment is that it demonstrates that light is a *wave*: on performing the experiment, you find an *interference pattern* of alternating light and dark bands on the screen. At any point P on the screen, the distance from one slit will be different from the distance from the other slit; this difference in distances will be $d \sin \theta$, where d is the separation distance between the slits, and θ is the angle from the midpoint of the slits to the point P . If the path length difference is an integral number of wavelengths, the interference will be constructive, and a *bright fringe* will be observed on the screen:

$$d \sin \theta = m\lambda \quad (m = 0, 1, 2, \dots) \quad (\text{bright fringes}) \quad (51.1)$$

Here m is called the *order* of the fringe. In between the bright fringes, one will see *dark fringes*:

$$d \sin \theta = \left(m + \frac{1}{2}\right)\lambda \quad (m = 0, 1, 2, \dots) \quad (\text{dark fringes}) \quad (51.2)$$

51.1 Quantum Effects

Young's experiment may be used to demonstrate some very odd *quantum mechanical* effects. (*Quantum mechanics* is the theory of mechanics that describes particles at very small distance scales — say at the size of an atom or smaller.) Light is — in some way we don't entirely understand — both an electromagnetic wave and a particle (called a *photon*) at the same time. It's possible to send light through Young's experiment one photon at a time, in which case you would expect the interference pattern to disappear. After all, the interference pattern is caused by light from one slit interfering with light from the other slit, but the photon goes through only one of the two slits. But if we do this experiment, we discover that the photons, one by one, will build up the same interference pattern.

Now if we try to determine *which* slit the photon went through (by bouncing another photon off of it near the slit, for example), the interference pattern disappears: the photon we used to make the determination messes up the experiment in such a way that it destroys the interference pattern. We might try to fix this by

using a lower-energy photon that will minimize the disturbance to the photon we're observing — and if we do this, the interference pattern does indeed return. But a low-energy photon also has a long wavelength, and the wavelength is now sufficiently long that it's no longer possible to tell which slit the original photon went through. It's as though Nature conspires against us to prevent us from determining which slit the photon goes through.

Chapter 52

Diffraction

The bending of waves (including light waves) around obstacles is called *diffraction*. Light has a very short wavelength, but it is possible to observe diffraction in light waves without too much trouble.

One such experiment involves a setup similar to Young's experiment, but using only *one* slit. Light from one part of the slit will interfere with light coming from another part of the slit, creating a *diffraction pattern* as the light waves coming from different parts of the slit interfere with each other. This phenomenon is called *single-slit diffraction*. The positions of the *dark fringes* in single-slit diffraction are given by

$$a \sin \theta = m\lambda \quad (m = 1, 2, 3, \dots) \quad (\text{dark fringes}) \quad (52.1)$$

where a is the slit width, θ is the angle between the midpoint of the slit and the m -th order dark fringe, and λ is the wavelength of light.

In a real Young's experiment, you observe *both* the interference pattern (due to the two slits) *and* single-slit diffraction (due to the finite width of the slits): you will see the interference pattern modulated by an "envelope" of single-slit diffraction.

A similar diffraction effect may be observed when light is incident on a *circular* aperture. In this case, the resulting diffraction pattern is a single central bright circle, surrounded by alternating dark and light rings. The radius of the first dark ring (which can be taken as the radius of the central maximum) subtends an angle

$$\theta_r = 1.22 \frac{\lambda}{D}, \quad (52.2)$$

where θ_r is in *radians*, λ is the wavelength of the light, and D is the diameter of the aperture.¹

52.1 The Rayleigh Criterion

Single-slit diffraction limits the resolving power of astronomical instruments: that is, it places limits on how close two point sources of light can be to each other and still be distinguished as separate points of light. For example, suppose an astronomical telescope is used to observe two stars that are close together. Each star is essentially a point source of light, and will produce a single-slit diffraction pattern as seen through the telescope aperture. If the two diffraction patterns are far apart, you will see two stars. But if the two diffraction patterns are too close together, they will overlap and the image will blur together and look like a single star (Fig. 52.1).

There is a threshold where the stars will be as close together as they can be, and still be distinguished as two separate stars. This threshold is given by the *Rayleigh criterion*. It states that the *minimum* angular

¹The coefficient 1.22 in this equation is the first zero of the Bessel function $J_1(x)$, divided by π . A closer value is 1.2196698912665.

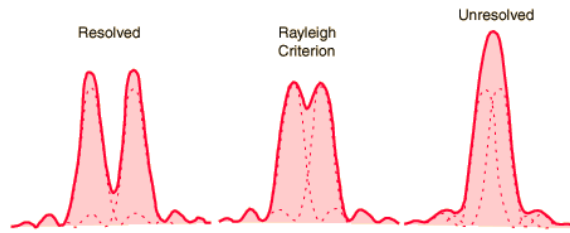


Figure 52.1: Overlapping diffraction patterns and the Rayleigh criterion.

separation of two point sources of light that allows the sources to still be distinguished as two separate points is given by:

$$\Delta\theta = \begin{cases} 1.22 \frac{\lambda}{D} & \text{(circular aperture)} \\ \frac{\lambda}{a} & \text{(rectangular aperture)} \end{cases} \quad (52.3)$$

Here $\Delta\theta$ is the minimum angular separation (in radians), λ is the wavelength of light, D is the diameter of a circular aperture, and a is the width of a rectangular aperture.

52.2 Floaters in the Eye

52.3 The Diffraction Grating

Chapter 53

Optics of the Hubble Space Telescope

53.1 The Hubble Space Telescope

To illustrate the workings of a real optical instrument, let's examine some of the optical details of the Hubble Space Telescope (HST). Hubble is arguably the most successful and productive astronomical instrument of all time, so a study of some of its inner workings will be instructive.

Overview

Hubble is a large astronomical telescope that was placed in orbit around the Earth on April 25, 1990. It is about the size of a school bus, and has a mass of about 11,000 kg. Hubble is in a low-Earth orbit (so it can be serviced by the Space Shuttle), and orbits the Earth about once every 96 minutes. Each orbit is about 1 hour in sunlight (orbit day) and 1/2 hour in darkness (orbit night).

Hubble is designed to make observations of astronomical objects in visible light, near infrared, and near ultraviolet wavelengths—it can observe wavelengths in the range of 100–2500 nm. (Visible light lies within this range, from 400–700 nm.)

The reason Hubble is in orbit around the Earth, rather than on the ground, is to get above the Earth's atmosphere. Turbulence in the Earth's atmosphere causes blurring of the images, which is avoided when the telescope is above the atmosphere. Also, the atmosphere absorbs some wavelengths of light, a complication that is also avoided by being in orbit. Finally, some light is lost when it passes through the atmosphere. By being in orbit above the atmosphere, Hubble avoids this light loss and can see very faint objects.

The Hubble Space Telescope can see objects fainter than magnitude 30 (see Section 50.6) — which is very faint.

Instruments

Unlike ground-based amateur telescopes, there is nobody looking at Hubble's images directly through an eyepiece. Instead, the images observed by Hubble are sent to a complement of scientific instruments (cameras and spectrometers), each of which can perform its own analysis and relay the resulting spectra and images to the ground by radio. The five instruments currently on board Hubble are:

- Wide Field Camera 3 (WFC3)
- Space Telescope Imaging Spectrograph (STIS)
- Cosmic Origins Spectrograph (COS)

- Advanced Camera for Surveys (ACS)
- Near Infrared Camera and Multi-Object Spectrometer (NICMOS)

53.2 HST Optics Overview

The Hubble Space Telescope's optics is all based on *mirrors* (no lenses). Lenses are generally not suitable for large astronomical telescopes for a number of reasons. First, a large lens requires a large solid piece of glass, which are subject to bubbles and other irregularities that degrade the image. Also, some light is always lost when passing through a lens, no matter how carefully the lens is made. Weight is another issue: large lenses are very heavy, but they can only be supported from around the edges, which can cause them to sag under gravity. Finally, lens designers are at the mercy of the optical properties of the glass (such as dispersion) over which they have little control, except for inserting additional corrective lenses. Nevertheless, some lens-based astronomical telescopes (called *refracting telescopes*) are still in use; the largest is the 40-inch diameter telescope at the Yerkes observatory in Wisconsin.

Mirrors, on the other hand, have numerous advantages. They have only one optical surface, so the back of the mirror can be hollowed out to make the mirror lighter. The mirror can be supported along the edges and along the back, so there are fewer problems with sagging. Also, mirrors don't suffer from some optical issues like chromatic aberration that plague lens designers, and don't have the light loss issues that lenses do. For these reasons, most modern large astronomical telescopes use mirrors; these are called *reflecting telescopes*.

The simplest design of a reflecting telescope is a *Newtonian* telescope, in which a single parabolic mirror (the *primary mirror*) forms an image, which is reflected out of the side of the telescope with a flat *secondary mirror* and into an eyepiece. A more compact design, used by many larger reflecting telescopes, is a *Cassegrain* telescope. In this design, light first strikes a curved primary mirror, reflects to a curved secondary mirror, and back through a hole in the primary mirror to the eyepiece. This design allows for a primary mirror with a long focal length to be placed in a relatively small space, since the optical path is "folded" on itself.

The Hubble Space Telescope is a reflecting telescope that is a variation of the Cassegrain design, called a *Ritchey-Chrétien Cassegrain* design. In this design, both the large primary mirror and the smaller secondary mirror are sections of hyperboloids of two sheets. The two hyperboloids work together to focus an image just behind the hole in the primary mirror.

Hubble's primary mirror has a diameter of $D = 2.4$ meters (94.5 inches), and has a focal length of $f = 57.6$ meters. Another parameter often used to characterize astronomical telescopes is the so-called *f*-number, which is defined to be the ratio of the focal length to the aperture diameter:

$$f\text{-number} = \frac{f}{D} \quad (53.1)$$

For Hubble, the primary mirror has an *f*-number of $f/24$.

53.3 Resolution

Because of single-slit diffraction, any astronomical object observed through a telescope with a finite aperture will create a diffraction pattern, and this diffraction effect limits the resolution of the image. In general, the larger the aperture of the telescope, the better the resolution (and also the fainter the objects it can see, since it can collect more light).

The *resolution* of an astronomical telescope (or other optical device) is defined to be the smallest angular separation of two point sources of light that will still allow them to be resolved as individual point sources, despite their overlapping diffraction patterns. The exact point at which two adjacent diffraction patterns are overlapping "too much" is a bit vague, but one commonly used definition is the *Rayleigh criterion*. Under the

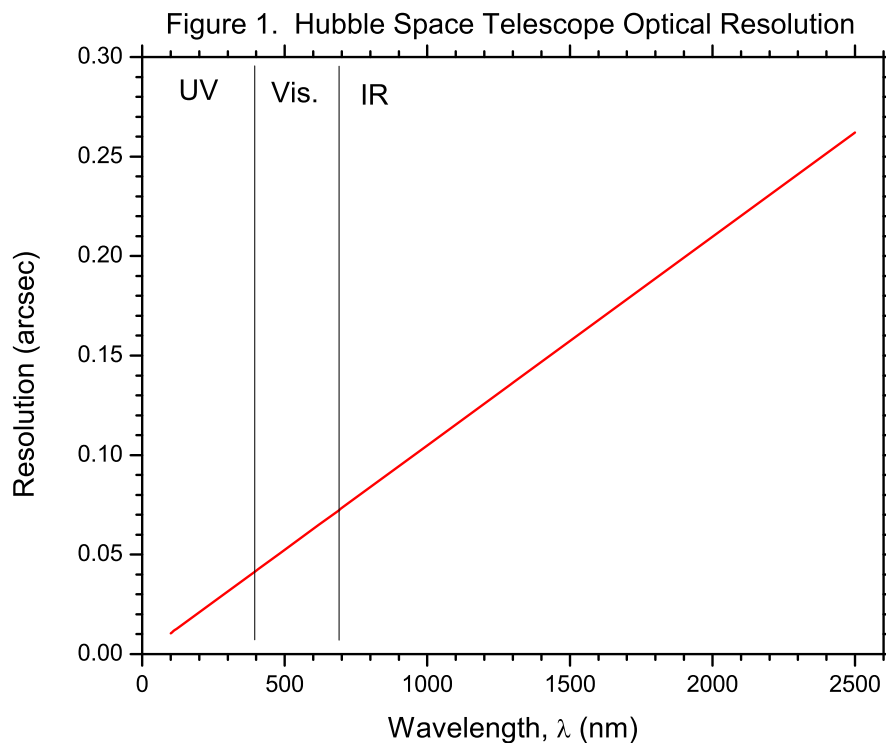


Figure 53.1: Resolution of the Hubble Space Telescope.

Rayleigh criterion, the smallest angular separation θ that two point sources can have and still be resolvable as two individual point sources is

$$\theta = 1.22 \frac{\lambda}{D} \quad (53.2)$$

where θ is the angular resolution in radians, λ is the wavelength of the light, and D is the diameter of the aperture of the instrument. For the Hubble Space Telescope, $D = 2.4$ meters, and λ varies between 100 and 2500 nanometers. Using equation (2) for the Rayleigh criterion, we can plot the angular resolution of Hubble as a function of wavelength (Figure 1).

As you can see from the figure, Hubble's resolution in visible light is about 0.05 arcseconds (where 1 arcsecond = 1/60 arcminute = 1/3600 degree). To give some idea of what this means, if the Hubble Space Telescope were in Washington DC, it could distinguish two objects in New York City if they were separated by a distance of just 3 inches:

$$\begin{aligned}
 s &= r\theta \\
 &= (331321 \text{ m}) \left[(0.05 \text{ arcsec}) \left(\frac{1 \text{ deg}}{3600 \text{ arcsec}} \right) \left(\frac{\pi \text{ rad}}{180 \text{ deg}} \right) \right] \\
 &= 0.080 \text{ m} \\
 &= 3 \text{ inches}
 \end{aligned}$$

53.4 Spherical Aberration

Shortly after its launch in 1990, it was discovered that Hubble's primary mirror had a *spherical aberration*, in the sense that it had not been ground exactly to the required hyperbolic shape. It turned out that the outer edges had been made too flat by about $2\text{ }\mu\text{m}$ —about 1/50 the thickness of a human hair, but enough to severely degrade the images. Light striking the primary mirror near the edges focused at a different point than light striking the mirror near the center, resulting in a significant blurring of the images.

Some mathematical techniques were developed to partially compensate for this, but the real issue was that the optics needed to be fixed. This was done during the Hubble First Servicing Mission in 1993, when a set of corrective optics called COSTAR (for “Corrective Optics Space Telescope Axial Replacement”) was installed. COSTAR consisted of a set of mirrors (one for each instrument) that were curved in such a way that they corrected for the spherical aberration in the primary mirror. The light path then became one where light would first strike the primary mirror, then reflect off of the secondary mirror, then down through the hole in the primary mirror where it would strike a COSTAR corrective mirror, then on to the instruments. Since the installation of COSTAR, the Hubble Space Telescope has operated right at the theoretical limit of resolution imposed by single-slit diffraction effects.

Since the First Servicing Mission, all new Hubble instruments that have been installed have included their own built-in corrective optics. By the time of Servicing Mission 3B in 2002, all the instruments had their own corrective optics built in, and COSTAR was no longer required. COSTAR was finally removed during Servicing Mission 4 in 2009, freeing up room for another scientific instrument to be installed during this mission, the Cosmic Origins Spectrograph.

Chapter 54

Dispersion

Recall that the index of refraction n of a transparent material is the ratio of the speed of light in vacuum to the speed of light in the material: $n = c/v$. In general, the index of refraction will vary somewhat with wavelength; this phenomenon is called *dispersion*. Dispersion can be an unwanted effect in lenses, since it causes chromatic aberration. But it can also be useful in prisms, in that it allows “white” light (light of all wavelengths) to be separated into its component colors. The same phenomenon occurs in Nature, where the dispersion properties of water allows sunlight to be separated into its component colors by water droplets, resulting in a *rainbow*.

For example, the

54.1 Cauchy Dispersion Formula

One simple model for dispersion in materials is the *Cauchy dispersion formula*:

$$n(\lambda) = a_0 + \frac{a_1}{\lambda^2} + \frac{a_2}{\lambda^4} + \frac{a_3}{\lambda^6} + \cdots \quad (54.1)$$

One often uses just the first two terms of the Cauchy dispersion formula:

$$n(\lambda) = a_0 + \frac{a_1}{\lambda^2}, \quad (54.2)$$

where λ is the wavelength, and the constants a_n depend on the material.

For water (20°C), $a_0 = 1.31494$, $a_1 = 4537.99465 \text{ nm}^2$.

Three terms:

$$n(\lambda) = a_0 + \frac{a_1}{\lambda^2} + \frac{a_2}{\lambda^4}, \quad (54.3)$$

where λ is the wavelength, and the constants a_n depend on the material.

For water (20°C), $a_0 = 1.32692$, $a_1 = 1610.845 \text{ nm}^2$, $a_2 = 95402300 \text{ nm}^4$.

54.2 Sellmeier Dispersion Formula

A more complex dispersion model is called the *Sellmeier dispersion formula*:

$$n(\lambda) = \left(1 + \frac{B_1\lambda^2}{\lambda^2 - C_1} + \frac{B_2\lambda^2}{\lambda^2 - C_2} + \frac{B_3\lambda^2}{\lambda^2 - C_3} \right)^{1/2}, \quad (54.4)$$

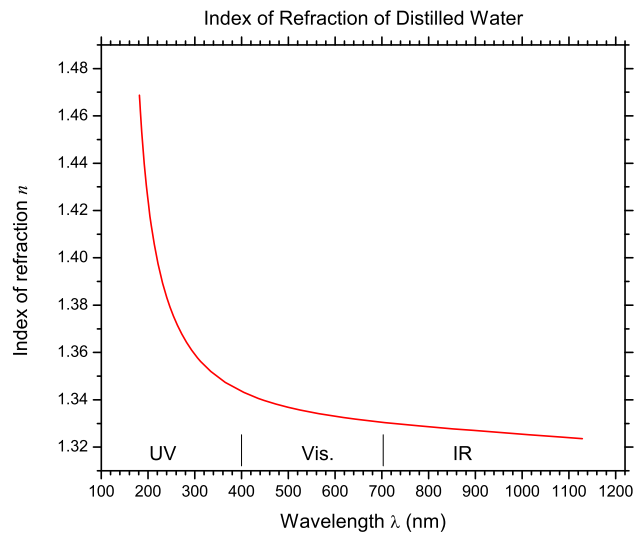


Figure 54.1: Dispersion in distilled water, showing how the index of refraction varies with wavelength.

where λ is the wavelength and constants B_1 , C_1 , B_2 , C_2 , B_3 , and C_3 depend on the material.

For water (20°C), $B_1 = 0.35074$, $C_1 = 8725.74686 \text{ nm}^2$, $B_2 = 0.04212$, $C_2 = 25765.38176 \text{ nm}^2$, $B_3 = 0.36067$, and $C_3 = 8739.18542 \text{ nm}^2$.

Chapter 55

Polarization

In normal white light, the plane of the electric field vector occurs in random directions for different wave trains; such light is said to be *unpolarized*. In *polarized* light, the electric field vector for all waves is in the same plane.

Light may be polarized by several different methods:

1. Selective absorption
2. Reflection
3. Scattering
4. Birefringence

55.1 Selective Absorption

In *selective absorption*, unpolarized light is passed through a material called a *polarizer*. The polarizing material has polymers embedded in it that absorb light whose electric vector is parallel to the polymers. The light that passes through the polarizer has its electric vector in one plane only: the plane perpendicular to the polymer direction. The direction of the plane of polarization (the plane of the electric vector) is called the *axis of polarization* of the polarizer.

If light passing through a polarizer is passed through a second piece of polarizing material (sometimes called an *analyzer*), the amount of light leaving the analyzer depends on the angle between the polarization axes of the polarizer and analyzer. If the polarization axes are in the same direction, all of the light leaving the polarizer passes through the analyzer. If the polarization axes are at right angles, the analyzer blocks all the light from the polarizer, and no light goes through. In general, if the polarizer and analyzer are at an angle θ with respect to each other, the intensity I of light leaving the analyzer is given by *Malus's law*:

$$I = I_0 \cos^2 \theta \quad (55.1)$$

where I_0 is the intensity of light leaving the polarizer, before it goes through the analyzer.

The intensity of *unpolarized* light is cut in half after passing through a single polarizer.

55.2 Reflection; Brewster's Law

Light may also be (partially) polarized by reflection from a reflecting surface (a linoleum floor or a glass window, for example). In this case, light will be polarized in a direction perpendicular to the plane of incidence:

light reflecting from a reflecting floor or swimming pool will be horizontally polarized, and light reflecting from a window will be vertically polarized.

Reflected light will, in general, be only *partially* polarized. At one particular angle of incidence, though, the reflected light will be not just partially polarized, it will be *completely* polarized. That incidence angle is called the *polarization angle*, and is given by *Brewster's law*: Brewster's law

$$\tan \theta_p = \frac{n}{n_{\text{air}}} \quad (55.2)$$

Here θ_p is the polarization angle, n is the index of refraction of the reflecting material, and $n_{\text{air}} = 1$ is the index of refraction of air.

Since light reflecting from a horizontal surface light a swimming pool will be at least partially horizontally polarized, polarizing sunglasses are designed to have their polarization axis in the *vertical* direction to block the reflected light.

55.3 Scattering

Light may be polarized by *scattering* of light. This may be seen by observing a clear blue sky through polarizing sunglasses; by rotating the sunglasses you can see the sky getting brighter and darker, as the sunglasses's polarization direction changes with respect to the direction of polarized skylight.

55.4 Birefringence

Another method of polarization is *birefringence*. This notably occurs in the mineral Iceland spar, which is a transparent crystalline form of *calcite*, or calcium carbonate (CaCO_3). If Iceland spar is placed on top of a page of printed text, you will see the image of the text is doubled (i.e. there will be two images of each letter). Each image is polarized in a different direction, as you can verify by rotating a polarizing material in front of the Iceland spar.



Figure 55.1: Birefringence in a sample of Iceland spar. (Credit: Jo Edkins, <http://gwydir.demon.co.uk/jo/minerals/index.htm>)

Chapter 56

Color

The human eye is capable of both color and black-and-white vision. Under conditions of very low illumination, a set of very light-sensitive *rods* on the retina of the eye allow us to see in black and white. Under higher illumination, a different set of light receptors called *cones* become active that permit color vision. The retina contains three types of cones, each of which is mostly sensitive to a different color: red, green, and blue. Combinations of these three primary colors allow us to see all the other colors.

What we perceive as “white” light is actually a combination of all colors of light. We can split white light into its component colors using a prism or a diffraction grating; the resulting colors and their approximate wavelengths are shown in Table 56-1. (The sequence of colors can be remembered from the mnemonic ROY G. BIV.)

Table 56-1. Approximate wavelengths of colors in the spectrum.

Color	Wavelength (nm)
Red	650
Orange	590
Yellow	570
Green	510
Blue	475
Indigo	445
Violet	400

Our perception of color is a complicated process. It depends partly on the wavelength of light received by the eye; but also the brain is able to distinguish colors by comparing the brightness of an object to other nearby objects, as seen by all three colors of cones on the retina. This complicated process (called the *Land effect*) allows us to perceive objects to be the same color, even under very different lighting conditions. (Notice, for example, that objects appear to have the same color indoors under a fluorescent light as they do outdoors under sunlight.) This phenomenon is called *color constancy*.

56.1 Lights

There are three *primary colors* of light: *red*, *green*, and *blue*. Other colors of light can be made by combining these three primary colors in different proportions. Equal proportions of red and green light make *yellow*

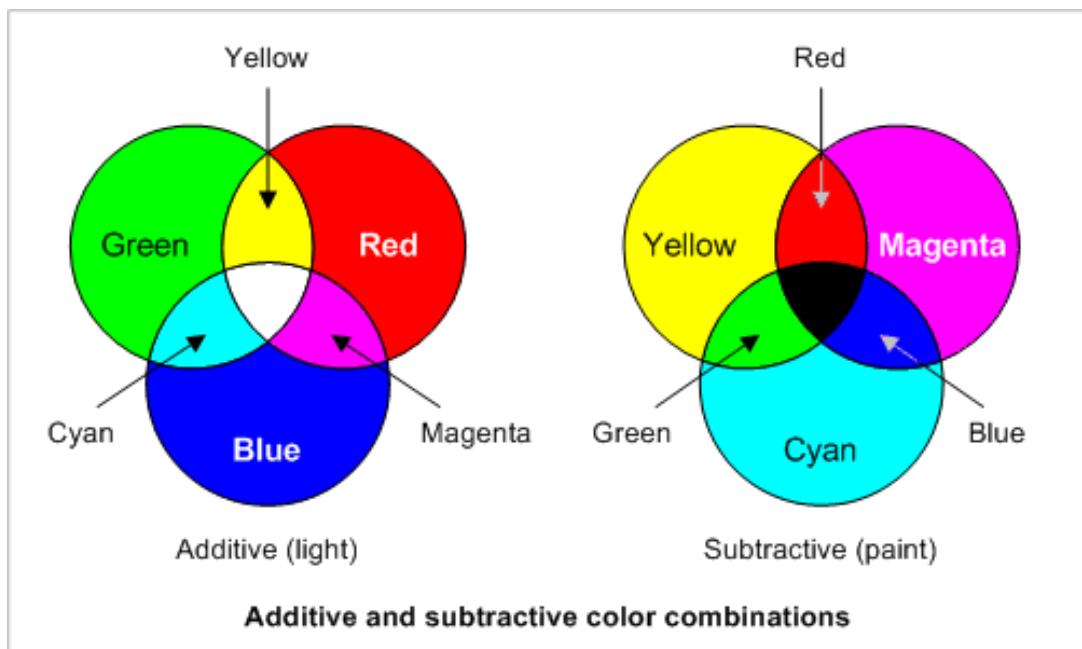


Figure 56.1: Addition and subtraction of primary colors.

light; equal proportions of green and blue light make a greenish-blue color called *cyan*; and equal proportions of red and blue light make a purplish color called *magenta*. All three primary colors combined in equal proportions make *white* light (Fig. 56.1).

If one of the primary colors is removed from white light, the remaining colors combine to form a *secondary color* that is said to be the *complement* of the missing color. For example, removing the blue light component from white light leaves yellow light, so yellow is said to be the complement of blue—in a sense, blue is “anti-yellow” and yellow is “anti-blue”. Similarly, the complement of green is magenta, and the complement of red is cyan (Table 56-2).

In summary, *for lights*, the primary colors are red, green, and blue; the secondary colors are cyan, magenta, and yellow.

Table 56-2. Complementary colors.

Color	Complement
red	cyan
green	magenta
blue	yellow

These properties of light colors are used in devices we encounter every day. For example, if you enlarge a color television screen or computer monitor with a magnifying glass, you will see that the image is made up of an array of small, adjacent red, green, and blue pixels which are combined in different proportions to form different colors. For example, if you enlarge a part of the screen that contains a yellow image, you will

see the red and green pixels turned on, and the blue pixels turned off. To form orange, the red pixels will be on and bright, the green pixels on but dim, and the blue pixels will be off.

56.2 Pigments

Colored *pigments* like paints and inks work differently from lights. If a colored pigment is illuminated with white light, it will absorb some colors and reflect others; the combination of the reflected light colors determines the color of the pigment.

Students of art are often taught that the primary colors for paints are red, yellow, and blue, but this isn't quite right. For pigments like paints, the primary colors are the same as the *secondary* colors of light: magenta, yellow, and cyan. Magenta paint absorbs its complementary color (green) and reflects red and blue light; yellow paint absorbs its complement (blue) and reflects red and green light; and cyan paint absorbs its complement (red) and reflects green and blue light (Table 56-3).

Table 56-3. Pigment colors.

Pigment	Made by combining	Absorbs	Reflects
red	magenta & yellow	green & blue	red
green	cyan & yellow	red & blue	green
blue	cyan & magenta	red & green	blue
cyan	cyan	red	green & blue
magenta	magenta	green	red & blue
yellow	yellow	blue	red & green
white	none	none	all
black	all	all	none

Likewise, the *secondary* colors for pigments are the same as the *primary* colors for lights: red, green, and blue. In each case, a pigment of one of the secondary colors reflects only that color, and absorbs the others.

The color that results by mixing pigments can generally be predicted¹ by assuming that the mixture will absorb the colors of its components, and reflect everything else. For example, what happens if we mix cyan and yellow paint? The cyan pigment absorbs red, the yellow pigment absorbs blue, and so the mixture should absorb both red and blue, and reflect green; thus cyan and yellow pigments mixed together make green.

Table 56-3 shows the colors resulting by combining colors in *equal amounts*. Other colors can be created by combining pigments in ways that reflect the primary light colors in unequal amounts. For example, suppose we combine red and yellow pigments. The red component of the mixture will absorb green and blue light, while the yellow component will absorb blue light. The mixture will then absorb some green light, and lots of blue light — resulting in the reflection of lots of red light and some green light, and an *orange* color.

56.3 Spectral Colors

If white light is split into its component colors (a *spectrum*) using a prism or diffraction grating, we observe the colors listed in Table 56-1, called the *spectral colors*. The spectrum includes the primary colors of light (red, green, and blue), along with the secondary colors yellow and cyan (located between green and blue).

¹ Assuming the pigments do not react chemically.

But the spectrum does *not* include the color magenta, which is a combination of two colors on opposite ends of the spectrum (red and blue). Magenta is an example of a class of colors called *purples*, that are formed by combining blue/violet with red in different proportions. Purples are *not* spectral colors, and do not appear in the spectrum of white light.

There is an important distinction between *purple* and *violet*. Purple is a *non*-spectral color formed by combining blue/violet light with red light. *Violet*, however, is a spectral color, and appears at the short-wavelength end of the spectrum.

56.4 The Chromaticity Diagram

Figure 56.2 shows the *CIE chromaticity diagram*.² It is a figure upon which may be plotted every color visible to the human eye. Its unusual shape is because of the way it is defined; see Appendix S for details.

The curved, horseshoe-shaped edge of the chromaticity diagram is where the pure spectral colors lie. Colors along this edge are the brightest and most vivid that we see them. Moving from the edge toward the center of the figure, the colors become more and more washed-out, finally becoming white at point $E = (0.3333, 0.3333)$, the equal-energy point.

The straight line from $(x, y) = (0.17, 0.00)$ to $(0.73, 0.26)$ is called the *line of purples*. The non-spectral colors (magenta and other purples) lie along this line.

Whether you're looking at Figure 56.2 on a color monitor or on paper (printed from a color printer), you're not *really* seeing the diagram the way it actually looks. That's because both color monitors and color printers are limited in the range of colors they can display. The white triangle in Fig. 56.2 shows the range of colors visible on a typical color monitor. If you look at the figure on a color monitor, you'll notice the colors look relatively constant moving along a line the white triangle to the curved edge; this is because of limitations in the color monitor.

Figure 56.3 illustrates some properties of the chromaticity diagram. Fig. 56.3(a) shows that if you connect any two points (colors) A and B with a straight line, then all points along the line represent colors that can be formed by combining colors A and B in different proportions. Another property is illustrated by Fig. 56.3(b): choose any color point on the edge of the diagram, and draw a straight line from that point, through the center point E , to the edge of the opposite side of the figure. This point at the opposite edge of the figure is the *complement* of the original point.

²CIE is the International Commission on Illumination; its initials are an abbreviation for its French name, Commission Internationale de l'Éclairage.

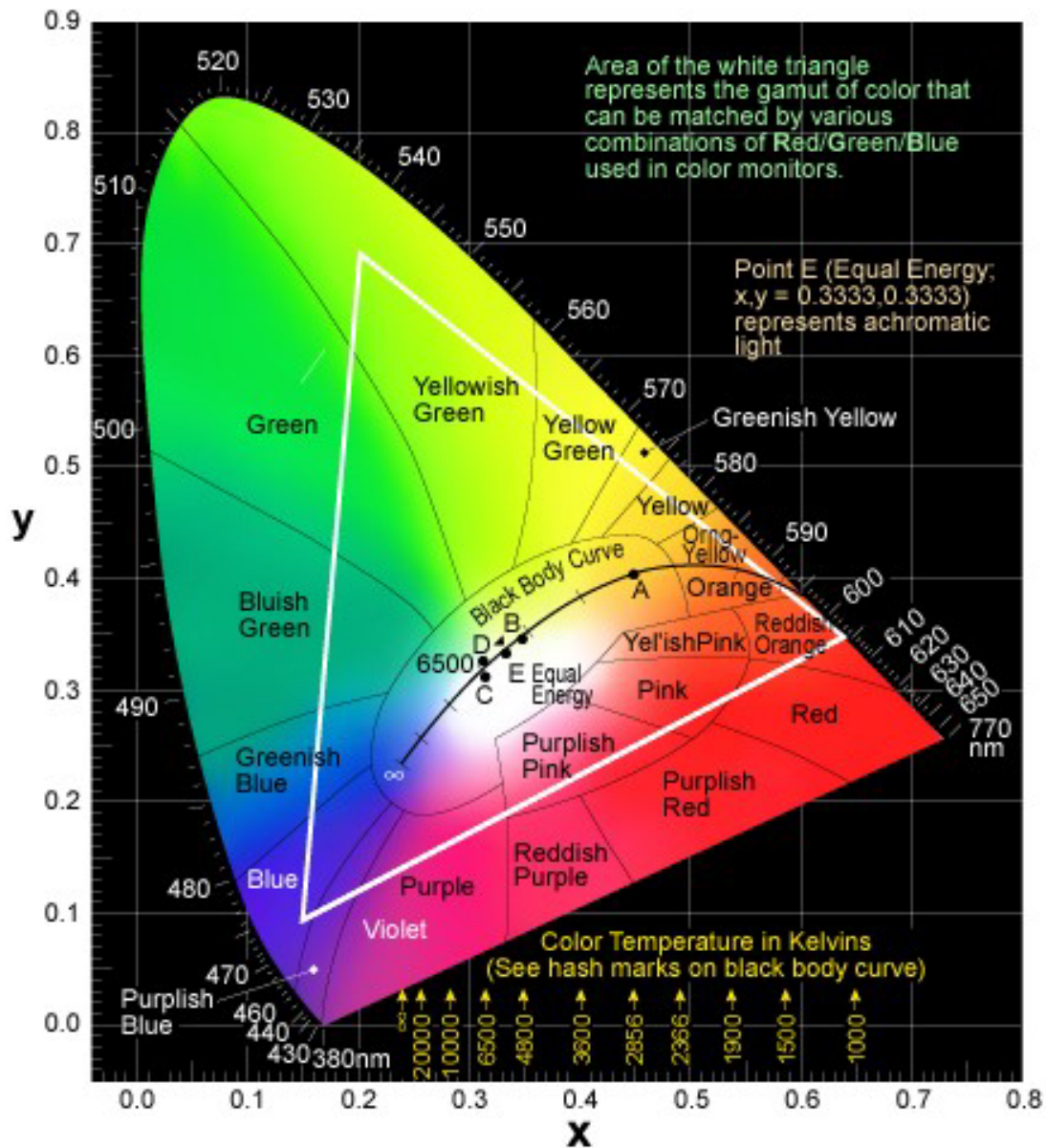


Figure 56.2: The CIE 1931 chromaticity diagram. The white triangle shows the range of colors that can be displayed on a color computer monitor. The curved line in the middle shows the color of blackbody radiation at various temperatures. Points A, B, C, and D are standard light sources (A: tungsten, 2856 K; B: Illuminant B; C: Illuminant C; D: D65, 6500 K.) Point E is the equal-energy point.

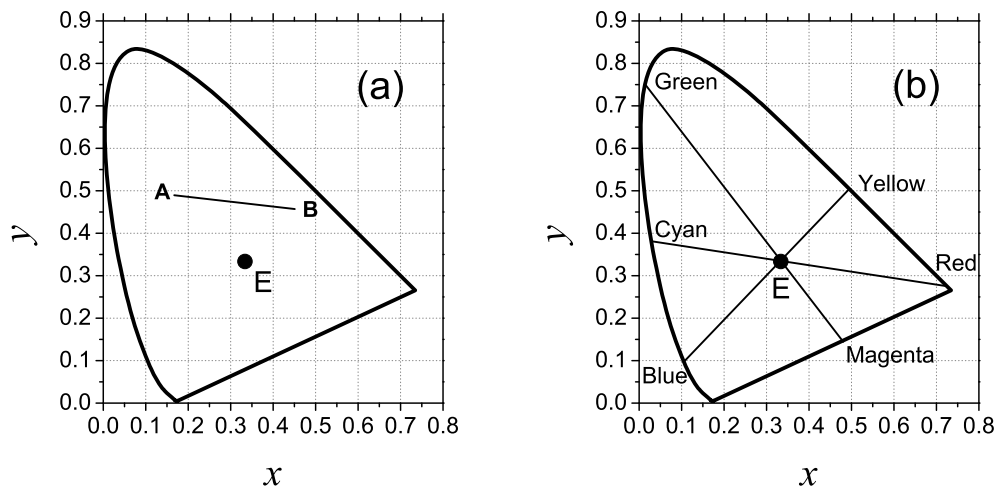


Figure 56.3: Some properties of the chromaticity diagram. (a) If A and B are two points (colors) on the diagram, then any color along the line connecting A and B can be formed by combining A and B in different proportions. (b) Complementary colors lie on opposite sides of a line passing through the equal-energy point E .

Chapter 57

The Rainbow

The *rainbow*, one of the most beautiful and striking objects seen in Nature, is typically visible during a late afternoon rain shower, when the Sun is low in the sky and shining at the same time (Fig. 57.1). Understanding all of the features of the rainbow requires many different ideas from optics.

57.1 Colors

The most obvious feature of the rainbow is its selection of colors. The phenomenon responsible for the colors is *dispersion* (Chapter 54). White light from the Sun enters each raindrop, refracts into the interior of the drop, reflects once via total internal reflection, and refracts back out of the drop again. The angles of refraction are determined by Snell's law and the index of refraction. But because of dispersion, the index of refraction (and therefore the angle of refraction) is different for each color of light (Table 57-1).

Table 57-1. Indices of refraction of water for different colors.

Color	Wavelength (nm)	n_{water}
Red	650	1.3317
Orange	590	1.3333
Yellow	570	1.3340
Green	510	1.3364
Blue	475	1.3381
Indigo	445	1.3400
Violet	400	1.3436

When you see a rainbow in Nature, you can often see *two* bows: a bright *primary rainbow*, and above it a fainter *secondary rainbow* (Fig. 49.1). The secondary bow is due to light reflecting a second time inside the raindrop due to total internal reflection.

57.2 The Primary Rainbow

In the brighter primary rainbow, red appears on the outside edge and violet on the inside edge.¹ The primary rainbow is due to light reflecting a single time inside the raindrops due to total internal reflection.

¹The next time you see a drawing of a rainbow, check to see whether the artist put the colors in the correct order, with red on the outside edge. Drawings have the colors wrong as often as not.



Figure 57.1: A rainbow. The bright primary bow is over the barn; the dimmer secondary bow is to the right. Alexander's dark band is the dark region between the two bows. (Credit: Pennsylvania State University.)

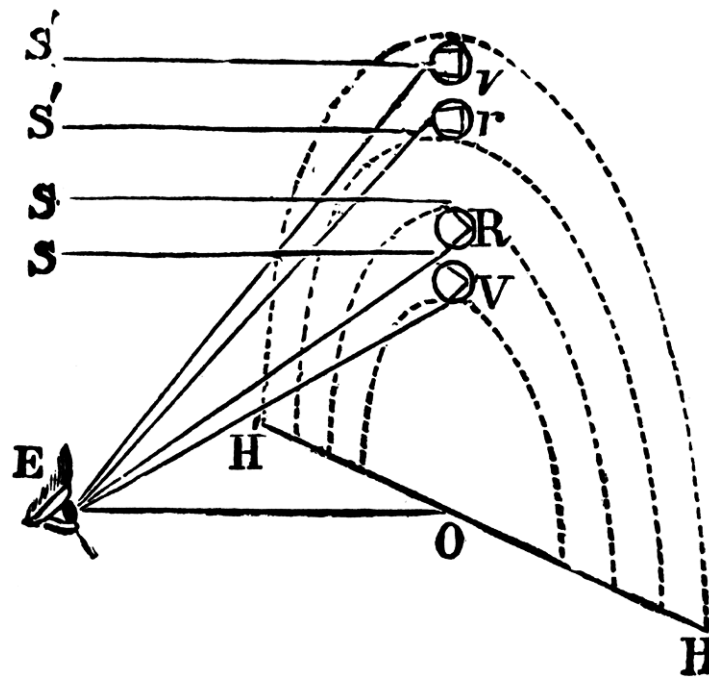


Figure 57.2: Observation of primary and secondary rainbows by an observer at point E . The Sun is behind the observer, as indicated by the lines S and S' . R and V show the locations of the red and violet bands (respectively) in the primary bow, while v and r show the locations of violet and red in the secondary bow. (Ref. [5])

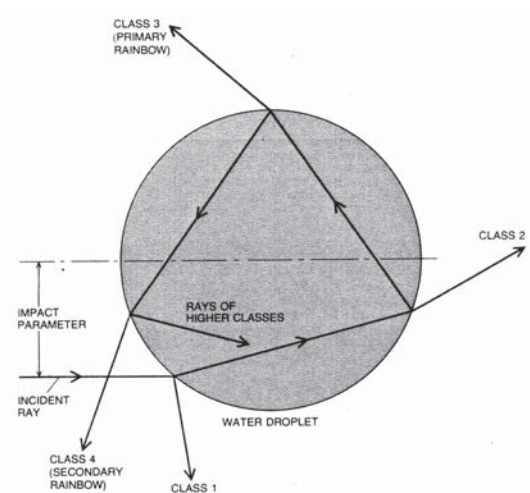


Figure 57.3: Impact parameter and the resulting scattered light rays. (From Nussenzveig, *Scientific American*, April 1977.)

The rainbow is in the shape of a partial circle, being a larger part of the circle the lower the Sun is in the sky. The rainbow would, in fact, be a complete circle if the ground weren't in the way; such complete rainbows may sometimes be seen from airplanes. The center of the rainbow's circle is directly opposite the direction of the Sun in the sky, so if the Sun is setting in the west and it's raining, look for a rainbow in the east.

57.3 The Secondary Rainbow

A fainter secondary rainbow appears above (outside) the primary rainbow, and its colors are reversed (violet on the outside edge and red on the inside edge). It is also a bit wider than the primary bow. The secondary bow is due to light reflecting *twice* inside the raindrop due to total internal reflection. Some light is lost during each reflection, so the secondary bow will be fainter than the primary bow.

57.4 Location of the Rainbow

What determines the location of the rainbow in the sky? The center of curvature of the rainbow is opposite the direction of the Sun, but what determines the angle from the sunline to the rainbow? By convention, we measure the angle between the Sun and the rainbow, as seen by the observer; this is called the *rainbow angle*. The primary rainbow has a rainbow angle of about 138° , while for the secondary bow the rainbow angle is 130° .

What determines these angles? Figure 57.3 shows the path of a light ray through a single (spherical) raindrop. The perpendicular distance between the light ray and the center of the drop is called the *impact parameter*, as shown in the figure. Of course, light rays are hitting the many raindrops at all different impact parameters, so the outgoing light rays are scattered over a range of angles. But from the principles of geometrical optics, we can calculate the angle of the outgoing light ray (the *scattering angle* as a function of impact parameter (Fig. 57.4). In the figure, you can see that the curve for the primary bow (upper curve) has a minimum for an impact parameter that is about 0.86 times the drop radius. Around this impact parameter, significant changes in the impact parameter result in nearly the same scattering angle — in essence, many

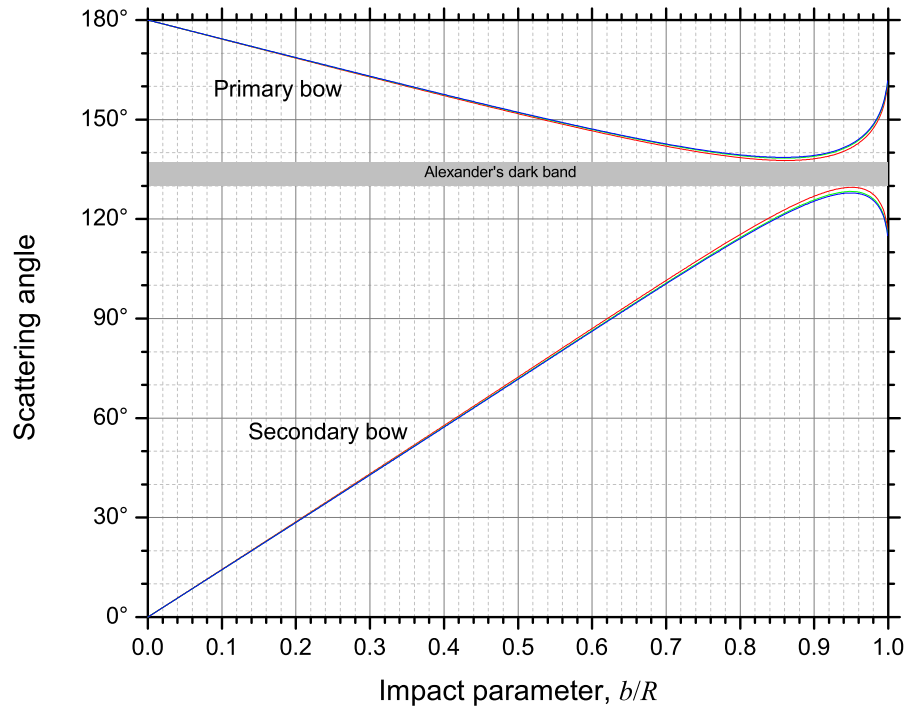


Figure 57.4: Scattering angle vs. impact parameter for the primary and secondary rainbows. (After Nussenzweig, *Scientific American*, April 1977.)

light rays hitting the drop at around this impact parameter will be scattered in the same direction, and this is where the rainbow will appear. According to calculations, light rays hitting the drop with an impact parameter of 0.86 of the drop radius will have a scattering angle of 138° , which is the rainbow angle for the primary bow.

Similarly, the curve for the secondary bow (two internal reflections) has a maximum at about 0.95 the radius of the drop. Therefore many light rays hitting the drop with an impact parameter around 0.95 of the drop radius will scatter at about the same angle, which is rainbow angle of the secondary bow, 130° .

57.5 Alexander's Dark Band

As seen in Figure 49.4, there is no impact parameter for either the primary or secondary bow that will lead to light being scattered between 130° and 138° . This results in a dark band between the primary and secondary bows, known as *Alexander's dark band* (Fig. 57.1).

57.6 Higher-Order Rainbows

Both the primary and secondary rainbows are easy to observe in Nature, but what about high-order bows, corresponding to three or more reflections of light rays inside each raindrop?

Confirmed observations of third- and fourth-order bows in Nature have only very recently been made for the first time, in 2011.² (See Figures 57.6 and 57.7.) There is at this time also evidence for observation of a fifth-order rainbow in 2015.³

It can be shown (Ref. [16]) that the rainbow angle of the k -th order rainbow (corresponding to k internal reflections in each drop) is given by

$$\theta_k = k(180^\circ) + 2\theta_{ik} - 2(k+1)\theta_{rk} \quad (57.1)$$

where the k -th angle of incidence θ_{ik} is given by

$$\theta_{ik} = \cos^{-1} \sqrt{\frac{n_w^2 - 1}{k(k+2)}} \quad (57.2)$$

and the k -th angle of refraction θ_{rk} is found from Snell's law:

$$\theta_{rk} = \sin^{-1} \left(\frac{1}{n_w} \sin \theta_{ik} \right) \quad (57.3)$$

Here n_w is the index of refraction of water. Since n_w varies depending on the color of light, these equations can be used to find the rainbow angle for both red and violet light, and from that deduce the width of each bow. The results of these calculations through the 20-th order rainbow are shown in Table 57-1, and illustrated in Figure 57.5. Notice that as the rainbow order k increases, the rainbows get both fainter and wider.

²See *Applied Optics*, **50**, 28, pp. F129-F141 (2011).

³See Edens, H.E. (2015) Photographic observation of a natural fifth-order rainbow, *Appl. Opt.* **54**, B26-34.

Table 57-1. The first 20 orders of rainbows of water, calculated from geometrical optics. This table shows the rainbow angles θ_k and bow widths $\Delta\theta$. Also shown are which side of the drop the incident light rays hits (T=top, B=bottom) and the rainbow “parity” (N=“normal” parity, with red on the outside and violet on the inside; R=“reversed” parity, with violet on the outside and red on the inside). ¹ The 12th order bow is split at the horizon, with red rays incident on the bottom of the drops and violet rays on the top (see Fig. 57.5).

k	Rainbow angle θ_k		Width $\Delta\theta$	Drop	
	red	violet		Side	Parity
1	137.63°	139.35°	1.72°	T	N
2	129.63°	126.52°	3.11°	B	R
3	42.47°	38.11°	4.37°	B	N
4	42.76°	48.34°	5.58°	T	R
5	127.08°	133.86°	6.78°	T	N
6	149.10°	141.13°	7.96°	B	R
7	65.59°	56.45°	9.14°	B	N
8	17.71°	28.02°	10.32°	T	R
9	100.86°	112.35°	11.49°	T	N
10	176.08°	163.43°	12.65°	B	R
11	93.11°	79.29°	13.82°	B	N
12	10.19°	4.79°	14.99°	(Note 1)	(Note 1)
13	72.68°	88.83°	16.15°	T	R
14	155.51°	172.82°	17.32°	T	N
15	121.69°	103.21°	18.48°	B	R
16	38.91°	19.27°	19.64°	B	N
17	43.84°	64.65°	20.80°	T	R
18	126.58°	148.55°	21.97°	T	N
19	150.69°	127.57°	23.13°	B	R
20	67.98°	43.69°	24.29°	B	N

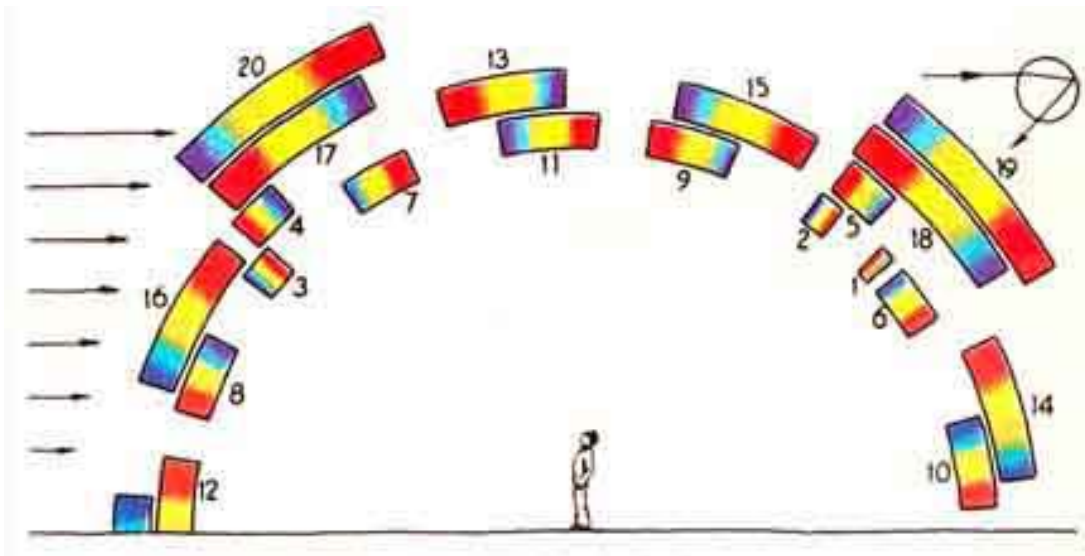


Figure 57.5: Locations of the first 20 orders of rainbows in the sky. Each order bow is dimmer and wider than the previous one. Only the 1st and 2nd order bows are visible in Nature, but the higher-order bows may be observed in laboratory experiments, as described in the *Amateur Scientist* column of *Scientific American*, July 1977.

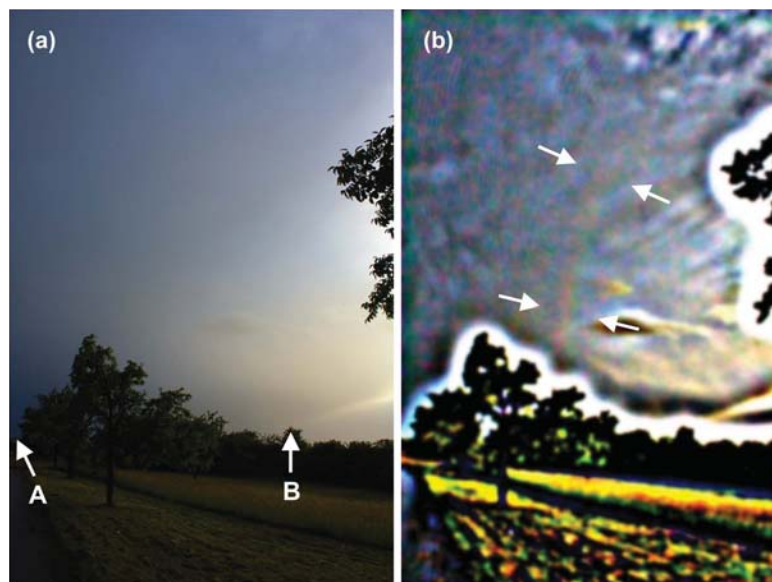


Figure 57.6: First ever photograph of a third-order (tertiary) rainbow, taken in southern Germany in 2011. (a) Original photograph. Points *A* and *B* are reference positions for image orientation. (b) Computer-enhanced version that shows the third-order bow. Arrows show the location of the rainbow image. The Sun is off to the right. (Großmann, Schmidt, and Haußmann, *Applied Optics*, **50**, 28, F134F141 (2011).)

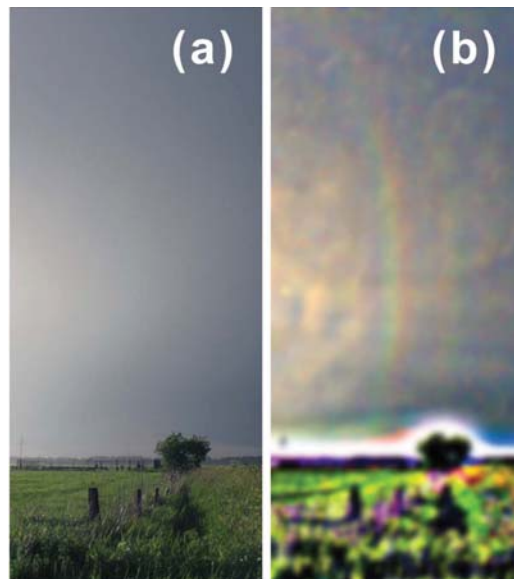


Figure 57.7: First ever photograph of a fourth-order (quaternary) rainbow, taken in northern Germany in 2011. (a) Original photograph. (b) Computer-enhanced version that shows both the third-order bow (left) and fourth-order bow (right). The Sun is off to the left. (Theusner, *Applied Optics*, **50**, 28, F129F133 (2011).)

Part VI

Modern Physics

Chapter 58

Special Relativity

58.1 Introduction

The classical mechanics described by Sir Isaac Newton begins to break down at very high velocities, i.e. at velocities near the speed of light $c = 299,792.458$ km/s. For bodies moving at a significant fraction of the speed of light, Newton's mechanics needs to be modified. The necessary modifications were developed by physicist Albert Einstein in the early 20th century.

58.2 Postulates

Einstein discovered that the necessary modifications to Newtonian mechanics could be derived by assuming two postulates:

1. Absolute uniform motion cannot be detected.
2. The speed of light is independent of the motion of the source.

The first postulate says that all motion is relative—that there is no reference frame that all observers can agree to be absolutely at rest. The second postulate says that light does not obey the usual laws of velocity addition. For example, if someone is moving toward you at 99% of the speed of light and turns on a flashlight in your direction, you will measure the light's speed to be the same as if that person were at rest.

Although these postulates seem quite reasonable, they lead to some surprising consequences. Let's examine a few of those consequences.

58.3 Time Dilation

It turns out that one consequence of Einstein's postulates is that time runs more slowly for someone moving relative to you; this effect is called *time dilation*. If someone is moving at speed v relative to you, then their clocks will run slower than yours. If a clock measures a time interval Δt_0 when it's at rest, then when it's moving at a speed v relative to you, you will measure that time interval to be longer by a factor γ :

$$\Delta t = \gamma \Delta t_0, \quad (58.1)$$

where Δt is the time interval measured by the moving clock, Δt_0 is the time interval measured on the clock when it's at rest, and γ is an abbreviation for the factor

$$\gamma \equiv \frac{1}{\sqrt{1 - v^2/c^2}}. \quad (58.2)$$

(Note that $\gamma \geq 1$.) The time interval Δt_0 , measured when you're at rest with respect to the clock, is called the *proper time*.

This effect means that time travel is possible—at least time travel into the future. One simply builds a spacecraft and travels close to the speed of light, then turns around and returns to Earth. (It is not clear whether time travel into the past is possible, but it might be possible under Einstein's *general* theory of relativity.)

58.4 Length Contraction

Another consequence of the postulates is that a moving body will appear to be shortened in the direction of motion; this effect is called *length contraction*. The length of a moving body will appear to be shortened by this same factor of γ :

$$L = \frac{L_0}{\gamma} \quad (58.3)$$

Here L_0 is the length of the body when it is at rest, and is called the *proper length*. Since $\gamma \geq 1$, the moving body will be shorter when it is moving.

58.5 An Example

As an example, let's imagine that a spacecraft is launched at high speed relative to Earth toward the nearest star, Alpha Centauri (which is about 4 light-years away). The ship travels at 80% of the speed of light during the trip. From Earth, we see that the whole trip takes 5 years. We also see the astronaut's clocks running more slowly than ours by a factor of $\gamma = 2.78$, so that when the astronauts arrive, they are only 1.8 years older.

What do the astronauts see from their point of view on the spacecraft? Their clocks run at what seems a normal rate for them, but they see that the *distance* to Alpha Centauri has been length-contracted by a factor of $\gamma = 2.78$. They're traveling at a speed of $0.80c$, but they only have to travel a distance of $(4 \text{ light-years})/\gamma = 1.44 \text{ light-years}$. When they arrive at Alpha Centauri, they're older by $(1.44 \text{ light-years})/0.80c = 1.8 \text{ years}$.

In summary, observers on Earth see the astronaut's clocks moving more slowly, but the astronauts have to travel the full 4 light-years. The astronauts see their clocks moving at normal speed, but the distance they have to travel is shorter. All observers agree that the astronauts are only 1.8 years older when they arrive.

58.6 Momentum

In Newton's classical mechanics, momentum is $\mathbf{p} = m\mathbf{v}$. Under special relativity, this is modified to be

$$\mathbf{p} = \gamma m\mathbf{v}. \quad (58.4)$$

Relativistically, it is this definition of momentum that is conserved. Newton's Second Law in the form $\mathbf{F} = m\mathbf{a}$ is no longer valid under special relativity, but Newton's original form $\mathbf{F} = d\mathbf{p}/dt$ is still valid, using this definition of momentum $\mathbf{p}$.

Notice that as $v \rightarrow c$, we have $\gamma \rightarrow \infty$ (by Eq. (58.2)), and so momentum $p \rightarrow \infty$. As a body goes faster, its momentum increases in such a way that it becomes increasingly difficult to make it go even faster. This means that it is not possible for a body to move faster than the speed of light in vacuum, c .

58.7 Addition of Velocities

Let's suppose that we have two bodies moving in one dimension. The first is moving at speed u , and the second is moving at speed v . What is the speed of the second relative to the first? In other words, what will you measure as the speed of the second body if you're sitting on the first body?

In classical Newtonian mechanics, the speed w of the second body relative to the first is simply

$$w = v - u. \quad (58.5)$$

For example, if the first body is moving to the right with speed $u = 10$ m/s, and the second body is moving toward it to the left with speed $v = -20$ m/s, then an observer on the first body will see the second body moving toward it with a speed of $w = 30$ m/s.

In the special theory of relativity, this seemingly self-evident equation for adding velocities must be modified as follows:

$$w = \frac{v - u}{1 - uv/c^2}. \quad (58.6)$$

This reduces to Eq. (58.5) unless the speeds involved are near the speed of light. For the above example, where $u = 10$ m/s and $v = -20$ m/s, Eq. (58.6) gives $w = 29.9999999999993324$ m/s, rather than $w = 30$ m/s given by Eq. (58.5). As you can see, for many applications, the difference between the classical formula (58.5) and the exact relativistic formula (58.6) is not enough to justify the extra complexity of using the relativistic formula.

But for speeds near the speed of light, using the relativistic formula is important. For example, if $u = 0.99c$ and $v = -0.99c$, then the classical formula of Eq. (58.5) would give $w = 1.98c > c$, in violation of special relativity; but using the exact expression in Eq. (58.6) gives the correct answer, $w = 0.9999494975c$.

Eq. (58.6) makes it impossible for the relative speeds to be greater than the speed of light c . In the extreme case $u = c$ and $v = -c$, Eq. (58.6) gives $w = c$, in agreement with the Einstein's second postulate.

58.8 Energy

Rest Energy

Einstein showed that mass is a form of energy, as shown by his most famous equation,

$$E_0 = mc^2. \quad (58.7)$$

E_0 is called the *rest energy* of the particle of mass m . The clearest illustration of this formula is the mutual annihilation of matter and *antimatter* (a kind of mirror-image of ordinary matter). When a particle of matter collides with a particle of antimatter, the mass of the two particles is converted completely to energy, the amount of energy liberated being given by Eq. (58.7).

As examples, the rest energy of the electron is 511 keV, and the rest energy of the proton is 938 MeV. (1 eV is one *electron volt*, and is equal to $1.6021766208 \times 10^{-19}$ J.)

Kinetic Energy

In classical Newtonian mechanics, the kinetic energy is given by $K = mv^2/2$. The relativistic version of this equation is

$$K = (\gamma - 1)mc^2. \quad (58.8)$$

It is not obvious that this reduces to the classical expression until we expand γ into a Taylor series:

$$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2} = 1 + \frac{1}{2} \frac{v^2}{c^2} + \frac{3}{8} \frac{v^4}{c^4} + \frac{5}{16} \frac{v^6}{c^6} + \frac{35}{128} \frac{v^8}{c^8} + \frac{63}{256} \frac{v^{10}}{c^{10}} + \frac{231}{1024} \frac{v^{12}}{c^{12}} + \dots \quad (58.9)$$

Substituting this series expansion for γ into Eq. (58.8), we get

$$K = \frac{1}{2}mv^2 + \frac{3}{8}m\frac{v^4}{c^2} + \frac{5}{16}m\frac{v^6}{c^4} + \frac{35}{128}m\frac{v^8}{c^6} + \frac{63}{256}m\frac{v^{10}}{c^8} + \frac{231}{1024}m\frac{v^{12}}{c^{10}} + \dots \quad (58.10)$$

Unless the speed v is near the speed of light c , all but the first term on the right will be very small and can be neglected, leaving the classical equation.

Total Energy

If the only forms of energy present are the rest energy E_0 and the kinetic energy K , then the total energy E will be the sum of these:

$$E = E_0 + K = \gamma mc^2. \quad (58.11)$$

It is often useful to know the total energy of a particle in terms of its momentum p rather than its velocity v . It can be shown that the total energy is given in terms of momentum by

$$E^2 = (pc)^2 + (mc^2)^2. \quad (58.12)$$

In the case where the total energy is much larger than the rest energy ($E \gg E_0$), we may neglect the second term on the right, and use

$$E \approx pc. \quad (58.13)$$

Chapter 59

Superfluids

When liquid helium-4 (^4He) is cooled below a critical temperature of 2.17 K (called the *lambda point*), a sudden phase transition occurs, and the helium becomes an exotic fluid called *helium II*.¹ Helium II is the best-known example of a *superfluid*—a fluid with odd properties that are governed by the laws of quantum mechanics.

As helium I is cooled toward the lambda point, it boils violently; but when the lambda point is reached, the boiling suddenly stops. This is due to a sudden increase in the thermal conductivity of the liquid when it transitions to the superfluid state. The thermal conductivity of superfluid helium II is more than a million times greater than that of liquid helium I, and helium II is a better conductor of heat than any metal.

Superfluid helium II is perhaps best known for its unusual viscosity. One method for measuring the viscosity of a liquid is to allow it to flow through a thin tube or channel called a *capillary*: the more viscous the liquid, the larger the diameter of the capillary needed to permit the liquid to flow. Helium II can flow through capillaries much less than $1\text{ }\mu\text{m}$ in diameter, and in such experiments behaves as though it has *zero* viscosity. This ability of helium II to flow through very tiny capillaries is called *superflow*.

Another method for measuring viscosity is to rotate a small cylinder inside the liquid; viscosity will cause the liquid to be dragged along with the cylinder, and a small rotatable paddle placed near the axis of the rotating cylinder will show whether the rotating cylinder is causing the liquid to rotate. In such experiments, helium II *does* exhibit some viscosity. No ordinary liquid exhibits this sort of dual behavior with respect to viscosity.

A common model to explaining this odd behavior is called the *two-fluid model*. In this model, liquid helium II is thought of as consisting of two interpenetrating components: a *normal* (viscous) component, and a *superfluid* (nonviscous) component. In the capillary experiment, only the superfluid component flows through the tiny capillaries, but in the rotating-cylinder experiment, the normal component is dragged along with the cylinder, causing circulation in the liquid.

Another unusual phenomenon observed in helium II is called the *fountain effect* (Fig. 59.1). A tube with a porous plug in the bottom is placed inside a bath of helium II. A superflow of helium is observed to flow through the tiny ($\ll 1\text{ }\mu\text{m}$) capillaries *toward* the heater; upon being heated, the superfluid component is converted to a normal component, and the fluid is unable to flow back out through the fine capillaries in the plug. Pressure builds in the tube until the helium squirts out of the capillary in the top of the tube, creating a “helium fountain”. Since the second law of thermodynamics states that heat cannot flow from lower to higher temperatures, this implies that the superfluid component carries no heat: any heat in the helium II must be in the normal component.

Yet another interesting property of helium II is the formation of a very thin film called a *Rollin film* when the liquid is placed in a container. The Rollin film will creep up the sides of the container, and if the container

¹ Above 2.17 K, liquid helium is a (mostly) ordinary liquid called *helium I*.

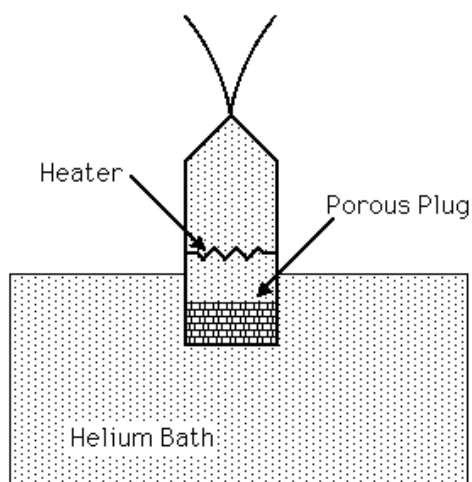


Figure 59.1: The fountain effect in superfluid liquid helium II. (Credit: NASA.)

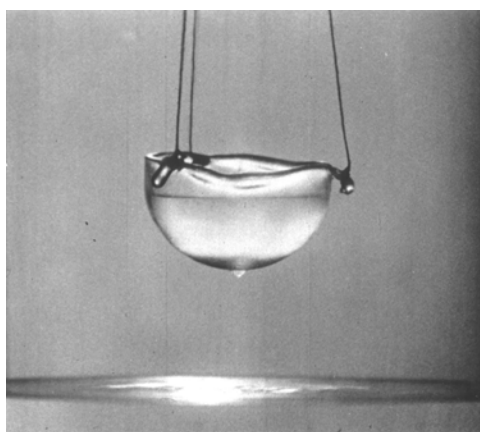


Figure 59.2: A Rollin film of helium II. The film creeps up the sides of the container and back down the outside, collecting in small drops at the bottom. (Credit: *Liquid Helium II: The Superfluid*, University of Michigan.)

is open, it will creep back down the outside, so that the helium II will spontaneously creep out of the container (Fig. 59.2). The Rollin film is much less than $1\text{ }\mu\text{m}$ in thickness; its creeping speed is slow just below the lambda point, but may reach a speed as high as 35 cm/s at lower temperatures.

Finally, helium II exhibits an unusual way of conducting heat. Normally, substances conduct heat through diffusion, where the rate of heat flow is proportional to the temperature difference; but in superfluid helium II, heat is conducted by *waves*. This phenomenon is called *second sound*, and no other substance exhibits this behavior. The speed of second sound is small just below the lambda point; at a lower temperature of 1.6 K, it is about 20 m/s.

It should be kept in mind that the two-fluid model of helium II discussed here is simply a *model*—a convenient way of thinking about the behavior of the liquid. Superfluid helium II is a quantum liquid, and a complete description of its behavior requires the application of quantum mechanics.

Chapter 60

The Standard Model

The *Standard Model* of particle physics is our current best theory of how the Universe is put together at its most fundamental level. It describes the fundamental nature of both matter and forces. This is still very much at the frontier of physics research, so it's not clear how much of our understanding of this is correct.

60.1 Matter

All of (ordinary) matter is found to be made of two types of particles: *quarks* and *leptons*. There are six types of quarks (called *up*, *down*, *charmed*, *strange*, *top*, and *bottom*) and six types of leptons (the *electron*, *muon*, *tau lepton*, and their associated *neutrinos*.) (Table 60-1.)

Table 60-1. The basic particles of matter.

Quarks	Leptons
Up (u)	Electron (e^-)
Down (d)	Electron neutrino (ν_e^0)
Charmed (c)	Muon (μ^-)
Strange (s)	Muon neutrino (ν_μ^0)
Top (t)	Tau lepton (τ^-)
Bottom (b)	Tau neutrino (ν_τ^0)

Quarks are never observed in isolation: they occur only as a system of three quarks (called a *baryon*), or as a quark-antiquark pair (called a *meson*). (An antiquark is a form of *antimatter*, described below.) Examples of baryons are the *proton* (which consists of two “up” quarks and one “down” quark) and the *neutron* (which consists of two “down” quarks and one “up” quark). Baryons and mesons together are collectively known as *hadrons*, so a hadron refers to a collection of bound quarks.

Quarks are held together in hadrons by a very strong force that becomes stronger the farther apart the quarks are separated. This is why they are not observed in isolation.

Leptons consist of the electron, the muon (which acts like a heavy electron), and the tau lepton (which acts like a very heavy electron). Each of these particles has a charge of $-e$. In reactions in which these particles are produced, there is generally also a neutrino particle. Neutrinos are very light particles with almost no mass, and for the most part they pass right through ordinary matter; in fact, there are billions of them passing through your body right now. Only very rarely do they interact with ordinary matter, but occasionally they do. Physicists have built neutrino “telescopes” to detect them; these telescopes consist of underground pools

filled with cleaning fluid surrounded by light detectors. In the rare event that a neutrino interacts with ordinary matter, it emits a brief flash of light which is detected and recorded.

Both quarks and leptons are, as far as we can observe, point masses. None of them has any internal structure that we're currently aware of.

60.2 Antimatter

Each quark and lepton has a corresponding mirror-image particle that has the same mass but opposite charge; such particles are called *antimatter*. The antimatter counterpart of the electron is called the *positron* (e^+); for other particles, you just add the prefix *anti-* (e.g. *anti-proton*, *anti-neutron*, etc.)

Whenever a particle of ordinary matter comes in contact with its antimatter counterpart, the two particles are destroyed and converted to energy in the form of gamma rays. The amount of energy created is given by Einstein's famous formula, $E_0 = mc^2$, where m is the sum of the particle masses and c is the speed of light in vacuum.

60.3 Forces

We know of four fundamental forces in Nature: the *gravitational force*, the *electromagnetic force*, and two *nuclear forces* (Table 60-2.) We're all familiar with the gravitational force (which is keeping you attached to the ground as you read this). Most of the other forces you encounter in everyday life are electromagnetic in nature. The strong nuclear force is responsible for holding atomic nuclei together against the mutual electrostatic repulsion of protons, and is also responsible for nuclear fusion reactions that occur in the Sun and in hydrogen bombs. The weak nuclear force is responsible for a process called β decay, in which a neutron in an atomic nucleus decays into a proton, electron, and anti-neutrino, and the electron escapes from the atom in the process.

Table 60-2. The four forces.

Force	Vector boson
Gravitational	Graviton (?)
Electromagnetic	Photon
Strong nuclear	Gluon
Weak nuclear	W, Z

According to the Standard Model, each of these forces is mediated by a particle called a *vector boson*. In effect, each force is thought to be caused by the exchange of these particles.¹

The electromagnetic and weak nuclear forces have been (somewhat) unified into a combined "electroweak theory", although this theory is not entirely complete. Many physicists believe that the electromagnetic, strong nuclear, and weak nuclear forces can be shown to be different aspects of a single underlying force, and thus all covered by a single "Grand Unified Theory". No Grand Unified Theory has yet been discovered.

Our best theory of gravity to date is Einstein's General Theory of Relativity, and has so far been shown to be consistent with experimental results. However, general relativity says that the gravitational force is due to the curvature of space-time; this is at odds with the Standard Model view, which is that gravity is caused by the exchange of particles called *gravitons*. No experiment has yet detected the existence of gravitons, and it's uncertain whether or not general relativity is the correct final theory of gravity.

¹The gravitational force is not considered to be part of the Standard Model.

Some physicists believe that it may be possible to show that *all four* forces (including gravity) are aspects of a single underlying force, and covered by a theory called the “Theory of Everything”. Such a theory (which is essentially a grand unified theory plus gravity) has not yet been found, nor is it known whether such a theory even exists. Some theories such as *string theory* have been proposed, but are far from being experimentally verified. These are issues to be worked out by future generations of physicists.

60.4 The Higgs Boson

A key piece of the Standard Model is *Higgs field*, which is responsible for giving particles their mass. The Higgs field fill all of space, even in places where there would otherwise be a vacuum. The degree to which a particle interacts with the Higgs field determines its mass: particles interacting weakly with the Higgs field are light, while those that interact strongly with the Higgs field are heavy. Particles that don't interact with the Higgs field at all, like the photon, are massless.

The Standard Model predicts that fields that fill all space should be associated with a particle — for example, as we've seen each of the four fundamental forces is associated with a vector boson particle.² The particle associated with the Higgs field is the *Higgs boson*. The Higgs boson was detected experimentally at the CERN particle physics accelerator³ in 2015, thus confirming the existence of the Higgs field and giving increased confidence in the Standard Model.⁴

²Except, perhaps, for gravity.

³CERN stands for Conseil Européen pour la Recherche Nucléaire, and is a facility located on the border between France and Switzerland.

⁴See http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/popular-physicsprize2013.pdf

Appendices

Appendix A

Further Reading

General

- *The Feynman Lectures on Physics* by R.P. Feynman, R.B. Leighton, and M.L. Sands (Addison-Wesley, 1963).

This collection of physics lectures was delivered by Nobel laureate Richard Feynman at the California Institute of Technology in the 1960s, and is known to every physicist. It is regarded by many as one of the best, clearest surveys of physics ever written. These lectures have recently re-released in a “New Millennium Edition”, and the audio recordings of the lectures have been released on CD as well.

- *Feynman’s Tips on Physics: A Problem-Solving Supplement to the Feynman Lectures on Physics* by R.P. Feynman (Addison-Wesley, 2005).

Supplementary material for the *Feynman Lectures on Physics*, in which Feynman gives his advice on strategies for solving physics problems.

Mathematics

- *How to Enjoy Calculus* by Eli S. Pine (Geyer Instructional AIDS Co., 1983).

The best introduction to the calculus, bar none. Also very brief (150 pages).

Waves (Part I)

- *Vibrations and Waves* by A.P. French (Norton, 1971).

One of the four volumes of the *MIT Introductory Physics Series*, this calculus-based book gives a fairly detailed presentation of vibrations and waves.

Acoustics (Part II)

- *The Physics of Sound* (3rd ed.) by R.E. Berg and D.G. Stork (Benjamin Cummings, 2004).

A text on acoustics and music for non-scientists, written by authors from the University of Maryland.

Music (Chapter 14)

- *Horns, Strings, and Harmony* by Arthur H. Benade (Dover, 1992).
A good survey of the physics of music and musical instruments.
- *Good Vibrations: The Physics of Music* by Barry Parker (Johns Hopkins, 2009).
A recent non-mathematical book on the physics of music.
- *Musical Acoustics* (3rd ed.) by Donald E. Hall (Brooks/Cole, 2002).
An undergraduate textbook on the physics of music.

Electricity and Magnetism (Part III)

- *Fundamentals of Electric Waves* by Hugh Hildreth Skilling (Krieger, 1948).
A brief, very clear book on electric waves (but requires a background in the calculus).
- *The Lightning Discharge* by Martin A. Uman (Dover, 2001).
A good book on the science of lightning by a well-known researcher.
- “A Bolt Out of the Blue” by Joseph R. Dwyer, *Scientific American*, May 2005.
A recent article on some of the latest developments in lightning research.

Electronics (Part III)

- *Getting Started in Electronics* by Forrest M. Mims III (Master Publishing, 2000).
This is a very brief (128 pp.), informal, hand-written (!) book on analog and digital electronics, aimed mainly at electronics hobbyists. Lots of good information on both theory and practical electronics, and easy to read.
- *Electronic Principles* (6th ed.) by Albert P. Malvino (Glencoe McGraw-Hill, 1999).
A standard, well-regarded undergraduate text on electronics, at roughly the level of this course.
- *The Art of Electronics* (3rd ed.) by Paul Horowitz and Winfield Hill (Cambridge, 2015).
An advanced book on analog and digital electronics, covering basically anything you would ever want to know about electronics. This book is widely regarded as a standard reference in the field. The book has a Web site at <http://www.artofelectronics.com/>.
- *Lessons in Electric Circuits* is a free electronic book, available on the Internet at: <http://www.allaboutcircuits.com/textbook/>. This book starts with the basics, yet covers a lot of material. The entire book is in six volumes, and is over 2700 pages long.
- *Bebop to the Boolean Boogie* (3rd ed.) by Clive “Max” Maxfield (Newnes, 2009).
An informal, easy-to-read introductory book on digital electronics.
- *Digital Fundamentals* (10th ed.) by Thomas L. Floyd (Pearson Prentice Hall, 2009).
A standard undergraduate text on digital electronics, at roughly the level of this course.

Radio (Chapter 43)

- *The Science of Radio: With MATLAB and Electronics Workbench Demonstrations* (2nd ed.) by Paul J. Nahin (Sprinter, 2001).
- *The Electronics of Radio* by David Rutledge (Cambridge, 1999).
- *The ARRL Handbook for Radio Communications* is published in a new edition each year by the American Radio Relay League, which is the association of radio amateurs in the United States: <http://www.arrl.org>.
- The Xtal Set Society (<http://www.midnightscience.com>) is a society devoted to crystal radios. They have a number of kits and publications, and issue a newsletter.

Optics (Part IV)

- *Optics* (4th ed.) by Eugene Hecht (Addison-Wesley, 2001).
A standard undergraduate text on optics.
- *Principles of Optics* (7th ed.) by Max Born and Emil Wolf (Cambridge, 1999).
An advanced, graduate-level book on optics.

Color (Chapter 56)

- *Light and Color in Nature and Art* by Samuel J. Williamson and Herman Z. Cummins (Wiley, 1983).
An excellent book on color—quite readable, yet contains a lot of technical information.
- *The Physics and Chemistry of Color* (2nd ed.) by Kurt Nassau (Wiley, 2001).
A good undergraduate text on color, somewhat more advanced than the Williamson and Cummins text.
- *Color Science: Concepts and Methods, Quantitative Data and Formulae* (2nd ed.) by G. Wyszecki and W.S. Stiles (Wiley, 2000).
A standard advanced text on color theory.

The Rainbow (Chapter 57)

- “The Theory of the Rainbow” by H. Moysés Nussenzveig, *Scientific American*, April 1977, 116–127.
- “The Amateur Scientist: How to Create and Observe a Dozen Rainbows in a Single Drop of Water” by Jearl Walker, *Scientific American*, July 1977.
- “Multiple rainbows from single drops of water and other liquids” by Jearl D. Walker, *Am. J. Phys.*, May 1976, 421–433.
- *The Rainbow: From Myth to Mathematics* by Carl B. Boyer (Princeton, 1987).

Modern Physics (Part V)

- *The Road to Reality* by Roger Penrose (Knopf, 2004).
A recent survey of modern physics by a famous physicist.
- *QED: The Strange Theory of Light and Matter* by Richard P. Feynman (Princeton, 1988).
A famous Nobel laureate explains the theory of quantum electrodynamics at a level accessible to the general public.

Just for Fun

- *Physics of the Impossible* by Michio Kaku (Doubleday, 2008). A noted physicist discusses the possibility of time travel, force fields, invisibility cloaks, transporters, etc.
- *The Disappearing Spoon* by Sam Kean (Little, Brown & Co., 2010). A very entertaining collection of stories surrounding the periodic table of the elements.
- *Mr. Tompkins in Paperback* by George Gamow (Cambridge, 1993). A famous Russian physicist wrote these stories of a world in which the speed of light is just 30 mph so relativistic effects are visible, and more stories of a world where Planck's constant is so large that quantum effects are visible. An updated version has also been written, *The New World of Mr. Tompkins* (Cambridge, 2001).
- *Dragon's Egg* by Robert L. Forward (Del Rey, 2000). Physicist Robert Forward wrote this novel about humans who discover a civilization of creatures living on the surface of a neutron star.

Appendix B

Greek Alphabet

Table B-1. The Greek alphabet.

Letter	Name
A α	Alpha
B β	Beta
Γ γ	Gamma
Δ δ	Delta
E ε	Epsilon
Z ζ	Zeta
H η	Eta
Θ θ	Theta
I ι	Iota
K κ	Kappa
Λ λ	Lambda
M μ	Mu
N ν	Nu
Ξ ξ	Xi
O $\omicron$	Omicron
Π π	Pi
P ρ	Rho
Σ σ	Sigma
T τ	Tau
Υ υ	Upsilon
Φ φ	Phi
X χ	Chi
Ψ ψ	Psi
Ω ω	Omega

(Alternate forms: $\delta = \beta$, $\partial = \delta$, $\epsilon = \varepsilon$, $\vartheta = \theta$, $\kappa = \kappa$, $\varpi = \pi$, $\varrho = \rho$, $\varsigma = \sigma$, $\phi = \varphi$.)

Appendix C

Trigonometry

Basic Formulæ

$$\sin^2 \theta + \cos^2 \theta \equiv 1$$

$$\sec^2 \theta \equiv 1 + \tan^2 \theta$$

$$\csc^2 \theta \equiv 1 + \cot^2 \theta$$

Angle Addition Formulæ

$$\sin(\alpha \pm \beta) \equiv \sin \alpha \cos \beta \pm \cos \alpha \sin \beta$$

$$\cos(\alpha \pm \beta) \equiv \cos \alpha \cos \beta \mp \sin \alpha \sin \beta$$

$$\tan(\alpha \pm \beta) \equiv \frac{\tan \alpha \pm \tan \beta}{1 \mp \tan \alpha \tan \beta}$$

Double-Angle Formulæ

$$\sin 2\theta \equiv 2 \sin \theta \cos \theta$$

$$\cos 2\theta \equiv \cos^2 \theta - \sin^2 \theta \equiv 1 - 2 \sin^2 \theta \equiv 2 \cos^2 \theta - 1$$

$$\tan 2\theta \equiv \frac{2 \tan \theta}{1 - \tan^2 \theta}$$

Half-Angle Formulæ

$$\sin \frac{\theta}{2} \equiv \pm \sqrt{\frac{1 - \cos \theta}{2}}$$

$$\cos \frac{\theta}{2} \equiv \pm \sqrt{\frac{1 + \cos \theta}{2}}$$

$$\tan \frac{\theta}{2} \equiv \frac{\sin \theta}{1 + \cos \theta} \equiv \frac{1 - \cos \theta}{\sin \theta}$$

Products of Sines and Cosines

$$\sin \alpha \cos \beta \equiv \frac{1}{2} [\sin(\alpha + \beta) + \sin(\alpha - \beta)]$$

$$\cos \alpha \sin \beta \equiv \frac{1}{2} [\sin(\alpha + \beta) - \sin(\alpha - \beta)]$$

$$\cos \alpha \cos \beta \equiv \frac{1}{2} [\cos(\alpha + \beta) + \cos(\alpha - \beta)]$$

$$\sin \alpha \sin \beta \equiv -\frac{1}{2} [\cos(\alpha + \beta) - \cos(\alpha - \beta)]$$

Sums and Differences of Sines and Cosines

$$\sin \alpha + \sin \beta \equiv 2 \sin \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2}$$

$$\sin \alpha - \sin \beta \equiv 2 \cos \frac{\alpha + \beta}{2} \sin \frac{\alpha - \beta}{2}$$

$$\cos \alpha + \cos \beta \equiv 2 \cos \frac{\alpha + \beta}{2} \cos \frac{\alpha - \beta}{2}$$

$$\cos \alpha - \cos \beta \equiv -2 \sin \frac{\alpha + \beta}{2} \sin \frac{\alpha - \beta}{2}$$

Other Formulæ

$$\sin^2 \theta \equiv \frac{1}{2} (1 - \cos 2\theta)$$

$$\cos^2 \theta \equiv \frac{1}{2} (1 + \cos 2\theta)$$

$$\tan \theta \equiv \cot \theta - 2 \cot 2\theta$$

Exact values of trigonometric functions at 3° intervals. (Ref. [9])

θ	$\sin \theta$	$\cos \theta$	$\tan \theta$
$0^\circ = 0\pi$	0	1	0
$3^\circ = \frac{\pi}{60}$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} - 1) - 2(\sqrt{3} - 1)\sqrt{5 + \sqrt{5}}]$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 + \sqrt{5}} + (\sqrt{6} - \sqrt{2})(\sqrt{5} - 1)]$	$\frac{1}{4} (\sqrt{5} - \sqrt{3})(\sqrt{3} - 1)(\sqrt{10 + 2\sqrt{5}} - \sqrt{5} - 1)$
$6^\circ = \frac{\pi}{30}$	$\frac{1}{8} (\sqrt{30 - 6\sqrt{5}} - \sqrt{5} - 1)$	$\frac{1}{8} (\sqrt{15} + \sqrt{3} + \sqrt{10 - 2\sqrt{5}})$	$\frac{1}{2} (\sqrt{10 - 2\sqrt{5}} - \sqrt{15} + \sqrt{3})$
$9^\circ = \frac{\pi}{20}$	$\frac{1}{8} (\sqrt{10} + \sqrt{2} - 2\sqrt{5 - \sqrt{5}})$	$\frac{1}{8} (\sqrt{10} + \sqrt{2} + 2\sqrt{5 - \sqrt{5}})$	$\sqrt{5} + 1 - \sqrt{5} + 2\sqrt{5}$
$12^\circ = \frac{\pi}{15}$	$\frac{1}{8} (\sqrt{10 + 2\sqrt{5}} - \sqrt{15} + \sqrt{3})$	$\frac{1}{8} (\sqrt{30 + 6\sqrt{5}} + \sqrt{5} - 1)$	$\frac{1}{2} (3\sqrt{3} - \sqrt{15} - \sqrt{50 - 22\sqrt{5}})$
$15^\circ = \frac{\pi}{12}$	$\frac{1}{4} (\sqrt{6} - \sqrt{2})$	$\frac{1}{4} (\sqrt{6} + \sqrt{2})$	$2 - \sqrt{3}$
$18^\circ = \frac{\pi}{10}$	$\frac{1}{4} (\sqrt{5} - 1)$	$\frac{1}{4} \sqrt{10 + 2\sqrt{5}}$	$\frac{1}{5} \sqrt{25 - 10\sqrt{5}}$
$21^\circ = \frac{7\pi}{60}$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 - \sqrt{5}} - (\sqrt{6} - \sqrt{2})(\sqrt{5} + 1)]$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} + 1) + 2(\sqrt{3} - 1)\sqrt{5 - \sqrt{5}}]$	$\frac{1}{4} (\sqrt{5} - \sqrt{3})(\sqrt{3} + 1)(\sqrt{10 - 2\sqrt{5}} - \sqrt{5} + 1)$
$24^\circ = \frac{2\pi}{15}$	$\frac{1}{8} (\sqrt{15} + \sqrt{3} - \sqrt{10 - 2\sqrt{5}})$	$\frac{1}{8} (\sqrt{30 - 6\sqrt{5}} + \sqrt{5} + 1)$	$\frac{1}{2} (\sqrt{50 + 22\sqrt{5}} - 3\sqrt{3} - \sqrt{15})$
$27^\circ = \frac{3\pi}{20}$	$\frac{1}{8} (2\sqrt{5 + \sqrt{5}} - \sqrt{10} + \sqrt{2})$	$\frac{1}{8} (2\sqrt{5 + \sqrt{5}} + \sqrt{10} - \sqrt{2})$	$\sqrt{5} - 1 - \sqrt{5} - 2\sqrt{5}$
$30^\circ = \frac{\pi}{6}$	$\frac{1}{2}$	$\frac{1}{2} \sqrt{3}$	$\frac{1}{3} \sqrt{3}$
$33^\circ = \frac{11\pi}{60}$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} - 1) + 2(\sqrt{3} - 1)\sqrt{5 + \sqrt{5}}]$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 + \sqrt{5}} - (\sqrt{6} - \sqrt{2})(\sqrt{5} - 1)]$	$\frac{1}{4} (\sqrt{5} - \sqrt{3})(\sqrt{3} - 1)(\sqrt{10 + 2\sqrt{5}} + \sqrt{5} + 1)$
$36^\circ = \frac{\pi}{5}$	$\frac{1}{4} \sqrt{10 - 2\sqrt{5}}$	$\frac{1}{4} (\sqrt{5} + 1)$	$\sqrt{5} - 2\sqrt{5}$
$39^\circ = \frac{13\pi}{60}$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} + 1) - 2(\sqrt{3} - 1)\sqrt{5 - \sqrt{5}}]$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 - \sqrt{5}} + (\sqrt{6} - \sqrt{2})(\sqrt{5} + 1)]$	$\frac{1}{4} (\sqrt{5} + \sqrt{3})(\sqrt{3} - 1)(\sqrt{10 - 2\sqrt{5}} - \sqrt{5} + 1)$
$42^\circ = \frac{7\pi}{30}$	$\frac{1}{8} (\sqrt{30 + 6\sqrt{5}} - \sqrt{5} + 1)$	$\frac{1}{8} (\sqrt{10 + 2\sqrt{5}} + \sqrt{15} - \sqrt{3})$	$\frac{1}{2} (\sqrt{15} + \sqrt{3} - \sqrt{10 + 2\sqrt{5}})$
$45^\circ = \frac{\pi}{4}$	$\frac{1}{2} \sqrt{2}$	$\frac{1}{2} \sqrt{2}$	1
$48^\circ = \frac{4\pi}{15}$	$\frac{1}{8} (\sqrt{10 + 2\sqrt{5}} + \sqrt{15} - \sqrt{3})$	$\frac{1}{8} (\sqrt{30 + 6\sqrt{5}} - \sqrt{5} + 1)$	$\frac{1}{2} (3\sqrt{3} - \sqrt{15} + \sqrt{50 - 22\sqrt{5}})$
$51^\circ = \frac{17\pi}{60}$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 - \sqrt{5}} + (\sqrt{6} - \sqrt{2})(\sqrt{5} + 1)]$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} + 1) - 2(\sqrt{3} - 1)\sqrt{5 - \sqrt{5}}]$	$\frac{1}{4} (\sqrt{5} - \sqrt{3})(\sqrt{3} + 1)(\sqrt{10 - 2\sqrt{5}} + \sqrt{5} - 1)$
$54^\circ = \frac{3\pi}{10}$	$\frac{1}{4} (\sqrt{5} + 1)$	$\frac{1}{4} \sqrt{10 - 2\sqrt{5}}$	$\frac{1}{5} \sqrt{25 + 10\sqrt{5}}$
$57^\circ = \frac{19\pi}{60}$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 + \sqrt{5}} - (\sqrt{6} - \sqrt{2})(\sqrt{5} - 1)]$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} - 1) + 2(\sqrt{3} - 1)\sqrt{5 + \sqrt{5}}]$	$\frac{1}{4} (\sqrt{5} + \sqrt{3})(\sqrt{3} + 1)(\sqrt{10 + 2\sqrt{5}} - \sqrt{5} - 1)$
$60^\circ = \frac{\pi}{3}$	$\frac{1}{2} \sqrt{3}$	$\frac{1}{2}$	$\sqrt{3}$
$63^\circ = \frac{7\pi}{20}$	$\frac{1}{8} (2\sqrt{5 + \sqrt{5}} + \sqrt{10} - \sqrt{2})$	$\frac{1}{8} (2\sqrt{5 + \sqrt{5}} - \sqrt{10} + \sqrt{2})$	$\sqrt{5} - 1 + \sqrt{5} - 2\sqrt{5}$
$66^\circ = \frac{11\pi}{30}$	$\frac{1}{8} (\sqrt{30 - 6\sqrt{5}} + \sqrt{5} + 1)$	$\frac{1}{8} (\sqrt{15} + \sqrt{3} - \sqrt{10 - 2\sqrt{5}})$	$\frac{1}{2} (\sqrt{10 - 2\sqrt{5}} + \sqrt{15} - \sqrt{3})$
$69^\circ = \frac{23\pi}{60}$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} + 1) + 2(\sqrt{3} - 1)\sqrt{5 - \sqrt{5}}]$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 - \sqrt{5}} - (\sqrt{6} - \sqrt{2})(\sqrt{5} + 1)]$	$\frac{1}{4} (\sqrt{5} + \sqrt{3})(\sqrt{3} - 1)(\sqrt{10 - 2\sqrt{5}} + \sqrt{5} - 1)$
$72^\circ = \frac{2\pi}{5}$	$\frac{1}{4} \sqrt{10 + 2\sqrt{5}}$	$\frac{1}{4} (\sqrt{5} - 1)$	$\sqrt{5} + 2\sqrt{5}$
$75^\circ = \frac{5\pi}{12}$	$\frac{1}{4} (\sqrt{6} + \sqrt{2})$	$\frac{1}{4} (\sqrt{6} - \sqrt{2})$	$2 + \sqrt{3}$
$78^\circ = \frac{13\pi}{30}$	$\frac{1}{8} (\sqrt{30 + 6\sqrt{5}} + \sqrt{5} - 1)$	$\frac{1}{8} (\sqrt{10 + 2\sqrt{5}} - \sqrt{15} + \sqrt{3})$	$\frac{1}{2} (\sqrt{15} + \sqrt{3} + \sqrt{10 + 2\sqrt{5}})$
$81^\circ = \frac{19\pi}{20}$	$\frac{1}{8} (\sqrt{10} + \sqrt{2} + 2\sqrt{5 - \sqrt{5}})$	$\frac{1}{8} (\sqrt{10} + \sqrt{2} - 2\sqrt{5 - \sqrt{5}})$	$\sqrt{5} + 1 + \sqrt{5} + 2\sqrt{5}$
$84^\circ = \frac{7\pi}{15}$	$\frac{1}{8} (\sqrt{15} + \sqrt{3} + \sqrt{10 - 2\sqrt{5}})$	$\frac{1}{8} (\sqrt{30 - 6\sqrt{5}} - \sqrt{5} - 1)$	$\frac{1}{2} (\sqrt{50 + 22\sqrt{5}} + 3\sqrt{3} + \sqrt{15})$
$87^\circ = \frac{29\pi}{60}$	$\frac{1}{16} [2(\sqrt{3} + 1)\sqrt{5 + \sqrt{5}} + (\sqrt{6} - \sqrt{2})(\sqrt{5} - 1)]$	$\frac{1}{16} [(\sqrt{6} + \sqrt{2})(\sqrt{5} - 1) - 2(\sqrt{3} - 1)\sqrt{5 + \sqrt{5}}]$	$\frac{1}{4} (\sqrt{5} + \sqrt{3})(\sqrt{3} + 1)(\sqrt{10 + 2\sqrt{5}} + \sqrt{5} + 1)$
$90^\circ = \frac{\pi}{2}$	1	0	∞

θ	$\sec \theta$	$\csc \theta$	$\cot \theta$
$0^\circ = 0\pi$	1	∞	∞
$3^\circ = \frac{\pi}{60}$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(\sqrt{5+2\sqrt{5}}-2+\sqrt{3})$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(2+\sqrt{3}+\sqrt{5+2\sqrt{5}})$	$\frac{1}{4}(\sqrt{5}+\sqrt{3})(\sqrt{3}+1)(\sqrt{10+2\sqrt{5}}+\sqrt{5}+1)$
$6^\circ = \frac{\pi}{30}$	$\sqrt{3}-\sqrt{5-2\sqrt{5}}$	$\sqrt{15+6\sqrt{5}}+\sqrt{5}+2$	$\frac{1}{2}(\sqrt{50+22\sqrt{5}}+3\sqrt{3}+\sqrt{15})$
$9^\circ = \frac{\pi}{20}$	$\frac{1}{2}(3\sqrt{2}+\sqrt{10}-2\sqrt{5+\sqrt{5}})$	$\frac{1}{2}(3\sqrt{2}+\sqrt{10}+2\sqrt{5}+\sqrt{5})$	$\sqrt{5}+1+\sqrt{5+2\sqrt{5}}$
$12^\circ = \frac{\pi}{15}$	$\sqrt{15-6\sqrt{5}}-\sqrt{5}+2$	$\sqrt{5+2\sqrt{5}}+\sqrt{3}$	$\frac{1}{2}(\sqrt{15}+\sqrt{3}+\sqrt{10+2\sqrt{5}})$
$15^\circ = \frac{\pi}{12}$	$\sqrt{6}-\sqrt{2}$	$\sqrt{6}+\sqrt{2}$	$2+\sqrt{3}$
$18^\circ = \frac{\pi}{10}$	$\frac{1}{2}\sqrt{50-10\sqrt{5}}$	$\sqrt{5}+1$	$\sqrt{5+2\sqrt{5}}$
$21^\circ = \frac{7\pi}{60}$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(2+\sqrt{3}-\sqrt{5-2\sqrt{5}})$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(\sqrt{5-2\sqrt{5}}+2-\sqrt{3})$	$\frac{1}{4}(\sqrt{5}+\sqrt{3})(\sqrt{3}-1)(\sqrt{10-2\sqrt{5}}+\sqrt{5}-1)$
$24^\circ = \frac{2\pi}{15}$	$\sqrt{15+6\sqrt{5}}-\sqrt{5}-2$	$\sqrt{3}+\sqrt{5-2\sqrt{5}}$	$\frac{1}{2}(\sqrt{10-2\sqrt{5}}+\sqrt{15}-\sqrt{3})$
$27^\circ = \frac{3\pi}{20}$	$\frac{1}{2}(2\sqrt{5}-\sqrt{5}-3\sqrt{2}+\sqrt{10})$	$\frac{1}{2}(2\sqrt{5}-\sqrt{5}+3\sqrt{2}-\sqrt{10})$	$\sqrt{5}-1+\sqrt{5-2\sqrt{5}}$
$30^\circ = \frac{\pi}{6}$	$\frac{2}{3}\sqrt{3}$	2	$\sqrt{3}$
$33^\circ = \frac{11\pi}{60}$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(\sqrt{5+2\sqrt{5}}+2-\sqrt{3})$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(2+\sqrt{3}-\sqrt{5+2\sqrt{5}})$	$\frac{1}{4}(\sqrt{5}+\sqrt{3})(\sqrt{3}+1)(\sqrt{10+2\sqrt{5}}-\sqrt{5}-1)$
$36^\circ = \frac{\pi}{5}$	$\sqrt{5}-1$	$\frac{1}{2}\sqrt{50+10\sqrt{5}}$	$\frac{1}{2}\sqrt{25+10\sqrt{5}}$
$39^\circ = \frac{13\pi}{60}$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(\sqrt{5-2\sqrt{5}}-2+\sqrt{3})$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(2+\sqrt{3}+\sqrt{5-2\sqrt{5}})$	$\frac{1}{4}(\sqrt{5}-\sqrt{3})(\sqrt{3}+1)(\sqrt{10-2\sqrt{5}}+\sqrt{5}-1)$
$42^\circ = \frac{7\pi}{30}$	$\sqrt{5+2\sqrt{5}}-\sqrt{3}$	$\sqrt{15-6\sqrt{5}}+\sqrt{5}-2$	$\frac{1}{2}(3\sqrt{3}-\sqrt{15}+\sqrt{50-22\sqrt{5}})$
$45^\circ = \frac{\pi}{4}$	$\sqrt{2}$	$\sqrt{2}$	1
$48^\circ = \frac{4\pi}{15}$	$\sqrt{15-6\sqrt{5}}+\sqrt{5}-2$	$\sqrt{5+2\sqrt{5}}-\sqrt{3}$	$\frac{1}{2}(\sqrt{15}+\sqrt{3}-\sqrt{10+2\sqrt{5}})$
$51^\circ = \frac{17\pi}{60}$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(2+\sqrt{3}+\sqrt{5-2\sqrt{5}})$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(\sqrt{5-2\sqrt{5}}-2+\sqrt{3})$	$\frac{1}{4}(\sqrt{5}+\sqrt{3})(\sqrt{3}-1)(\sqrt{10-2\sqrt{5}}-\sqrt{5}+1)$
$54^\circ = \frac{3\pi}{10}$	$\frac{1}{2}\sqrt{50+10\sqrt{5}}$	$\sqrt{5}-1$	$\sqrt{5-2\sqrt{5}}$
$57^\circ = \frac{19\pi}{60}$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(2+\sqrt{3}-\sqrt{5+2\sqrt{5}})$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(\sqrt{5+2\sqrt{5}}+2-\sqrt{3})$	$\frac{1}{4}(\sqrt{5}-\sqrt{3})(\sqrt{3}-1)(\sqrt{10+2\sqrt{5}}+\sqrt{5}+1)$
$60^\circ = \frac{\pi}{3}$	2	$\frac{2}{3}\sqrt{3}$	$\frac{1}{2}\sqrt{3}$
$63^\circ = \frac{7\pi}{20}$	$\frac{1}{2}(2\sqrt{5}-\sqrt{5}+3\sqrt{2}-\sqrt{10})$	$\frac{1}{2}(2\sqrt{5}-\sqrt{5}-3\sqrt{2}+\sqrt{10})$	$\sqrt{5}-1-\sqrt{5-2\sqrt{5}}$
$66^\circ = \frac{11\pi}{30}$	$\sqrt{3}+\sqrt{5-2\sqrt{5}}$	$\sqrt{15+6\sqrt{5}}-\sqrt{5}-2$	$\frac{1}{2}(\sqrt{50+22\sqrt{5}}-3\sqrt{3}-\sqrt{15})$
$69^\circ = \frac{23\pi}{60}$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(\sqrt{5-2\sqrt{5}}+2-\sqrt{3})$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(2+\sqrt{3}-\sqrt{5-2\sqrt{5}})$	$\frac{1}{4}(\sqrt{5}-\sqrt{3})(\sqrt{3}+1)(\sqrt{10-2\sqrt{5}}-\sqrt{5}+1)$
$72^\circ = \frac{2\pi}{5}$	$\sqrt{5}+1$	$\frac{1}{2}\sqrt{50-10\sqrt{5}}$	$\frac{1}{2}\sqrt{25-10\sqrt{5}}$
$75^\circ = \frac{5\pi}{12}$	$\sqrt{6}+\sqrt{2}$	$\sqrt{6}-\sqrt{2}$	$2-\sqrt{3}$
$78^\circ = \frac{13\pi}{30}$	$\sqrt{5+2\sqrt{5}}+\sqrt{3}$	$\sqrt{15-6\sqrt{5}}-\sqrt{5}+2$	$\frac{1}{2}(3\sqrt{3}-\sqrt{15}-\sqrt{50-22\sqrt{5}})$
$81^\circ = \frac{19\pi}{60}$	$\frac{1}{2}(3\sqrt{2}+\sqrt{10}+2\sqrt{5}+\sqrt{5})$	$\frac{1}{2}(3\sqrt{2}+\sqrt{10}-2\sqrt{5}+\sqrt{5})$	$\sqrt{5}+1-\sqrt{5+2\sqrt{5}}$
$84^\circ = \frac{7\pi}{15}$	$\sqrt{15+6\sqrt{5}}+\sqrt{5}+2$	$\sqrt{3}-\sqrt{5-2\sqrt{5}}$	$\frac{1}{2}(\sqrt{10-2\sqrt{5}}-\sqrt{15}+\sqrt{3})$
$87^\circ = \frac{29\pi}{60}$	$\frac{1}{2}(\sqrt{10}+\sqrt{6})(2+\sqrt{3}+\sqrt{5+2\sqrt{5}})$	$\frac{1}{2}(\sqrt{10}-\sqrt{6})(\sqrt{5+2\sqrt{5}}-2+\sqrt{3})$	$\frac{1}{4}(\sqrt{5}-\sqrt{3})(\sqrt{3}-1)(\sqrt{10+2\sqrt{5}}-\sqrt{5}-1)$
$90^\circ = \frac{\pi}{2}$	∞	1	0

Appendix D

Useful Series

The first four series are valid if $|x| < 1$; the fifth is valid for $x^2 < a^2$; and the last three are valid for all real x .

$$(1+x)^{1/2} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{128}x^4 + \frac{7}{256}x^5 - \frac{21}{1024}x^6 + \frac{33}{2048}x^7 - \frac{429}{32768}x^8 + \dots \quad (\text{D.1})$$

$$(1-x)^{1/2} = 1 - \frac{1}{2}x - \frac{1}{8}x^2 - \frac{1}{16}x^3 - \frac{5}{128}x^4 - \frac{7}{256}x^5 - \frac{21}{1024}x^6 - \frac{33}{2048}x^7 - \frac{429}{32768}x^8 - \dots \quad (\text{D.2})$$

$$(1+x)^{-1/2} = 1 - \frac{1}{2}x + \frac{3}{8}x^2 - \frac{5}{16}x^3 + \frac{35}{128}x^4 - \frac{63}{256}x^5 + \frac{231}{1024}x^6 - \frac{429}{2048}x^7 + \frac{6435}{32768}x^8 - \dots \quad (\text{D.3})$$

$$(1-x)^{-1/2} = 1 + \frac{1}{2}x + \frac{3}{8}x^2 + \frac{5}{16}x^3 + \frac{35}{128}x^4 + \frac{63}{256}x^5 + \frac{231}{1024}x^6 + \frac{429}{2048}x^7 + \frac{6435}{32768}x^8 + \dots \quad (\text{D.4})$$

$$\frac{1}{a+x} = \frac{1}{a} - \frac{x}{a^2} + \frac{x^2}{a^3} - \frac{x^3}{a^4} + \frac{x^4}{a^5} - \frac{x^5}{a^6} + \dots \quad (\text{D.5})$$

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \frac{x^7}{5040} + \frac{x^8}{40320} + \frac{x^9}{362880} + \dots \quad (\text{D.6})$$

$$\sin x = x - \frac{x^3}{6} + \frac{x^5}{120} - \frac{x^7}{5040} + \frac{x^9}{362880} - \frac{x^{11}}{39916800} + \frac{x^{13}}{6227020800} - \dots \quad (\text{D.7})$$

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{24} - \frac{x^6}{720} + \frac{x^8}{40320} - \frac{x^{10}}{3628800} + \frac{x^{12}}{479001600} - \dots \quad (\text{D.8})$$

Appendix E

Table of Derivatives

$$\frac{d}{dx} a = 0 \quad (\text{E.1})$$

$$\frac{d}{dx} x = 1 \quad (\text{E.2})$$

$$\frac{d}{dx} x^n = nx^{n-1} \quad (\text{E.3})$$

$$\frac{d}{dx} \sqrt{x} = \frac{1}{2\sqrt{x}} \quad (\text{E.4})$$

$$\frac{d}{dx} \sin x = \cos x \quad (\text{E.5})$$

$$\frac{d}{dx} \cos x = -\sin x \quad (\text{E.6})$$

$$\frac{d}{dx} \tan x = \sec^2 x \quad (\text{E.7})$$

$$\frac{d}{dx} \sec x = \tan x \sec x \quad (\text{E.8})$$

$$\frac{d}{dx} \csc x = -\cot x \csc x \quad (\text{E.9})$$

$$\frac{d}{dx} \cot x = -\csc^2 x \quad (\text{E.10})$$

$$(\text{E.11})$$

$$\frac{d}{dx} e^x = e^x \quad (\text{E.12})$$

$$\frac{d}{dx} \ln x = \frac{1}{x} \quad (\text{E.13})$$

$$\frac{d}{dx} a^x = a^x \ln a \quad (\text{E.14})$$

$$\frac{d}{dx} \log_a x = \frac{1}{x \ln a} \quad (\text{E.15})$$

$$\frac{d}{dx} \sinh x = \cosh x \quad (\text{E.16})$$

$$\frac{d}{dx} \cosh x = \sinh x \quad (\text{E.17})$$

$$\frac{d}{dx} \tanh x = \text{sech}^2 x \quad (\text{E.18})$$

Appendix F

Table of Integrals

In the following table, an arbitrary constant C should be added to each result.

$$\int dx = x \tag{F.1}$$

$$\int a \, dx = ax \tag{F.2}$$

$$\int x^n \, dx = \frac{x^{n+1}}{n+1} \quad (n \neq -1) \tag{F.3}$$

$$\int \sqrt{x} \, dx = \frac{2}{3} \sqrt{x^3} \tag{F.4}$$

$$\int \frac{1}{x} \, dx = \ln |x| \tag{F.5}$$

$$\int \sin x \, dx = -\cos x \tag{F.6}$$

$$\int \cos x \, dx = \sin x \tag{F.7}$$

$$\int \tan x \, dx = \ln |\sec x| \tag{F.8}$$

$$\tag{F.9}$$

$$\int \sec x \, dx = \ln |\sec x + \tan x| \quad (\text{F.10})$$

$$\int \csc x \, dx = \ln |\csc x - \cot x| \quad (\text{F.11})$$

$$\int \cot x \, dx = \ln |\sin x| \quad (\text{F.12})$$

$$\int e^x \, dx = e^x \quad (\text{F.13})$$

$$\int \ln x \, dx = x \ln x - x \quad (\text{F.14})$$

$$\int a^x \, dx = \frac{a^x}{\ln a} \quad (\text{F.15})$$

$$\int \log_a x \, dx = \frac{x \ln x - x}{\ln a} \quad (\text{F.16})$$

$$\int \sinh x \, dx = \cosh x \quad (\text{F.17})$$

$$\int \cosh x \, dx = \sinh x \quad (\text{F.18})$$

$$\int \tanh x \, dx = \ln \cosh x \quad (\text{F.19})$$

Appendix G

Mathematical Subtleties

- When taking the square root of both sides of an equation, a $\pm$ sign must always be introduced. For example:

$$x^2 = a \quad \Rightarrow \quad x = \pm\sqrt{a}$$

Both roots may be valid, or, depending on the problem, it may be that one root or the other may be rejected on mathematical or physical grounds.

- Dividing an equation through by a variable may result in losing roots. For example, suppose we have

$$x^2 - ax = 0$$

Dividing through by the variable x will result in one solution, $x = a$; the solution $x = 0$ has been lost. Instead of dividing through by the variable x , the proper procedure is to *factor out* an x :

$$x(x - a) = 0$$

Since the product on the left-hand side is zero, it follows that either $x = 0$ or $x - a = 0$, and we retain both roots.

- The relation

$$\sqrt{x}\sqrt{y} = \sqrt{xy} \tag{G.1}$$

is valid only for $x, y \geq 0$.

- Some mathematical conventions:

- ★ 1 is *not* considered a prime number.
- ★ $0! = 1$
- ★ $0^0 = 1$
- ★ Towers of exponents are evaluated from the top down: $a^{b^c} = a^{(b^c)}$

- When taking an inverse trigonometric function, there will in general be *two* correct values; your calculator will give only one value, the *principal value* (P.V.). The other value is found using the table below.

Function	P.V.	Other value
$\arcsin$	θ	$\pi - \theta$
$\arccos$	θ	$-\theta$
$\arctan$	θ	$\pi + \theta$
arcsec	θ	$-\theta$
arccsc	θ	$\pi - \theta$
arccot	θ	$\pi + \theta$

For $\arctan (y/x)$, add π to the calculator's principal value answer if $x < 0$.

Appendix H

SI Units

Table H-1. SI base units.

Name	Symbol	Quantity
meter	m	length
kilogram	kg	mass
second	s	time
ampere	A	electric current
kelvin	K	temperature
mole	mol	amount of substance
candela	cd	luminous intensity

Table H-2. Derived SI units.

Name	Symbol	Definition	Base Units	Quantity
radian	rad	m / m	—	plane angle
steradian	sr	m ² / m ²	—	solid angle
newton	N	kg m s ⁻²	kg m s ⁻²	force
joule	J	N m	kg m ² s ⁻²	energy
watt	W	J / s	kg m ² s ⁻³	power
pascal	Pa	N / m ²	kg m ⁻¹ s ⁻²	pressure
hertz	Hz	s ⁻¹	s ⁻¹	frequency
coulomb	C	A s	A s	electric charge
volt	V	J / C	kg m ² A ⁻¹ s ⁻³	electric potential
ohm	Ω	V / A	kg m ² A ⁻² s ⁻³	electrical resistance
siemens	S	A / V	kg ⁻¹ m ⁻² A ² s ³	electrical conductance
farad	F	C / V	kg ⁻¹ m ⁻² A ² s ⁴	capacitance
weber	Wb	V s	kg m ² A ⁻¹ s ⁻²	magnetic flux
tesla	T	Wb / m ²	kg A ⁻¹ s ⁻²	magnetic induction
henry	H	Wb / A	kg m ² A ⁻² s ⁻²	induction
lumen	lm	cd sr	cd sr	luminous flux
lux	lx	lm / m ²	cd sr m ⁻²	illuminance
becquerel	Bq	s ⁻¹	s ⁻¹	radioactivity
gray	Gy	J / kg	m ² s ⁻²	absorbed dose
sievert	Sv	J / kg	m ² s ⁻²	dose equivalent
katal	kat	mol / s	mol s ⁻¹	catalytic activity

Table H-3. SI prefixes.

Prefix	Symbol	Definition	English
yotta-	Y	10^{24}	septillion
zetta-	Z	10^{21}	sextillion
exa-	E	10^{18}	quintillion
peta-	P	10^{15}	quadrillion
tera-	T	10^{12}	trillion
giga-	G	10^9	billion
mega-	M	10^6	million
kilo-	k	10^3	thousand
hecto-	h	10^2	hundred
deka-	da	10^1	ten
deci-	d	10^{-1}	tenth
centi-	c	10^{-2}	hundredth
milli-	m	10^{-3}	thousandth
micro-	μ	10^{-6}	millionth
nano-	n	10^{-9}	billionth
pico-	p	10^{-12}	trillionth
femto-	f	10^{-15}	quadrillionth
atto-	a	10^{-18}	quintillionth
zepto-	z	10^{-21}	sextillionth
yocto-	y	10^{-24}	septillionth

Table H-4. Prefixes for *computer use only*.

Prefix	Symbol	Definition
yobi-	Yi	$2^{80} = 1,208,925,819,614,629,174,706,176$
zebi-	Zi	$2^{70} = 1,180,591,620,717,411,303,424$
exbi-	Ei	$2^{60} = 1,152,921,504,606,846,976$
pebi-	Pi	$2^{50} = 1,125,899,906,842,624$
tebi-	Ti	$2^{40} = 1,099,511,627,776$
gibi-	Gi	$2^{30} = 1,073,741,824$
mebi-	Mi	$2^{20} = 1,048,576$
kibi-	Ki	$2^{10} = 1,024$

Appendix I

Gaussian Units

Table I-1. Gaussian base units.

Name	Symbol	Quantity
centimeter	cm	length
gram	g	mass
second	s	time
kelvin	K	temperature
mole	mol	amount of substance
candela	cd	luminous intensity

Table I-2. Derived Gaussian units.

Name	Symbol	Definition	Base Units	Quantity
radian	rad	m / m	—	plane angle
steradian	sr	m ² / m ²	—	solid angle
dyne	dyn	g cm s ⁻²	g cm s ⁻²	force
erg	erg	dyn cm	g cm ² s ⁻²	energy
statwatt	statW	erg / s	g cm ² s ⁻³	power
barye	ba	dyn / cm ²	g cm ⁻¹ s ⁻²	pressure
galileo	Gal	cm / s ²	cm s ⁻²	acceleration
poise	P	g / (cm s)	g cm ⁻¹ s ⁻¹	dynamic viscosity
stokes	St	cm ² / s	cm ² s ⁻¹	kinematic viscosity
hertz	Hz	s ⁻¹	s ⁻¹	frequency
statcoulomb	statC		g ^{1/2} cm ^{3/2} s ⁻¹	electric charge
franklin	Fr	statC	g ^{1/2} cm ^{3/2} s ⁻¹	electric charge
statampere	statA	statC / s	g ^{1/2} cm ^{3/2} s ⁻²	electric current
statvolt	statV	erg / statC	g ^{1/2} cm ^{1/2} s ⁻¹	electric potential
statohm	statΩ	statV / statA	s cm ⁻¹	electrical resistance
statfarad	statF	statC / statV	cm	capacitance
maxwell	Mx	statV cm	g ^{1/2} cm ^{3/2} s ⁻¹	magnetic flux
gauss	G	Mx / cm ²	g ^{1/2} cm ^{-1/2} s ⁻¹	magnetic induction
oersted	Oe	statA s / cm ²	g ^{1/2} cm ^{-1/2} s ⁻¹	magnetic intensity
gilbert	Gb	statA	g ^{1/2} cm ^{3/2} s ⁻²	magnetomotive force
unit pole	pole	dyn / Oe	g ^{1/2} cm ^{3/2} s ⁻¹	magnetic pole strength
stathenry	statH	erg / statA ²	s ² cm ⁻¹	induction
lumen	lm	cd sr	cd sr	luminous flux
phot	ph	lm / cm ²	cd sr cm ⁻²	illuminance
stilb	sb	cd / cm ²	cd cm ⁻²	luminance
lambert	Lb	1/π cd / cm ²	cd cm ⁻²	luminance
kayser	K	1 / cm	cm ⁻¹	wave number
becquerel	Bq	s ⁻¹	s ⁻¹	radioactivity
katal	kat	mol / s	mol s ⁻¹	catalytic activity

Appendix J

Units of Physical Quantities

Table J-1. Units of physical quantities.

Quantity	SI Units	Gaussian Units
Absorbed dose	Gy	erg g ⁻¹
Acceleration	m s ⁻²	cm s ⁻²
Amount of substance	mol	mol
Angle (plane)	rad	rad
Angle (solid)	sr	sr
Angular acceleration	rad s ⁻²	rad s ⁻²
Angular momentum	N m s	dyn cm s
Angular velocity	rad s ⁻¹	rad s ⁻¹
Area	m ²	cm ²
Bulk modulus	Pa	ba
Catalytic activity	kat	kat
Coercivity	A m ⁻¹	Oe
Crackle	m s ⁻⁵	cm s ⁻⁵
Density	kg m ⁻³	g cm ⁻³
Distance	m	cm
Dose equivalent	Sv	erg g ⁻¹
Elastic modulus	N m ⁻²	dyn cm ⁻²
Electric capacitance	F	statF
Electric charge	C	statC
Electric conductance	S	statΩ ⁻¹
Electric conductivity	S m ⁻¹	statΩ ⁻¹ cm ⁻¹
Electric current	A	statA
Electric dipole moment	C m	statC cm
Electric displacement (<i>D</i>)	C m ⁻²	statC cm ⁻²
Electric elastance	F ⁻¹	statF ⁻¹
Electric field (<i>E</i>)	V m ⁻¹	statV cm ⁻¹
Electric flux	V m	statV cm
Electric permittivity	F m ⁻¹	—
Electric polarization (<i>P</i>)	C m ⁻²	statC cm ⁻²
Electric potential	V	statV
Electric resistance	Ω	statΩ
Electric resistivity	Ω m	statΩ cm

Table J-1 (cont'd). Units of physical quantities.

Quantity	SI Units	Gaussian Units
Energy	J	erg
Enthalpy	J	erg
Entropy	J K ⁻¹	erg K ⁻¹
Force	N	dyn
Frequency	Hz	Hz
Heat	J	erg
Heat capacity	J K ⁻¹	erg K ⁻¹
Illuminance	lx	ph
Impulse	N s	dyn s
Inductance	H	statH
Jerk	m s ⁻³	cm s ⁻³
Jounce	m s ⁻⁴	cm s ⁻⁴
Latent heat	J kg ⁻¹	erg g ⁻¹
Length	m	cm
Luminance	cd m ⁻²	sb
Luminous flux	lm	lm
Luminous intensity	cd	cd
Magnetic flux	Wb	Mx
Magnetic induction (B)	T	G
Magnetic intensity (H)	A m ⁻¹	Oe
Magnetic dipole moment (B convention)	A m ²	pole cm
Magnetic dipole moment (H convention)	Wb m	pole cm
Magnetic permeability	H m ⁻¹	—
Magnetic permeance	H	s
Magnetic pole strength (B convention)	A m	unit pole
Magnetic pole strength (H convention)	Wb	unit pole
Magnetic potential (scalar)	A	Oe cm
Magnetic potential (vector)	T m	G cm
Magnetic reluctance	H ⁻¹	s ⁻¹
Magnetization (M)	A m ⁻¹	Mx cm ⁻²
Magnetomotive force	A	Gb
Mass	kg	g
Memristance	Ω	stat Ω
Molality	mol kg ⁻¹	mol g ⁻¹
Molarity	mol m ⁻³	mol cm ⁻³
Moment of inertia	kg m ²	g cm ²
Momentum	N s	dyn s
Pop	m s ⁻⁶	cm s ⁻⁶
Power	W	statW
Pressure	Pa	ba
Radioactivity	Bq	Bq
Remanence	T	G
Retentivity	T	G

Table J-1 (cont'd). Units of physical quantities.

Quantity	SI Units	Gaussian Units
Shear modulus	N m^{-2}	dyn cm^{-2}
Snap	m s^{-4}	cm s^{-4}
Specific heat	$\text{J K}^{-1} \text{kg}^{-1}$	$\text{erg K}^{-1} \text{g}^{-1}$
Strain	—	—
Stress	N m^{-2}	dyn cm^{-2}
Temperature	K	K
Tension	N	dyn
Time	s	s
Torque	N m	dyn cm
Velocity	m s^{-1}	cm s^{-1}
Viscosity (dynamic)	Pa s	P
Viscosity (kinematic)	$\text{m}^2 \text{s}^{-1}$	St
Volume	m^3	cm^3
Wave number	m^{-1}	kayser
Weight	N	dyn
Work	J	erg
Young's modulus	N m^{-2}	dyn cm^{-2}

Appendix K

Physical Constants

Table K-1. Fundamental physical constants (CODATA 2010).

Description	Symbol	Value
Speed of light (vacuum)	c	$2.99792458 \times 10^8 \text{ m/s}$
Gravitational constant	G	$6.67408 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$
Elementary charge	e	$1.6021766208 \times 10^{-19} \text{ C}$
Permittivity of free space	ϵ_0	$8.85418781762038985 \dots \times 10^{-12} \text{ F/m}$
Permeability of free space	μ_0	$4\pi \times 10^{-7} \text{ N/A}^2$
Coulomb constant ($1/(4\pi\epsilon_0)$)	k_c	$8.9875517873681764 \times 10^9 \text{ m/F}$
Electron mass	m_e	$9.10938356 \times 10^{-31} \text{ kg}$
Proton mass	m_p	$1.672621898 \times 10^{-27} \text{ kg}$
Neutron mass	m_n	$1.674927471 \times 10^{-27} \text{ kg}$
Atomic mass unit	u	$1.660539040 \times 10^{-27} \text{ kg}$
Planck constant	h	$6.626070040 \times 10^{-34} \text{ J s}$
Planck constant $\div 2\pi$	$\hbar$	$1.054571800 \times 10^{-34} \text{ J s}$
Boltzmann constant	k_B	$1.38064852 \times 10^{-23} \text{ J/K}$

Table K-2. Other physical constants.

Description	Symbol	Value
Acceleration due to gravity at Earth surface	g	9.80 m/s^2
Speed of sound in air (20°C)	v_{snd}	343 m/s
Density of air (sea level)	ρ_{air}	1.29 kg/m^3
Density of water	ρ_{w}	$1 \text{ g/cm}^3 = 1000 \text{ kg/m}^3$
Index of refraction of water	n_{w}	1.33
Resistivity of copper (20°C)	ρ_{Cu}	$1.68 \times 10^{-8} \Omega \text{ m}$

Appendix L

Astronomical Data

Table L-1. Astronomical constants.

Description	Symbol	Value
Astronomical unit	AU	$1.49597870 \times 10^{11}$ m
Obliquity of ecliptic (J2000)	ε	23°4392911
Solar mass	$M_{\odot}$	1.9891×10^{30} kg
Solar radius	$R_{\odot}$	696,000 km

Table L-2. Planetary Data.

Planet	Mass (Yg)	Eq. radius (km)	Orbit semi-major axis (Gm)
Mercury	330.2	2439.7	57.91
Venus	4868.5	6051.8	108.21
Earth	5973.6	6378.1	149.60
Mars	641.85	3396.2	227.92
Jupiter	1,898,600	71,492	778.57
Saturn	568,460	60,268	1433.53
Uranus	86,832	25,559	2872.46
Neptune	102,430	24,764	4495.06
Pluto	12.5	1195	5906.38

Appendix M

Unit Conversion Tables

Time

1 day = 24 hours = 1440 minutes = 86400 seconds
1 hour = 60 minutes = 3600 seconds
1 year = 31 557 600 seconds $\approx \pi \times 10^7$ seconds

Length

1 mile = 8 furlongs = 80 chains = 320 rods = 1760 yards = 5280 feet = 1.609344 km
1 yard = 3 feet = 36 inches = 0.9144 meter
1 foot = 12 inches = 0.3048 meter
1 inch = 2.54 cm
1 nautical mile = 1852 meters = 1.15077944802354 miles
1 fathom = 6 feet
1 parsec = 3.26156376188 light-years = 206264.806245 AU = $3.08567756703 \times 10^{16}$ meters
1 ångström = 0.1 nm = 10^5 fermi = 10^{-10} meter

Mass

1 kilogram = 2.20462262184878 lb
1 pound = 16 oz = 0.45359237 kg
1 slug = 32.1740485564304 lb = 14.5939029372064 kg
1 short ton = 2000 lb
1 long ton = 2240 lb
1 metric ton = 1000 kg

Velocity

15 mph = 22 fps
1 mph = 0.44704 m/s
1 knot = 1.15077944802354 mph = 0.514444444444444 m/s

Area

$$1 \text{ acre} = 43560 \text{ ft}^2 = 4840 \text{ yd}^2 = 4046.8564224 \text{ m}^2$$

$$1 \text{ mile}^2 = 640 \text{ acres} = 2.589988110336 \text{ km}^2$$

$$1 \text{ are} = 100 \text{ m}^2$$

$$1 \text{ hectare} = 10^4 \text{ m}^2 = 2.47105381467165 \text{ acres}$$

Volume

$$1 \text{ liter} = 1 \text{ dm}^3 = 10^{-3} \text{ m}^3 \approx 1 \text{ quart}$$

$$1 \text{ m}^3 = 1000 \text{ liters}$$

$$1 \text{ cm}^3 = 1 \text{ mL}$$

$$1 \text{ ft}^3 = 1728 \text{ in}^3 = 7.48051948051948 \text{ gal} = 28.316846592 \text{ liters}$$

$$1 \text{ gallon} = 231 \text{ in}^3 = 4 \text{ quarts} = 8 \text{ pints} = 16 \text{ cups} = 3.785411784 \text{ liters}$$

$$1 \text{ cup} = 8 \text{ floz} = 16 \text{ tablespoons} = 48 \text{ teaspoons}$$

$$1 \text{ tablespoon} = 3 \text{ teaspoons} = 4 \text{ fluidrams}$$

$$1 \text{ dry gallon} = 268.8025 \text{ in}^3 = 4.40488377086 \text{ liters}$$

$$1 \text{ imperial gallon} = 4.54609 \text{ liters}$$

$$1 \text{ bushel} = 4 \text{ pecks} = 8 \text{ dry gallons}$$

Density

$$1 \text{ g/cm}^3 = 1000 \text{ kg/m}^3 = 8.34540445201933 \text{ lb/gal} = 1.043175556502416 \text{ lb/pint}$$

Force

$$1 \text{ lbf} = 4.44822161526050 \text{ newtons} = 32.1740485564304 \text{ poundals}$$

$$1 \text{ newton} = 10^5 \text{ dynes}$$

Energy

$$1 \text{ calorie} = 4.1868 \text{ joules}$$

$$1 \text{ BTU} = 1055.05585262 \text{ joules}$$

$$1 \text{ ft-lb} = 1.35581794833140 \text{ joules}$$

$$1 \text{ kW-hr} = 3.6 \text{ MJ}$$

$$1 \text{ eV} = 1.6021766208 \times 10^{-19} \text{ joules}$$

$$1 \text{ joule} = 10^7 \text{ ergs}$$

Power

$$1 \text{ horsepower} = 745.69987158227022 \text{ watts}$$

$$1 \text{ statwatt} = 1 \text{ abwatt} = 1 \text{ erg/s} = 10^{-7} \text{ watt}$$

Angle

$$\text{rad} = \text{deg} \times \frac{\pi}{180} \quad \text{deg} = \text{rad} \times \frac{180}{\pi}$$

$$1 \text{ deg} = 60 \text{ arcmin} = 3600 \text{ arcsec}$$

Temperature

$$^{\circ}\text{C} = (^{\circ}\text{F} - 32) \times \frac{5}{9} \quad ^{\circ}\text{F} = (^{\circ}\text{C} \times \frac{9}{5}) + 32$$

$$\text{K} = ^{\circ}\text{C} + 273.15$$

$$^{\circ}\text{R} = ^{\circ}\text{F} + 459.67$$

Pressure

$$\begin{aligned} 1 \text{ atm} &= 101325 \text{ Pa} = 1.01325 \text{ bar} = 1013.25 \text{ millibar} = 760 \text{ torr} \\ &= 760 \text{ mmHg} = 29.9212598425197 \text{ inHg} = 14.6959487755134 \text{ psi} \\ &= 2116.21662367394 \text{ lb/ft}^2 = 1.05810831183697 \text{ ton/ft}^2 \\ &= 1013250 \text{ dyne/cm}^2 = 1013250 \text{ barye} \end{aligned}$$

Electromagnetism

$$1 \text{ statcoulomb} = 3.335640951981520 \times 10^{-10} \text{ coulomb}$$

$$1 \text{ abcoulomb} = 10 \text{ coulombs}$$

$$1 \text{ statvolt} = 299.792458 \text{ volts}$$

$$1 \text{ abvolt} = 10^{-8} \text{ volt}$$

$$1 \text{ maxwell} = 10^{-8} \text{ weber}$$

$$1 \text{ gauss} = 10^{-4} \text{ tesla}$$

$$1 \text{ oersted} = 250/\pi (= 79.5774715459477) \text{ A/m}$$

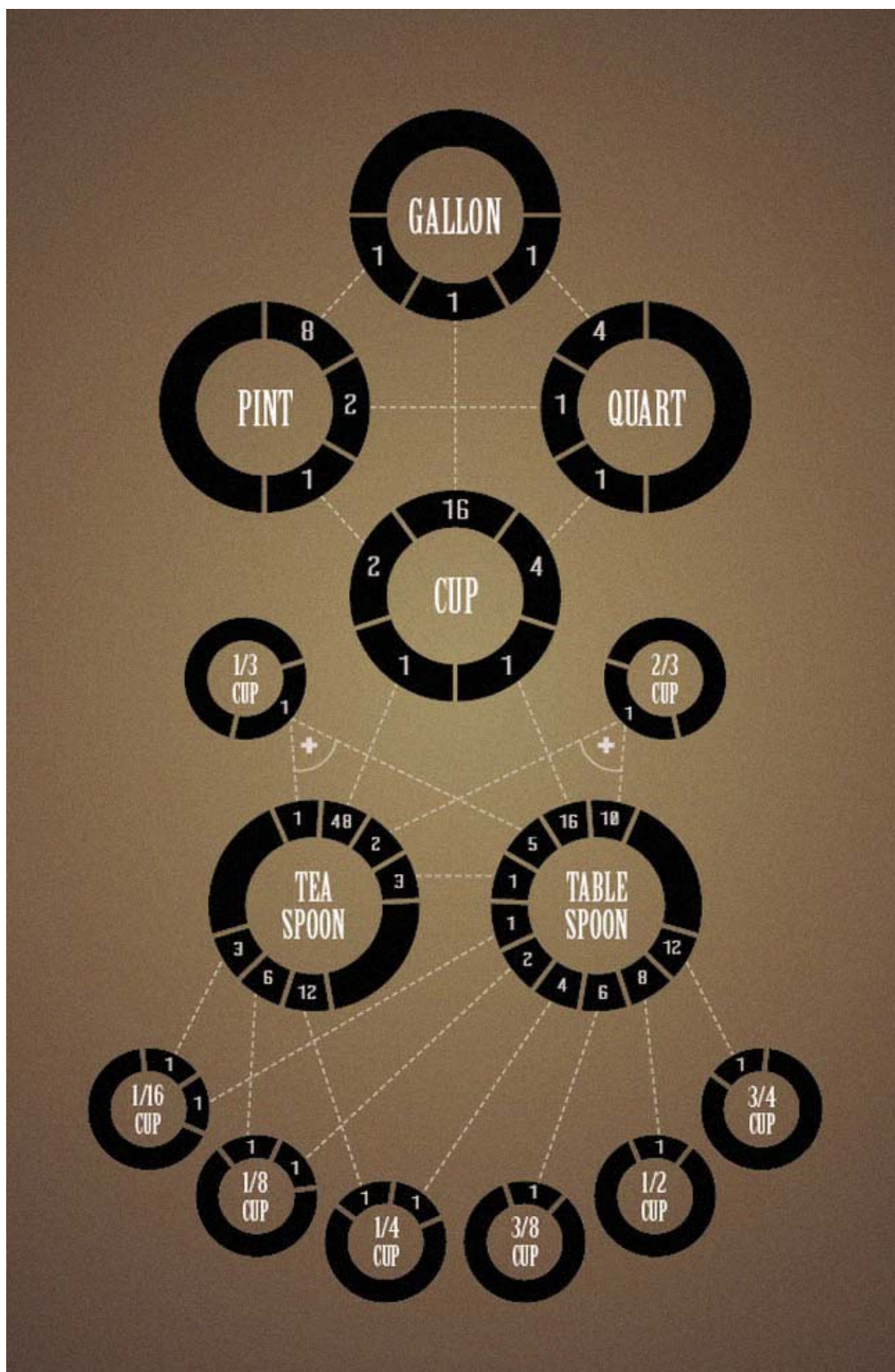


Figure M.1: Conversion chart for kitchen measurements. (Credit: S.B. Lattin Design.)

Appendix N

Angular Measure

N.1 Plane Angle

The most common unit of measure for plane angle is the *degree* (°), which is 1/360 of a full circle. Therefore a circle is 360°, a semicircle is 180°, and a right angle is 90°.

A similar unit (seldom used nowadays) is a sort of “metric” angle called the *grad*, defined so that a right angle is 100 grads, and so a full circle is 400 grads.

The SI unit of plane angle is the *radian* (rad), which is defined to be the angle that subtends an arc length equal to the radius of the circle. By this definition, a full circle subtends an angle equal to the arc length of a full circle ($2\pi r$) divided by its radius r — and so a full circle is 2π radians.

Since a hemisphere is 180° or π radians, the conversion factors are:

$$\text{rad} = \frac{\pi}{180} \times \text{deg} \quad (\text{N.1})$$

$$\text{deg} = \frac{180}{\pi} \times \text{rad} \quad (\text{N.2})$$

Subunits of the Degree

For small angles, a degree may be subdivided into 60 *minutes* (′), and a minute into 60 *seconds* (″). Thus a minute is 1/60 degree, and a second is 1/3600 degree.¹ Angles smaller than 1 second are sometimes expressed as *milli-arcseconds* (1/1000 arcsecond).²

N.2 Solid Angle

A *solid angle* is the three-dimensional version of a plane angle, and is subtended by the vertex of a cone. The SI unit of solid angle is the *steradian* (sr), which is defined to be the solid angle that subtends an area equal to the square of the radius of a circle. By this definition, a full sphere subtends an area equal to the area of a sphere ($4\pi r^2$) divided by the square of its radius (r^2) — so a full sphere is 4π steradians, and a hemisphere is 2π steradians.

¹Sometimes these units are called the *minute of arc* or *arcminute*, and the *second of arc* or *arcsecond* to distinguish them from the units of time that have the same name.

²In an old system (Ref. [14]), the second was further subdivided into 60 *thirds* (″″), the third into 60 *fourths* (″″″), etc. Under this system, 1 milli-arcsecond is 3.6 fourths of arc. This system is no longer used, though; today the second of arc is simply subdivided into decimals (e.g. 32.86473″).

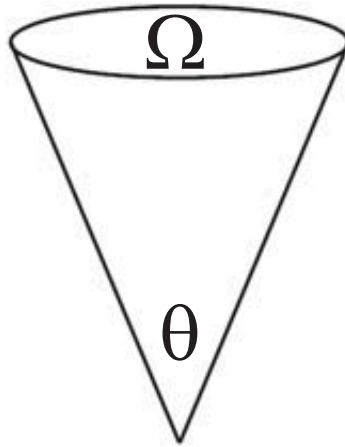


Figure N.1: Relation between plane angle θ and solid angle Ω for a right circular cone.

There is a simple relation between plane angle and solid angle for a right circular cone. If the vertex of the cone subtends an angle θ (the *aperture angle* of the cone), then the corresponding solid angle Ω is (Fig. N.1)

$$\Omega = 2\pi \left(1 - \cos \frac{\theta}{2} \right). \quad (\text{N.3})$$

Another unit of solid angle is the *square degree* (deg^2):

$$\text{sq. deg.} = \text{sr} \times \left(\frac{180}{\pi} \right)^2. \quad (\text{N.4})$$

In these units, a hemisphere is $20,626.48 \text{ deg}^2$, and a complete sphere is $41,252.96 \text{ deg}^2$.

Appendix O

Vector Arithmetic

A vector $\mathbf{A}$ may be written in cartesian (rectangular) form as

$$\mathbf{A} = A_x \mathbf{i} + A_y \mathbf{j} + A_z \mathbf{k}, \quad (\text{O.1})$$

where $\mathbf{i}$ is a *unit vector* (a vector of magnitude 1) in the x direction, $\mathbf{j}$ is a unit vector in the y direction, and $\mathbf{k}$ is a unit vector in the z direction. A_x , A_y , and A_z are called the x , y , and z *components* (respectively) of vector $\mathbf{A}$, and are the projections of the vector onto those axes.

The *magnitude* (“length”) of vector $\mathbf{A}$ is

$$|\mathbf{A}| = A = \sqrt{A_x^2 + A_y^2 + A_z^2}. \quad (\text{O.2})$$

For example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$, then $|\mathbf{A}| = A = \sqrt{3^2 + 5^2 + 2^2} = \sqrt{38}$.

In two dimensions, a vector has no $\mathbf{k}$ component: $\mathbf{A} = A_x \mathbf{i} + A_y \mathbf{j}$.

Addition and Subtraction

To add two vectors, you add their components. Writing a second vector as $\mathbf{B} = B_x \mathbf{i} + B_y \mathbf{j} + B_z \mathbf{k}$, we have

$$\mathbf{A} + \mathbf{B} = (A_x + B_x) \mathbf{i} + (A_y + B_y) \mathbf{j} + (A_z + B_z) \mathbf{k}. \quad (\text{O.3})$$

For example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$ and $\mathbf{B} = 2\mathbf{i} - \mathbf{j} + 4\mathbf{k}$, then $\mathbf{A} + \mathbf{B} = 5\mathbf{i} + 4\mathbf{j} + 6\mathbf{k}$.

Subtraction of vectors is defined similarly:

$$\mathbf{A} - \mathbf{B} = (A_x - B_x) \mathbf{i} + (A_y - B_y) \mathbf{j} + (A_z - B_z) \mathbf{k}. \quad (\text{O.4})$$

For example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$ and $\mathbf{B} = 2\mathbf{i} - \mathbf{j} + 4\mathbf{k}$, then $\mathbf{A} - \mathbf{B} = \mathbf{i} + 6\mathbf{j} - 2\mathbf{k}$.

Scalar Multiplication

To multiply a vector by a scalar, just multiply each component by the scalar. Thus if c is a scalar, then

$$c\mathbf{A} = cA_x \mathbf{i} + cA_y \mathbf{j} + cA_z \mathbf{k}. \quad (\text{O.5})$$

For example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$, then $7\mathbf{A} = 21\mathbf{i} + 35\mathbf{j} + 14\mathbf{k}$.

Dot Product

It is possible to multiply a vector by another vector, but there is more than one kind of multiplication between vectors. One type of vector multiplication is called the *dot product*, in which a vector is multiplied by another vector to give a *scalar* result. The dot product (written with a dot operator, as in $\mathbf{A} \cdot \mathbf{B}$) is

$$\mathbf{A} \cdot \mathbf{B} = AB \cos \theta = A_x B_x + A_y B_y + A_z B_z, \quad (\text{O.6})$$

where θ is the angle between vectors $\mathbf{A}$ and $\mathbf{B}$. For example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$ and $\mathbf{B} = 2\mathbf{i} - \mathbf{j} + 4\mathbf{k}$, then $\mathbf{A} \cdot \mathbf{B} = 6 - 5 + 8 = 9$.

The dot product can be used to find the angle between two vectors. To do this, we solve Eq. (O.6) for θ and find $\cos \theta = \mathbf{A} \cdot \mathbf{B} / (AB)$. Applying this to the previous example, we get $A = \sqrt{38}$ and $B = \sqrt{21}$, so $\cos \theta = 9/(\sqrt{38}\sqrt{21})$, and thus $\theta = 71.4^\circ$.

An immediate consequence of Eq. (O.6) is that two vectors are perpendicular if and only if their dot product is zero.

Cross Product

Another kind of multiplication between vectors, called the *cross product*, involves multiplying one vector by another and giving another *vector* as a result. The cross product is written with a cross operator, as in $\mathbf{A} \times \mathbf{B}$. It is defined by

$$\mathbf{A} \times \mathbf{B} = (AB \sin \theta) \mathbf{u} \quad (\text{O.7})$$

$$= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \quad (\text{O.8})$$

$$= (A_y B_z - A_z B_y) \mathbf{i} - (A_x B_z - A_z B_x) \mathbf{j} + (A_x B_y - A_y B_x) \mathbf{k}, \quad (\text{O.9})$$

where again θ is the angle between the vectors, and $\mathbf{u}$ is a unit vector pointing in a direction perpendicular to the plane containing $\mathbf{A}$ and $\mathbf{B}$, in a right-hand sense: if you curl the fingers of your right hand from $\mathbf{A}$ into $\mathbf{B}$, then the thumb of your right hand points in the direction of $\mathbf{A} \times \mathbf{B}$ (Fig. O.1). As an example, if $\mathbf{A} = 3\mathbf{i} + 5\mathbf{j} + 2\mathbf{k}$ and $\mathbf{B} = 2\mathbf{i} - \mathbf{j} + 4\mathbf{k}$, then $\mathbf{A} \times \mathbf{B} = (20 - (-2))\mathbf{i} - (12 - 4)\mathbf{j} + (-3 - 10)\mathbf{k} = 22\mathbf{i} - 8\mathbf{j} - 13\mathbf{k}$.

Rectangular and Polar Forms

A two-dimensional vector may be written in either *rectangular form* $\mathbf{A} = A_x \mathbf{i} + A_y \mathbf{j}$ described earlier, or in *polar form* $\mathbf{A} = A \angle \theta$, where A is the vector magnitude, and θ is the direction measured counterclockwise from the $+x$ axis. To convert from polar form to rectangular form, one finds

$$A_x = A \cos \theta \quad (\text{O.10})$$

$$A_y = A \sin \theta \quad (\text{O.11})$$

Inverting these equations gives the expressions for converting from rectangular form to polar form:

$$A = \sqrt{A_x^2 + A_y^2} \quad (\text{O.12})$$

$$\tan \theta = \frac{A_y}{A_x} \quad (\text{O.13})$$

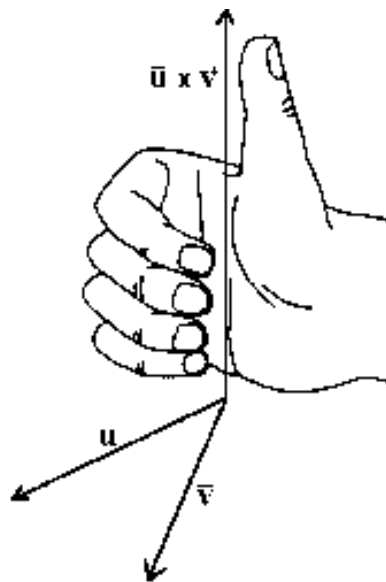


Figure O.1: The vector cross product $\mathbf{A} \times \mathbf{B}$ is perpendicular to the plane of $\mathbf{A}$ and $\mathbf{B}$, and in the right-hand sense. (Credit: "Connected Curriculum Project", Duke University.)

Appendix P

Matrix Properties

This appendix presents a brief summary of the properties of 2×2 and 3×3 matrices.

2×2 Matrices

Determinant

The determinant of a 2×2 matrix is given by the well-known formula:

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc. \quad (\text{P.1})$$

Matrix of Cofactors

The matrix of cofactors is the matrix of signed minors; for a 2×2 matrix, this is

$$\text{cof} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad (\text{P.2})$$

Inverse

Finally, the inverse of a matrix is the transpose of the matrix of cofactors divided by the determinant. For a 2×2 matrix,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad (\text{P.3})$$

3×3 Matrices

Determinant

The determinant of a 3×3 matrix is given by:

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg). \quad (\text{P.4})$$

Matrix of Cofactors

The matrix of cofactors is the matrix of signed minors; for a 3×3 matrix, this is

$$\text{cof} \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = \begin{pmatrix} ei - fh & fg - di & dh - eg \\ ch - bi & ai - cg & bg - ah \\ bf - ce & cd - af & ae - bd \end{pmatrix} \quad (\text{P.5})$$

Inverse

Finally, the inverse of a matrix is the transpose of the matrix of cofactors divided by the determinant. For a 3×3 matrix,

$$\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{a(ei - fh) - b(di - fg) + c(dh - eg)} \begin{pmatrix} ei - fh & ch - bi & bf - ce \\ fg - di & ai - cg & cd - af \\ dh - eg & bg - ah & ae - bd \end{pmatrix} \quad (\text{P.6})$$

Appendix Q

Moments of Inertia

The table below shows the moments of inertia of several common uniform bodies. A very helpful theorem to be used in conjunction with this table is the *parallel axis theorem*, which relates the moment of inertia I_{cm} about an axis A passing through the center of mass to the moment of inertia I about another axis parallel to A . If the two rotation axes are separated by a distance h , then

$$\boxed{I = I_{\text{cm}} + Mh^2} \quad (\text{Q.1})$$

where M is the mass of the body.

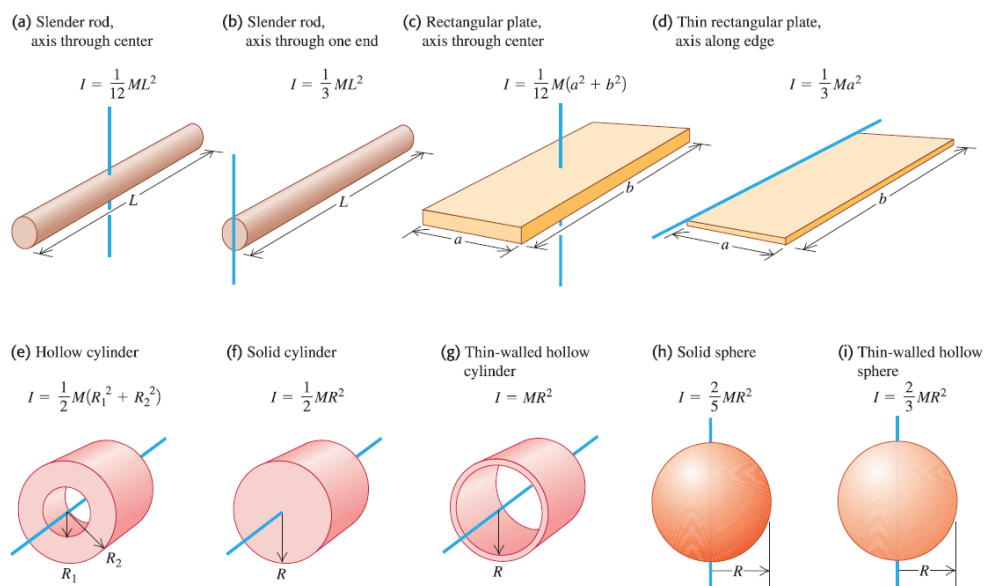


Figure Q.1: Table of moments of inertia of uniform bodies. (Credit: University of Pennsylvania.)

Appendix R

The Simple Plane Pendulum: Exact Solution

The solution to the simple plane pendulum problem described in Chapter 8 is only approximate; here we will examine the *exact* solution, which is surprisingly complicated. We will begin by deriving the differential equation of the motion, then find expressions for the angle θ from the vertical and the period T at any time t . We won't go through the derivations here—we'll just look at the results. Here we'll assume the amplitude of the motion $\theta_0 < \pi$, so that the pendulum does *not* spin in complete circles around the pivot, but simply oscillates back and forth.

The mathematics involved in the exact solution to the pendulum problem is somewhat advanced, but is included here so that you can see that even a very simple physical system can lead to some complicated mathematics.

R.1 Equation of Motion

To derive the differential equation of motion for the pendulum, we begin with Newton's second law in rotational form:

$$\tau = I\alpha = I \frac{d^2\theta}{dt^2}, \quad (\text{R.1})$$

where τ is the torque, I is the moment of inertia, α is the angular acceleration, and θ is the angle from the vertical. In the case of the pendulum, the torque is given by

$$\tau = -mgL \sin \theta, \quad (\text{R.2})$$

and the moment of inertia is

$$I = mL^2. \quad (\text{R.3})$$

Substituting these expressions for τ and I into Eq. (R.1), we get the second-order differential equation

$$-mgL \sin \theta = mL^2 \frac{d^2\theta}{dt^2}, \quad (\text{R.4})$$

which simplifies to give the differential equation of motion,

$$\frac{d^2\theta}{dt^2} = -\frac{g}{L} \sin \theta. \quad (\text{R.5})$$

R.2 Solution, $\theta(t)$

If the amplitude θ_0 is small, we can approximate $\sin \theta \approx \theta$, and find the position $\theta(t)$ at any time t is given by Eq. (8.3) in Chapter 8. But when the amplitude is not necessarily small, the angle θ from the vertical at any time t is found (by solving Eq. (R.5)) to be a more complicated function:

$$\theta(t) = 2 \sin^{-1} \left\{ k \operatorname{sn} \left[\sqrt{\frac{g}{L}} (t - t_0); k \right] \right\}, \quad (\text{R.6})$$

where $\operatorname{sn}(x; k)$ is a *Jacobian elliptic function* with modulus $k = \sin(\theta_0/2)$. The time t_0 is a time at which the pendulum is vertical ($\theta = 0$) and moving in the $+\theta$ direction.

The Jacobian elliptic function is one of a number of so-called “special functions” that often appear in mathematical physics. In this case, the function $\operatorname{sn}(x; k)$ is defined as a kind of inverse of an integral. Given the function

$$u(y; k) = \int_0^y \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \quad (\text{R.7})$$

the Jacobian elliptic function is defined as:

$$\operatorname{sn}(u; k) = y. \quad (\text{R.8})$$

Values of $\operatorname{sn}(x; k)$ may be found in tables of functions or computed by specialized mathematical software libraries.

R.3 Period

As found in Chapter 8, the approximate period of a pendulum for small amplitudes is given by

$$T_0 = 2\pi \sqrt{\frac{L}{g}}. \quad (\text{R.9})$$

This equation is really only an *approximate* expression for the period of a simple plane pendulum; the smaller the amplitude of the motion, the better the approximation. An *exact* expression for the period is given by

$$T = 4 \sqrt{\frac{L}{g}} \int_0^1 \frac{dt}{\sqrt{(1-t^2)(1-k^2t^2)}}, \quad (\text{R.10})$$

which is a type of integral known as a *complete elliptic integral of the first kind*.

The integral in Eq. (R.10) cannot be evaluated in closed form, but it *can* be expanded into an infinite series. The result is

$$T = 2\pi \sqrt{\frac{L}{g}} \left\{ 1 + \sum_{n=1}^{\infty} \left[\frac{(2n-1)!!}{(2n)!!} \right]^2 \sin^{2n} \left(\frac{\theta_0}{2} \right) \right\} \quad (\text{R.11})$$

$$= 2\pi \sqrt{\frac{L}{g}} \left\{ 1 + \sum_{n=1}^{\infty} \left[\frac{(2n)!}{2^{2n}(n!)^2} \right]^2 \sin^{2n} \left(\frac{\theta_0}{2} \right) \right\}. \quad (\text{R.12})$$

We can explicitly write out the first few terms of this series; the result is

$$T = 2\pi \sqrt{\frac{L}{g}} \left[1 + \frac{1}{4} \sin^2\left(\frac{\theta_0}{2}\right) + \frac{9}{64} \sin^4\left(\frac{\theta_0}{2}\right) + \frac{25}{256} \sin^6\left(\frac{\theta_0}{2}\right) + \frac{1225}{16384} \sin^8\left(\frac{\theta_0}{2}\right) + \frac{3969}{65536} \sin^{10}\left(\frac{\theta_0}{2}\right) + \frac{53361}{1048576} \sin^{12}\left(\frac{\theta_0}{2}\right) + \frac{184041}{4194304} \sin^{14}\left(\frac{\theta_0}{2}\right) + \frac{41409225}{1073741824} \sin^{16}\left(\frac{\theta_0}{2}\right) + \frac{147744025}{4294967296} \sin^{18}\left(\frac{\theta_0}{2}\right) + \frac{2133423721}{68719476736} \sin^{20}\left(\frac{\theta_0}{2}\right) + \dots \right]. \quad (\text{R.13})$$

If we wish, we can write out a series expansion for the period in another form—one which does not involve the sine function, but only involves powers of the amplitude θ_0 . To do this, we expand $\sin(\theta_0/2)$ into a Taylor series:

$$\sin \frac{\theta_0}{2} = \sum_{n=1}^{\infty} \frac{(-1)^{n+1} \theta_0^{2n-1}}{2^{2n-1} (2n-1)!} \quad (\text{R.14})$$

$$= \frac{\theta_0}{2} - \frac{\theta_0^3}{48} + \frac{\theta_0^5}{3840} - \frac{\theta_0^7}{645120} + \frac{\theta_0^9}{185794560} - \frac{\theta_0^{11}}{81749606400} + \dots \quad (\text{R.15})$$

Now substitute this series into the series of Eq. (R.11) and collect terms. The result is

$$T = 2\pi \sqrt{\frac{L}{g}} \left(1 + \frac{1}{16} \theta_0^2 + \frac{11}{3072} \theta_0^4 + \frac{173}{737280} \theta_0^6 + \frac{22931}{1321205760} \theta_0^8 + \frac{1319183}{951268147200} \theta_0^{10} + \frac{233526463}{2009078326886400} \theta_0^{12} + \frac{2673857519}{265928913086054400} \theta_0^{14} + \frac{39959591850371}{44931349155019751424000} \theta_0^{16} + \frac{8797116290975003}{109991942731488351485952000} \theta_0^{18} + \frac{4872532317019728133}{668751011807449177034588160000} \theta_0^{20} + \dots \right). \quad (\text{R.16})$$

An entirely different formula for the exact period of a simple plane pendulum has appeared in a recent paper (Adlaj, 2012). According to Adlaj, the exact period of a pendulum may be calculated more efficiently using the *arithmetic-geometric mean*, by means of the formula

$$T = 2\pi \sqrt{\frac{L}{g}} \times \frac{1}{\text{agm}(1, \cos(\theta_0/2))} \quad (\text{R.17})$$

where $\text{agm}(x, y)$ denotes the arithmetic-geometric mean of x and y , which is found by computing the arithmetic and geometric means of x and y , then the arithmetic and geometric mean of those two means, then iterating this process over and over again until the two means converge:

$$a_{n+1} = \frac{a_n + g_n}{2} \quad (\text{R.18})$$

$$g_{n+1} = \sqrt{a_n g_n} \quad (\text{R.19})$$

Here a_n denotes an arithmetic mean, and g_n a geometric mean.

Shown in Fig. R.1 is a plot of the ratio of the pendulum's true period T to its small-angle period T_0 ($T/(2\pi \sqrt{L/g})$) vs. amplitude θ_0 for values of the amplitude between 0 and 180°, using Eq. (R.17). As you can see, the ratio is 1 for small amplitudes (as expected), and increasingly deviates from 1 for large amplitudes. The true period will always be longer than the small-angle period T_0 .

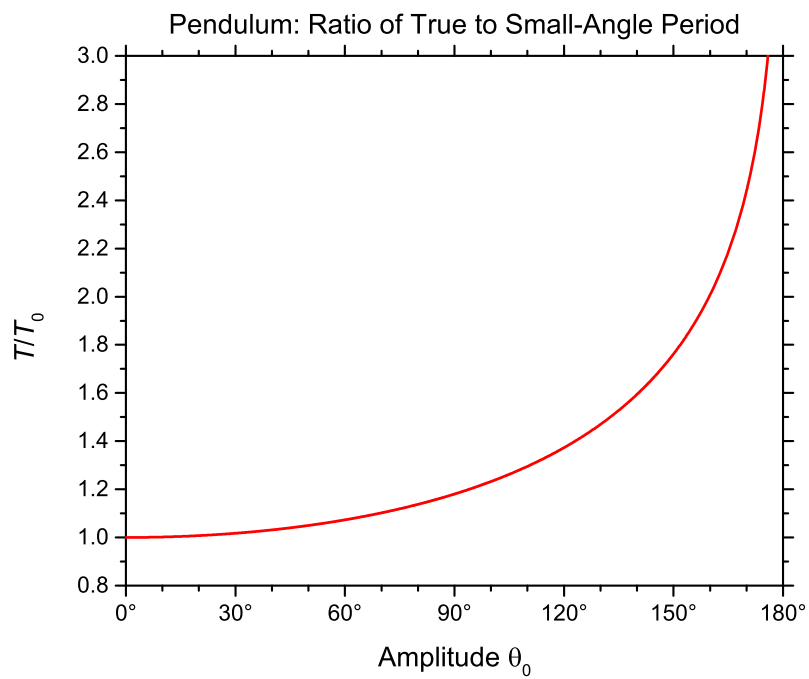


Figure R.1: Ratio of a pendulum's true period T to its small-angle period $T_0 = 2\pi\sqrt{L/g}$, as a function of amplitude θ_0 . For small amplitudes, this ratio is near 1; for larger amplitudes, the true period is longer than predicted by the small-angle approximation.

References

1. L.P. Fulcher and B.F. Davis, "Theoretical and experimental study of the motion of the simple pendulum", *Am. J. Phys.*, **44**, 51 (1976).
2. R.A. Nelson and M.G. Olsson, "The pendulum—Rich physics from a simple system", *Am. J. Phys.*, **54**, 112 (1986).
3. E.T. Whittaker, *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies* (Cambridge, New York, 1937), 4th ed., p. 73.
4. G.L. Baker and J.A. Blackburn, *The Pendulum: A Case Study in Physics* (Oxford, New York, 2005).
5. S. Adlaj, An Eloquent Formula for the Perimeter of an Ellipse. *Notices Amer. Math. Soc.*, **59**, 8, 1094 (September 2012).

Appendix S

CIE Chromaticity Coordinates

In this appendix, we'll look at some of the details of the CIE chromaticity diagram (Fig. 56.4) and how coordinates on the diagram are computed. The mathematics involves the integral calculus, so is outside the usual scope of this course.

Suppose we have a colored object, and we wish to find its coordinates (x, y) on the CIE chromaticity diagram. We begin by measuring the *spectral power distribution* $I(\lambda)$ of the object: this is the fraction I of light reflected from the object at each wavelength (λ) , under some standard illumination conditions. We also need a set of “weighting” functions called the *CIE color matching functions* $(\bar{x}, \bar{y}, \bar{z})$; these are defined as shown in Figure S.1. Then the *tristimulus values* (X, Y, Z) are given by

$$X = \int_0^\infty I(\lambda) \bar{x}(\lambda) d\lambda \quad (\text{S.1})$$

$$Y = \int_0^\infty I(\lambda) \bar{y}(\lambda) d\lambda \quad (\text{S.2})$$

$$Z = \int_0^\infty I(\lambda) \bar{z}(\lambda) d\lambda \quad (\text{S.3})$$

Roughly speaking, X measures the “redness” of the object, Y its “brightness” (or *luminance*), and Z its “blueness.” Normalizing these tristimulus values gives us the coordinates (x, y, z) on the chromaticity diagram:

$$x = \frac{X}{X + Y + Z} \quad (\text{S.4})$$

$$y = \frac{Y}{X + Y + Z} \quad (\text{S.5})$$

$$z = \frac{Z}{X + Y + Z} = 1 - x - y \quad (\text{S.6})$$

$$(\text{S.7})$$

Because of the normalization condition, knowing x and y automatically gives $z = 1 - x - y$; therefore only x and y are needed as the chromaticity coordinates.

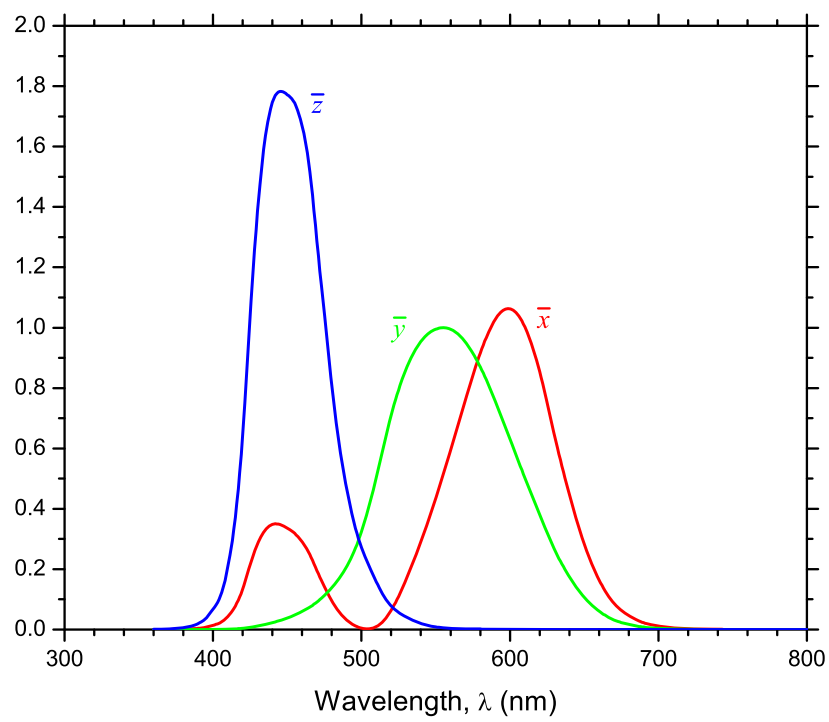


Figure S.1: CIE color matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$.

Appendix T

Calculator Programs

On the class Web site you will find several physics-related programs for a variety of electronic calculator models. The programs are available at:

<http://www.pgccphy.net/1020/software.html>

Contents

- 1. Projectile Problem**
- 2. Kepler's Equation**
- 3. Hyperbolic Kepler's Equation**
- 4. Barker's Equation**
- 5. Reduction of an Angle**
- 6. Helmert's Equation**
- 7. Pendulum Period**
- 8. 1D Perfectly Elastic Collisions**

Appendix U

Right-Hand Rules

- **Vector cross product.** Curl the fingers of your right hand from one vector **A** to a second vector **B**; then your right thumb points in the direction of the cross product $\mathbf{A} \times \mathbf{B}$.
- **Magnetic field in a long wire.** Point the thumb of your right hand in the direction of the current; then the fingers of your right hand curl in the direction of the magnetic field.
- **Magnetic field in a solenoid.** Curl the fingers of your right hand in the direction of the current; then the thumb of your right hand points in the direction of the magnetic field inside the solenoid.
- **Magnetic moment of a coil.** Curl the fingers of your right hand in the direction of the current flowing around the coil; then the thumb of your right hand points in the direction of the magnetic moment.
- **Gyro motion of a negative charge in a magnetic field.** Point the thumb of your right hand in the direction of the magnetic field; then the fingers of your right hand curl in the direction of gyro motion of a *negative* charge.

Appendix V

The Earth's Magnetosphere

The following pages on the Earth's magnetosphere were written by Dr. Sten Odenwald as part of the public education for NASA's IMAGE mission. IMAGE (Imager for Magnetopause-to-Aurora Global Exploration) was an Earth-orbiting spacecraft designed to produce images of various parts of the Earth's magnetosphere. The figures labeled 5-2, 5-3, and 5-4 were taken by the IMAGE spacecraft.

Source: http://solarb.msfc.nasa.gov/for_educators/learn/textbooks.html

9.0 Earth's Magnetism

An ordinary compass works because the Earth is itself a giant magnet with a north and a south pole. Navigators have known about the pole-seeking ability of magnetized compass needles and lodestone for thousands of years. During the last two centuries, much more has been learned about the geomagnetic field and how it shapes the environment of the Earth in space.

The geomagnetic field is believed to be generated by a **magnetic dynamo** process near the core of the Earth through the action of currents in its outer liquid region. Geologic evidence shows that it reverses its polarity every 250,000 to 500,000 years. In fact, the geomagnetic field is decreasing in strength by 5% per century, suggesting that in a few thousand years it may temporarily vanish as the next field reversal begins. Although the geomagnetic field deflects high-energy cosmic rays, past magnetic reversals have not caused obvious biological impacts traceable in the fossil record. Earth's atmosphere, by itself, is very effective in shielding the surface from cosmic rays able to do biological damage. The location of the magnetic poles at the surface also wanders over time at about 10 kilometers per year. Mapmakers periodically update their maps to accommodate this drift.

The domain of space controlled by Earth's magnetic field is called the **magnetosphere**. The geomagnetic field resembles the field of a bar magnet; however, there are important differences due to its interaction with the **solar wind**: an interplanetary flow of plasma from the Sun. The magnetosphere is shaped like a comet with Earth at its head. The field on the day side is compressed inwards by the pressure of the solar wind. A boundary called the **magnetopause** forms about 60,000 kilometers from Earth as the solar wind and geomagnetic field reach an approximate pressure balance. The field on the nightside of Earth is stretched into a long **geomagnetic tail** extending millions of kilometers from Earth. Above the polar regions, magnetic field lines from Earth can connect with field lines from the solar wind forming a **magnetospheric cusp** where plasma and energy from the solar wind may enter. Ionized gases from Earth's upper atmosphere can escape into the magnetosphere through the cusp in gas outflows called **polar fountains**. The magnetosphere is a complex system of circulating currents and changing magnetic

often affected by distant events on the Sun called "space weather." The conveyor belt for the worst of these influences is the ever-changing solar wind itself. Space weather "storms" can trigger changes in the magnetospheric environment, cause spectacular aurora in the polar regions, and lead to satellite damage and even electrical power outages.

9.1 Trapped Particles and Other Plasmas

Within the magnetosphere there are several distinct populations of neutral particles and plasmas. The **Van Allen Radiation Belts** were discovered in 1958 during the early days of the Space Age. The inner belts extend from an altitude of 700 up to 15,000 km and contain very high-energy protons trapped in the geomagnetic field. The outer belt extends 15,000 to 30,000 km and mostly consists of high-energy electrons. Geosynchronous satellites orbit Earth just outside the outer belt. Human space activity is confined to the zone within the inner edge of the inner belt. Space-suited astronauts exposed to the energetic particles in the Van Allen Belts would receive potentially lethal doses of radiation. The particles that make up the Van Allen Belts bounce along the north- and south-directed magnetic field lines to which they are trapped like water flowing in a pipe. At the same time, there is a slow drift of these particles to the west if they are positively charged, or east if they are negatively charged. There are also three additional systems of particles that share much the same space as the Van Allen Belts, but have much lower energies: the geocorona, the plasmasphere, and the ring current.

Extending thousands of kilometers above Earth is the continuation of its tenuous outer atmosphere called the **geocorona**. It is a comparatively cold, uncharged gas of hydrogen and helium atoms whose particles carry little energy. In the geocoronal region, there is a low-energy population of charged particles called the **plasmasphere**, which is a high-altitude extension of the ionosphere. Unlike the geocorona, the plasmasphere is a complex, ever-changing system controlled by electrical currents within the magnetosphere. These changes can cause this region to fill up with particles and empty over the course of hours or days.

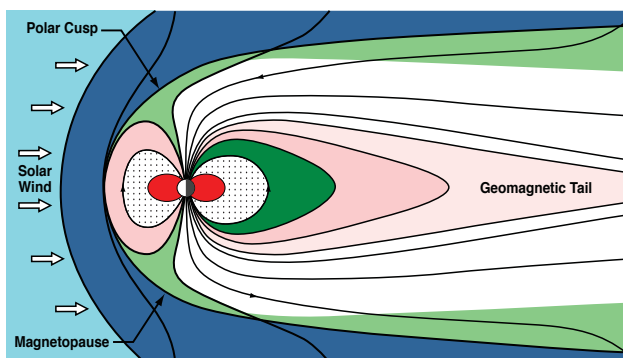


Figure 5-1 Earth's Magnetic Field.

The geomagnetic field resembles the field of an ordinary bar magnet. The north magnetic pole of Earth is located near the south geographic pole, while the south magnetic pole of Earth is located near the north geographic pole. The figure also shows the major regions of Earth's magnetosphere. The filled region shown in red is called the plasmasphere. The dotted region contains the Van Allen Radiation Belts and the ring current. The region shown in green just outside of the ring current zone contains the plasmasheath.

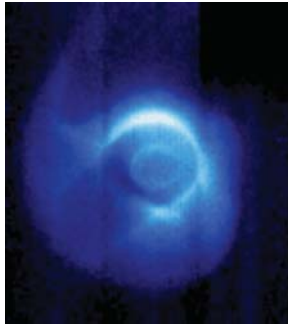


Figure 5-2 The Plasmasphere.
A view from above the North Pole of the plasmasphere illuminated by ultraviolet light from the Sun. The Sun is located beyond the upper right corner.

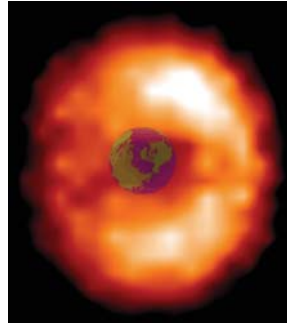


Figure 5-3 The Ring Current.
From above the North Pole, the current is seen flowing around the equatorial regions of the Earth.

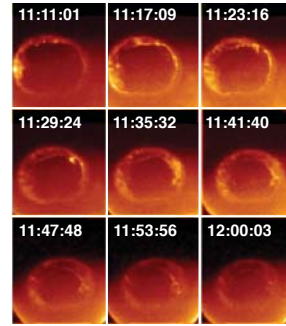


Figure 5-4 The Auroral Oval.
From space, the aurora borealis appears as a ring of light that changes its appearance from minute to minute.

During severe storms, compasses display incorrect bearings as the surface geomagnetic field changes its direction. In the equatorial regions, an actual decrease in the strength of the geomagnetic field can often be measured. This is generally attributed to the existence of a temporary river of charged particles flowing between 30,000 and 60,000 kilometers above ground: the **ring current**. These particles have energies between those within the plasmasphere and those in the Van Allen Belts. They appear to originate within the geomagnetic tail as charged particles that are injected deep into the magnetosphere. Most of the time there are few particles in the ring current, but during severe storms, it fills up with a current of millions of amperes, which spreads into an invisible ring encircling Earth. Just as a flow of current through a wire creates its own magnetic field, the ring current generates a local magnetic field that can reduce some of Earth's surface field by up to 2% over the equatorial regions.

In addition to these families of particles, there are also powerful currents of particles that appear during especially stormy conditions and lead to visually dramatic phenomena called the **aurora borealis** and the **aurora australis**: the northern and southern lights.

9.2 The Aurora

For thousands of years humans have been able to look up at the northern sky and see strange, colorful glows of light. By the early 1900's, spectroscopic studies had shown that auroral light was actually caused by excited oxygen and nitrogen atoms emitting light at only a few specific wavelengths. The source of the excitation was eventually traced to currents of electrons and protons flowing down the geomagnetic field lines into the polar regions where they collide with the atmospheric atoms. However, aurora are not produced directly by solar flares. Radio communications blackouts on the day side of Earth are triggered

by solar flares as these high-energy particles disturb the ionosphere. When directed toward Earth, expulsions of matter by the Sun called **coronal mass ejections** contribute to the conditions that cause some of the strongest aurora to light up the skies. At other times, a simple change in magnetic polarity of the solar wind from north-directed to south-directed seems to be enough to trigger aurora without any obvious solar disturbance.

Because of the existence of the magnetospheric cusp on the day side of Earth, solar wind particles can, under some conditions, flow down this entryway into the polar regions. This causes daytime aurora, and the diffuse red glows of night time aurora. This is, virtually, the only instance where solar wind particles can directly cause aurora. It is not, however, the cause of the spectacular nighttime polar aurora that are so commonly photographed. To understand how these aurora are produced, it is helpful to imagine yourself living inside a television picture tube. We don't see the currents of electrons guided by magnetic forces, but we do see them paint serpentine pictures on the atmosphere, which we then see as the aurora. The origin of these currents is in the distant geomagnetic tail region, not in the direct inflow of solar wind plasma.

When the polarity of the solar wind's magnetic field turns southward, its lines of force encounter the north-directed lines in Earth's equatorial regions on the dayside. The solar wind field lines then connect with Earth's field in a complex event that transfers particles and energy into Earth's magnetosphere. While this is happening near Earth, in the distant geomagnetic tail, other changes are causing the geomagnetic field to stretch like rubber bands and snap into new magnetic shapes. This causes billions of watts of energy to be transferred into the particles already trapped in the magnetosphere out in these distant regions. These particles, boosted in energy by thousands of volts, then flow down the field lines into the polar regions to cause the aurora, like the electrons in a television picture tube that paint a pattern on the phosphor screen.

Appendix W

Round-Number Handbook of Physics

The one-page *Round Number Handbook of Physics* on the following page is by Edward M. Purcell of Harvard University, and appeared in the January 1983 issue of the *American Journal of Physics*. It is intended as a brief reference for doing quick “back of the envelope”, order-of-magnitude calculations.

ROUND-NUMBER HANDBOOK OF PHYSICS

CONSTANTS

$$\begin{aligned}
 c &= 3 \times 10^{10} \text{ cm s}^{-1} \\
 \hbar &= 10^{-27} \text{ erg s} \\
 N_0 &= 6 \times 10^{23} \text{ mole}^{-1} \\
 n_0 &= 3 \times 10^{19} \text{ cm}^{-3} \\
 g &= 10^3 \text{ cm s}^{-2} \\
 e &= 4.8 \times 10^{-10} \text{ esu} \\
 &= 1.6 \times 10^{-19} \text{ C} \\
 k &= 1.4 \times 10^{-16} \text{ erg deg}^{-1} \\
 \alpha &= e^2/\hbar c = 1/137 \\
 (\mu_0/\epsilon_0)^{1/2} &= 377 \Omega \\
 G &= 7 \times 10^{-8} \text{ g cm}^{-4} \text{ s}^{-2} \\
 \mu_0 &= 4\pi \times 10^{-7} \text{ N A}^{-2} \\
 \epsilon_0 &= 8.8 \times 10^{-12} \text{ N}^{-1} \text{ A}^2 \text{ m}^{-2} \text{ s}^2 \\
 R &= 2 \text{ cal/mole deg}
 \end{aligned}$$

CONVERSIONS

$$\begin{aligned}
 1 \text{ cal} &= 4 \text{ J} = 4 \times 10^7 \text{ erg} \\
 1 \text{ N} &= 10^5 \text{ dyn} \\
 680 \text{ lumens} &= 1 \text{ W (5550 Å)} \\
 1 \text{ ft} &= 30 \text{ cm} \\
 1 \text{ lb} &= 4.4 \text{ N} \\
 1 \text{ ci} &= 4 \times 10^{10} \text{ disint/s} \\
 1 \text{ eV} &= 1.6 \times 10^{-12} \text{ erg} \\
 1 \Omega^{-1} &= 9 \times 10^{11} \text{ cm/s} \\
 \text{pc(eV)} &= 300 \text{ Br(G cm)}
 \end{aligned}$$

MASSES

$$\begin{aligned}
 m_e &= 10^{-27} \text{ g} \\
 m_{\text{pion}} &= 270 m_e \\
 m_{\text{kaon}} &= 1000 m_e \\
 m_{\text{nucleon}} &= 2000 m_e \\
 m_e c^2 &= 0.5 \text{ MeV} \\
 m_{\text{muon}} &= 200 m_e
 \end{aligned}$$

USEFUL NUMBERS

$$\begin{aligned}
 \text{classical electron radius} &= r_0 = e^2/m_e c^2 = 3 \times 10^{-13} \text{ cm} \\
 \text{Bohr radius} &= a_0 = \hbar^2/m_e e^2 = 5 \times 10^{-9} \text{ cm} \\
 \text{Rydberg wavelength} &= \lambda_R = \hbar^3 c/m_e e^4 = 7 \times 10^{-7} \text{ cm} \\
 \text{Compton wavelength} &= \lambda_c = \hbar/m_e c = 4 \times 10^{-11} \text{ cm} \\
 \text{Bohr magneton} &= e\hbar/2mc = 10^{-20} \text{ erg/G} \\
 \text{Stefan-Boltzman const} &= 6 \times 10^{-12} \text{ W/deg}^4 \text{ cm}^2 \\
 \text{Min. ionization loss} &= 2 \text{ MeV/g cm}^2 \\
 kT_{\text{room}} &= 0.025 \text{ eV} \\
 R_{\text{nuclear}} &= A^{1/3} \times 10^{-13} \text{ cm} \\
 e^2/a_0 &= 26 \text{ eV}
 \end{aligned}$$

$$h\nu(\text{visible}) = 2 \text{ eV}$$

$$\text{Band gaps: Si} = 1.1 \text{ eV; Ge} = 0.7 \text{ eV}$$

$$\text{Spin precession: } e: 3 \text{ MHz/G; } p: 4 \text{ kHz/G}$$

MATERIALS

$$\begin{aligned}
 \text{Resistivities in } \Omega \text{ cm: Cu: } 2 \times 10^{-6} \text{ (room temp.)} \\
 \text{H}_2\text{O(pure): } 2 \times 10^7; \text{ seawater: } 25 \Omega \text{ cm} \\
 \text{Specific heat (solid or liquid)} &= 0.5 \text{ cal/cm}^3 \text{ deg} \\
 \text{Linear expansion (solid or liquid)} &= 2 \times 10^{-5}/\text{deg} \\
 \text{Heat conduction (insulator)} &= 10^{-2} \text{ cal/s cm deg} \\
 \text{(metal)} &= 1.0(\rho_{\text{Cu}}/\rho_{\text{metal}}) \text{ cal/s cm deg} \\
 \text{Heat of combustion (food or fuel)} &= 10^4 \text{ cal/g} \\
 \text{Heat of vaporization} &= 10^4 \text{ cal/mole} \\
 \text{Elastic moduli (solids)} &= 10^{11}-10^{12} \text{ dyn/cm}^2 \\
 \text{Tensile strength (solids)} &= 10^8-10^{10} \text{ dyn/cm}^2 \\
 \text{Surface tension: H}_2\text{O} &= 50 \text{ dyn/cm} \\
 \text{Diffusion: H}_2\text{O } 10^{-5}, \text{ air: } 0.2 \text{ cm}^2/\text{s} \\
 \text{Viscosity: H}_2\text{O } 10^{-2}, \text{ air: } 2 \times 10^{-4} \text{ dyn s/cm}^2
 \end{aligned}$$

ASTRONOMICAL

$$\begin{aligned}
 1 \text{ pc} &= 3 \times 10^{18} \text{ cm} \\
 1 \text{ mag} &= -4 \text{ dB} \\
 m_{\text{abs}} &= m \text{ at } 10 \text{ pc} \\
 m_{\text{abs}}(\text{sun}) &= +5 \\
 B_{\text{Earth}}(\text{pole}) &= 0.5 \text{ G} \\
 M_{\text{Earth}} &= 6 \times 10^{27} \text{ g} \\
 R_{\text{Earth}} &= 6 \times 10^8 \text{ cm} \\
 M_{\odot} &= 2 \times 10^{33} \text{ g} \\
 R_{\odot} &= 8 \times 10^{10} \text{ cm} \\
 L_{\odot} &= 2 \times 10^{33} \text{ erg/s} = 1 \text{ kW/m}^2 \text{ at Earth} \\
 r_{\text{moon}} &= 4 \times 10^{10} \text{ cm} \\
 r_{\text{sun}} &= 1 \text{ AU} = 1.5 \times 10^{13} \text{ cm} \\
 M_{\text{Galaxy}} &= 2 \times 10^{44} \text{ g} \\
 \text{Distance to center of galaxy} &= 3 \times 10^{22} \text{ cm} \\
 \text{Distance between galaxies} &= 10^{25} \text{ cm} \\
 \text{Energy density: starlight} &= 10^{-12} \text{ erg/cm}^3 \\
 \text{Primary cosmic rays: } &1/\text{cm}^2 \text{ s} \\
 R_{\text{Universe}} &= 3000 \text{ Mpc}
 \end{aligned}$$

ATMOSPHERE (STP)

$$\begin{aligned}
 P_{\text{atm}} &= 10^6 \text{ dyn/cm}^2 = 15 \text{ psi} \\
 V_{\text{sound}} &= V_{\text{molec}} = 4 \times 10^4 \text{ cm/s} \\
 \text{Radiation length} &= 36 \text{ g/cm}^2 \\
 \text{Density} &= 10^{-3} \text{ g/cm}^3 \\
 \text{Mean free path} &= 7 \times 10^{-6} \text{ cm} \\
 \text{Scale height} &= 8 \text{ km}
 \end{aligned}$$

Appendix X

Short Glossary of Particle Physics

antineutron, the antimatter counterpart of the neutron.

antiproton, the antimatter counterpart of the proton.

baryon, a particle made up of three quarks.

boson, any particle that has integer spin.

electron, a lepton of negative charge, found to surround the atomic nucleus in atoms of ordinary matter.

fermion, any particle that has half-integer spin.

hadron, any particle that “feels” the strong nuclear force.

Higgs boson, the particle associated with the Higgs field, that gives mass to other particles.

lepton, one of six light fundamental particles: e^- , ν_e^0 , μ^- , ν_μ^0 , τ^- , ν_τ^0 .

meson, a particle consisting of a quark-antiquark pair.

neutrino, an uncharged lepton of very light mass, produced in some reactions.

neutron, an uncharged baryon, found in the nucleus of atoms of ordinary matter.

positron, the antimatter counterpart of the electron.

proton, a baryon of positive charge, found in the nucleus of atoms of ordinary matter.

quark, one of six heavy fundamental particles: u , d , c , s , t , b .

vector boson, a particle responsible for mediating a force.

Appendix Y

Fundamental Physical Constants — Extensive Listing

The following tables, published by the National Institutes of Science and Technology (NIST), give the current best estimates of a large number of fundamental physical constants. These values were determined by the Committee on Data for Science and Technology (CODATA) for 2014, and are a best fit of the constants to the latest experimental results.

Source: <http://physics.nist.gov/constants>

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
UNIVERSAL				
speed of light in vacuum	c, c_0	299 792 458	m s^{-1}	exact
magnetic constant	μ_0	$4\pi \times 10^{-7}$	N A^{-2}	exact
		$= 12.566\,370\,614... \times 10^{-7}$	N A^{-2}	
electric constant $1/\mu_0 c^2$	ϵ_0	$8.854\,187\,817... \times 10^{-12}$	F m^{-1}	exact
characteristic impedance of vacuum $\mu_0 c$	Z_0	376.730 313 461...	Ω	exact
Newtonian constant of gravitation	G	$6.674\,08(31) \times 10^{-11}$	$\text{m}^3 \text{kg}^{-1} \text{s}^{-2}$	4.7×10^{-5}
		$6.708\,61(31) \times 10^{-39}$	$(\text{GeV}/c^2)^{-2}$	4.7×10^{-5}
Planck constant	h	$6.626\,070\,040(81) \times 10^{-34}$	J s	1.2×10^{-8}
		$4.135\,667\,662(25) \times 10^{-15}$	eV s	6.1×10^{-9}
		$h/2\pi$	$\hbar$	$1.054\,571\,800(13) \times 10^{-34}$
$6.582\,119\,514(40) \times 10^{-16}$	eV s			6.1×10^{-9}
Planck mass $(\hbar c/G)^{1/2}$	m_{P}	197.326 9788(12)	MeV fm	6.1×10^{-9}
		$2.176\,470(51) \times 10^{-8}$	kg	2.3×10^{-5}
energy equivalent	$m_{\text{P}} c^2$	$1.220\,910(29) \times 10^{19}$	GeV	2.3×10^{-5}
Planck temperature $(\hbar c^5/G)^{1/2}/k$	T_{P}	$1.416\,808(33) \times 10^{32}$	K	2.3×10^{-5}
Planck length $\hbar/m_{\text{P}} c = (\hbar G/c^3)^{1/2}$	l_{P}	$1.616\,229(38) \times 10^{-35}$	m	2.3×10^{-5}
Planck time $l_{\text{P}}/c = (\hbar G/c^5)^{1/2}$	t_{P}	$5.391\,16(13) \times 10^{-44}$	s	2.3×10^{-5}
ELECTROMAGNETIC				
elementary charge	e	$1.602\,176\,6208(98) \times 10^{-19}$	C	6.1×10^{-9}
		$2.417\,989\,262(15) \times 10^{14}$	A J^{-1}	6.1×10^{-9}
magnetic flux quantum $h/2e$	Φ_0	$2.067\,833\,831(13) \times 10^{-15}$	Wb	6.1×10^{-9}
conductance quantum $2e^2/h$	G_0	$7.748\,091\,7310(18) \times 10^{-5}$	S	2.3×10^{-10}
inverse of conductance quantum	G_0^{-1}	12 906.403 7278(29)	Ω	2.3×10^{-10}
Josephson constant ¹ $2e/h$	K_{J}	$483\,597.8525(30) \times 10^9$	Hz V^{-1}	6.1×10^{-9}
von Klitzing constant ² $h/e^2 = \mu_0 c/2\alpha$	R_{K}	25 812.807 4555(59)	Ω	2.3×10^{-10}
Bohr magneton $e\hbar/2m_{\text{e}}$	μ_{B}	$927.400\,9994(57) \times 10^{-26}$	J T^{-1}	6.2×10^{-9}
		$5.788\,381\,8012(26) \times 10^{-5}$	eV T^{-1}	4.5×10^{-10}
	μ_{B}/h	$13.996\,245\,042(86) \times 10^9$	Hz T^{-1}	6.2×10^{-9}
	μ_{B}/hc	46.686 448 14(29)	$\text{m}^{-1} \text{T}^{-1}$	6.2×10^{-9}
	μ_{B}/k	0.671 714 05(39)	K T^{-1}	5.7×10^{-7}
	μ_{N}	$5.050\,783\,699(31) \times 10^{-27}$	J T^{-1}	6.2×10^{-9}
		$3.152\,451\,2550(15) \times 10^{-8}$	eV T^{-1}	4.6×10^{-10}
	μ_{N}/h	7.622 593 285(47)	MHz T^{-1}	6.2×10^{-9}
nuclear magneton $e\hbar/2m_{\text{p}}$	μ_{N}/hc	$2.542\,623\,432(16) \times 10^{-2}$	$\text{m}^{-1} \text{T}^{-1}$	6.2×10^{-9}
		$3.658\,2690(21) \times 10^{-4}$	K T^{-1}	5.7×10^{-7}
ATOMIC AND NUCLEAR				
General				
fine-structure constant $e^2/4\pi\epsilon_0\hbar c$	α	$7.297\,352\,5664(17) \times 10^{-3}$		2.3×10^{-10}
		137.035 999 139(31)		2.3×10^{-10}
inverse fine-structure constant	α^{-1}	137.035 999 139(31)		2.3×10^{-10}
Rydberg constant $\alpha^2 m_{\text{e}} c/2h$	R_{∞}	10 973 731.568 508(65)	m^{-1}	5.9×10^{-12}
	$R_{\infty} c$	$3.289\,841\,960\,355(19) \times 10^{15}$	Hz	5.9×10^{-12}
	$R_{\infty} hc$	$2.179\,872\,325(27) \times 10^{-18}$	J	1.2×10^{-8}
		13.605 693 009(84)	eV	6.1×10^{-9}
Bohr radius $\alpha/4\pi R_{\infty} = 4\pi\epsilon_0\hbar^2/m_{\text{e}} e^2$	a_0	$0.529\,177\,210\,67(12) \times 10^{-10}$	m	2.3×10^{-10}
Hartree energy $e^2/4\pi\epsilon_0 a_0 = 2R_{\infty} hc = \alpha^2 m_{\text{e}} c^2$	E_{h}	$4.359\,744\,650(54) \times 10^{-18}$	J	1.2×10^{-8}
		27.211 386 02(17)	eV	6.1×10^{-9}
quantum of circulation	$h/2m_{\text{e}}$	$3.636\,947\,5486(17) \times 10^{-4}$	$\text{m}^2 \text{s}^{-1}$	4.5×10^{-10}

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
	h/m_e	$7.273\,895\,0972(33) \times 10^{-4}$	$\text{m}^2 \text{s}^{-1}$	4.5×10^{-10}
	Electroweak			
Fermi coupling constant ³	$G_F/(\hbar c)^3$	$1.166\,3787(6) \times 10^{-5}$	GeV^{-2}	5.1×10^{-7}
weak mixing angle ⁴ θ_W (on-shell scheme)				
$\sin^2 \theta_W = s_W^2 \equiv 1 - (m_W/m_Z)^2$	$\sin^2 \theta_W$	$0.2223(21)$		9.5×10^{-3}
	Electron, e^-			
electron mass	m_e	$9.109\,383\,56(11) \times 10^{-31}$	kg	1.2×10^{-8}
		$5.485\,799\,090\,70(16) \times 10^{-4}$	u	2.9×10^{-11}
energy equivalent	$m_e c^2$	$8.187\,105\,65(10) \times 10^{-14}$	J	1.2×10^{-8}
		$0.510\,998\,9461(31)$	MeV	6.2×10^{-9}
electron-muon mass ratio	m_e/m_μ	$4.836\,331\,70(11) \times 10^{-3}$		2.2×10^{-8}
electron-tau mass ratio	m_e/m_τ	$2.875\,92(26) \times 10^{-4}$		9.0×10^{-5}
electron-proton mass ratio	m_e/m_p	$5.446\,170\,213\,52(52) \times 10^{-4}$		9.5×10^{-11}
electron-neutron mass ratio	m_e/m_n	$5.438\,673\,4428(27) \times 10^{-4}$		4.9×10^{-10}
electron-deuteron mass ratio	m_e/m_d	$2.724\,437\,107\,484(96) \times 10^{-4}$		3.5×10^{-11}
electron-triton mass ratio	m_e/m_t	$1.819\,200\,062\,203(84) \times 10^{-4}$		4.6×10^{-11}
electron-helion mass ratio	m_e/m_h	$1.819\,543\,074\,854(88) \times 10^{-4}$		4.9×10^{-11}
electron to alpha particle mass ratio	m_e/m_α	$1.370\,933\,554\,798(45) \times 10^{-4}$		3.3×10^{-11}
electron charge to mass quotient	$-e/m_e$	$-1.758\,820\,024(11) \times 10^{11}$	C kg^{-1}	6.2×10^{-9}
electron molar mass $N_A m_e$	$M(e), M_e$	$5.485\,799\,090\,70(16) \times 10^{-7}$	kg mol^{-1}	2.9×10^{-11}
Compton wavelength $h/m_e c$	λ_C	$2.426\,310\,2367(11) \times 10^{-12}$	m	4.5×10^{-10}
$\lambda_C/2\pi = \alpha a_0 = \alpha^2/4\pi R_\infty$	λ_C	$386.159\,267\,64(18) \times 10^{-15}$	m	4.5×10^{-10}
classical electron radius $\alpha^2 a_0$	r_e	$2.817\,940\,3227(19) \times 10^{-15}$	m	6.8×10^{-10}
Thomson cross section $(8\pi/3)r_e^2$	σ_e	$0.665\,245\,871\,58(91) \times 10^{-28}$	m^2	1.4×10^{-9}
electron magnetic moment	μ_e	$-928.476\,4620(57) \times 10^{-26}$	J T^{-1}	6.2×10^{-9}
to Bohr magneton ratio	μ_e/μ_B	$-1.001\,159\,652\,180\,91(26)$		2.6×10^{-13}
to nuclear magneton ratio	μ_e/μ_N	$-1838.281\,972\,34(17)$		9.5×10^{-11}
electron magnetic moment anomaly $ \mu_e /\mu_B - 1$	a_e	$1.159\,652\,180\,91(26) \times 10^{-3}$		2.3×10^{-10}
electron g -factor $-2(1 + a_e)$	g_e	$-2.002\,319\,304\,361\,82(52)$		2.6×10^{-13}
electron-muon magnetic moment ratio	μ_e/μ_μ	$206.766\,9880(46)$		2.2×10^{-8}
electron-proton magnetic moment ratio	μ_e/μ_p	$-658.210\,6866(20)$		3.0×10^{-9}
electron to shielded proton magnetic moment ratio (H_2O , sphere, 25°C)	μ_e/μ'_p	$-658.227\,5971(72)$		1.1×10^{-8}
electron-neutron magnetic moment ratio	μ_e/μ_n	$960.920\,50(23)$		2.4×10^{-7}
electron-deuteron magnetic moment ratio	μ_e/μ_d	$-2143.923\,499(12)$		5.5×10^{-9}
electron to shielded helion magnetic moment ratio (gas, sphere, 25°C)	μ_e/μ'_h	$864.058\,257(10)$		1.2×10^{-8}
electron gyromagnetic ratio $2 \mu_e /\hbar$	γ_e	$1.760\,859\,644(11) \times 10^{11}$	$\text{s}^{-1} \text{T}^{-1}$	6.2×10^{-9}
	$\gamma_e/2\pi$	$28\,024.951\,64(17)$	MHz T^{-1}	6.2×10^{-9}
	Muon, μ^-			
muon mass	m_μ	$1.883\,531\,594(48) \times 10^{-28}$	kg	2.5×10^{-8}
		$0.113\,428\,9257(25)$	u	2.2×10^{-8}
energy equivalent	$m_\mu c^2$	$1.692\,833\,774(43) \times 10^{-11}$	J	2.5×10^{-8}
		$105.658\,3745(24)$	MeV	2.3×10^{-8}
muon-electron mass ratio	m_μ/m_e	$206.768\,2826(46)$		2.2×10^{-8}
muon-tau mass ratio	m_μ/m_τ	$5.946\,49(54) \times 10^{-2}$		9.0×10^{-5}
muon-proton mass ratio	m_μ/m_p	$0.112\,609\,5262(25)$		2.2×10^{-8}

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
muon-neutron mass ratio	m_μ/m_n	0.112 454 5167(25)		2.2×10^{-8}
muon molar mass $N_A m_\mu$	$M(\mu), M_\mu$	$0.113\,428\,9257(25) \times 10^{-3}$	kg mol ⁻¹	2.2×10^{-8}
muon Compton wavelength $h/m_\mu c$	$\lambda_{C,\mu}$	$11.734\,441\,11(26) \times 10^{-15}$	m	2.2×10^{-8}
$\lambda_{C,\mu}/2\pi$	$\lambda_{C,\mu}$	$1.867\,594\,308(42) \times 10^{-15}$	m	2.2×10^{-8}
muon magnetic moment	μ_μ	$-4.490\,448\,26(10) \times 10^{-26}$	J T ⁻¹	2.3×10^{-8}
to Bohr magneton ratio	μ_μ/μ_B	$-4.841\,970\,48(11) \times 10^{-3}$		2.2×10^{-8}
to nuclear magneton ratio	μ_μ/μ_N	$-8.890\,597\,05(20)$		2.2×10^{-8}
muon magnetic moment anomaly				
$ \mu_\mu /(e\hbar/2m_\mu) - 1$	a_μ	$1.165\,920\,89(63) \times 10^{-3}$		5.4×10^{-7}
muon g -factor $-2(1 + a_\mu)$	g_μ	$-2.002\,331\,8418(13)$		6.3×10^{-10}
muon-proton magnetic moment ratio	μ_μ/μ_p	$-3.183\,345\,142(71)$		2.2×10^{-8}
Tau, τ^-				
tau mass ⁵	m_τ	$3.167\,47(29) \times 10^{-27}$	kg	9.0×10^{-5}
		1.907 49(17)	u	9.0×10^{-5}
energy equivalent	$m_\tau c^2$	$2.846\,78(26) \times 10^{-10}$	J	9.0×10^{-5}
		1776.82(16)	MeV	9.0×10^{-5}
tau-electron mass ratio	m_τ/m_e	3477.15(31)		9.0×10^{-5}
tau-muon mass ratio	m_τ/m_μ	16.8167(15)		9.0×10^{-5}
tau-proton mass ratio	m_τ/m_p	1.893 72(17)		9.0×10^{-5}
tau-neutron mass ratio	m_τ/m_n	1.891 11(17)		9.0×10^{-5}
tau molar mass $N_A m_\tau$	$M(\tau), M_\tau$	$1.907\,49(17) \times 10^{-3}$	kg mol ⁻¹	9.0×10^{-5}
tau Compton wavelength $h/m_\tau c$	$\lambda_{C,\tau}$	$0.697\,787(63) \times 10^{-15}$	m	9.0×10^{-5}
$\lambda_{C,\tau}/2\pi$	$\lambda_{C,\tau}$	$0.111\,056(10) \times 10^{-15}$	m	9.0×10^{-5}
Proton, p				
proton mass	m_p	$1.672\,621\,898(21) \times 10^{-27}$	kg	1.2×10^{-8}
		1.007 276 466 879(91)	u	9.0×10^{-11}
energy equivalent	$m_p c^2$	$1.503\,277\,593(18) \times 10^{-10}$	J	1.2×10^{-8}
		938.272 0813(58)	MeV	6.2×10^{-9}
proton-electron mass ratio	m_p/m_e	1836.152 673 89(17)		9.5×10^{-11}
proton-muon mass ratio	m_p/m_μ	8.880 243 38(20)		2.2×10^{-8}
proton-tau mass ratio	m_p/m_τ	0.528 063(48)		9.0×10^{-5}
proton-neutron mass ratio	m_p/m_n	0.998 623 478 44(51)		5.1×10^{-10}
proton charge to mass quotient	e/m_p	$9.578\,833\,226(59) \times 10^7$	C kg ⁻¹	6.2×10^{-9}
proton molar mass $N_A m_p$	$M(p), M_p$	$1.007\,276\,466\,879(91) \times 10^{-3}$	kg mol ⁻¹	9.0×10^{-11}
proton Compton wavelength $h/m_p c$	$\lambda_{C,p}$	$1.321\,409\,853\,96(61) \times 10^{-15}$	m	4.6×10^{-10}
$\lambda_{C,p}/2\pi$	$\lambda_{C,p}$	$0.210\,308\,910\,109(97) \times 10^{-15}$	m	4.6×10^{-10}
proton rms charge radius	r_p	$0.8751(61) \times 10^{-15}$	m	7.0×10^{-3}
proton magnetic moment	μ_p	$1.410\,606\,7873(97) \times 10^{-26}$	J T ⁻¹	6.9×10^{-9}
to Bohr magneton ratio	μ_p/μ_B	$1.521\,032\,2053(46) \times 10^{-3}$		3.0×10^{-9}
to nuclear magneton ratio	μ_p/μ_N	2.792 847 3508(85)		3.0×10^{-9}
proton g -factor $2\mu_p/\mu_N$	g_p	5.585 694 702(17)		3.0×10^{-9}
proton-neutron magnetic moment ratio	μ_p/μ_n	-1.459 898 05(34)		2.4×10^{-7}
shielded proton magnetic moment	μ'_p	$1.410\,570\,547(18) \times 10^{-26}$	J T ⁻¹	1.3×10^{-8}
(H ₂ O, sphere, 25 °C)				
to Bohr magneton ratio	μ'_p/μ_B	$1.520\,993\,128(17) \times 10^{-3}$		1.1×10^{-8}
to nuclear magneton ratio	μ'_p/μ_N	2.792 775 600(30)		1.1×10^{-8}
proton magnetic shielding correction				
$1 - \mu'_p/\mu_p$ (H ₂ O, sphere, 25 °C)	σ'_p	$25.691(11) \times 10^{-6}$		4.4×10^{-4}

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
proton gyromagnetic ratio $2\mu_p/\hbar$	γ_p	$2.675\,221\,900(18) \times 10^8$	$\text{s}^{-1} \text{T}^{-1}$	6.9×10^{-9}
	$\gamma_p/2\pi$	42.577 478 92(29)	MHz T^{-1}	6.9×10^{-9}
shielded proton gyromagnetic ratio $2\mu'_p/\hbar$ (H_2O , sphere, 25°C)	γ'_p	$2.675\,153\,171(33) \times 10^8$	$\text{s}^{-1} \text{T}^{-1}$	1.3×10^{-8}
	$\gamma'_p/2\pi$	42.576 385 07(53)	MHz T^{-1}	1.3×10^{-8}
Neutron, n				
neutron mass	m_n	$1.674\,927\,471(21) \times 10^{-27}$	kg	1.2×10^{-8}
		1.008 664 915 88(49)	u	4.9×10^{-10}
energy equivalent	$m_n c^2$	$1.505\,349\,739(19) \times 10^{-10}$	J	1.2×10^{-8}
		939.565 4133(58)	MeV	6.2×10^{-9}
neutron-electron mass ratio	m_n/m_e	1838.683 661 58(90)		4.9×10^{-10}
neutron-muon mass ratio	m_n/m_μ	8.892 484 08(20)		2.2×10^{-8}
neutron-tau mass ratio	m_n/m_τ	0.528 790(48)		9.0×10^{-5}
neutron-proton mass ratio	m_n/m_p	1.001 378 418 98(51)		5.1×10^{-10}
neutron-proton mass difference	$m_n - m_p$	$2.305\,573\,77(85) \times 10^{-30}$	kg	3.7×10^{-7}
		0.001 388 449 00(51)	u	3.7×10^{-7}
energy equivalent	$(m_n - m_p)c^2$	$2.072\,146\,37(76) \times 10^{-13}$	J	3.7×10^{-7}
		1.293 332 05(48)	MeV	3.7×10^{-7}
neutron molar mass $N_A m_n$	$M(\text{n}), M_n$	$1.008\,664\,915\,88(49) \times 10^{-3}$	kg mol^{-1}	4.9×10^{-10}
neutron Compton wavelength $h/m_n c$	$\lambda_{C,n}$	$1.319\,590\,904\,81(88) \times 10^{-15}$	m	6.7×10^{-10}
	$\lambda_{C,n}/2\pi$	$0.210\,019\,415\,36(14) \times 10^{-15}$	m	6.7×10^{-10}
neutron magnetic moment	μ_n	$-0.966\,236\,50(23) \times 10^{-26}$	J T^{-1}	2.4×10^{-7}
to Bohr magneton ratio	μ_n/μ_B	$-1.041\,875\,63(25) \times 10^{-3}$		2.4×10^{-7}
to nuclear magneton ratio	μ_n/μ_N	$-1.913\,042\,73(45)$		2.4×10^{-7}
neutron g -factor $2\mu_n/\mu_N$	g_n	$-3.826\,085\,45(90)$		2.4×10^{-7}
neutron-electron magnetic moment ratio	μ_n/μ_e	$1.040\,668\,82(25) \times 10^{-3}$		2.4×10^{-7}
neutron-proton magnetic moment ratio	μ_n/μ_p	$-0.684\,979\,34(16)$		2.4×10^{-7}
neutron to shielded proton magnetic moment ratio (H_2O , sphere, 25°C)	μ_n/μ'_p	$-0.684\,996\,94(16)$		2.4×10^{-7}
neutron gyromagnetic ratio $2 \mu_n /\hbar$	γ_n	$1.832\,471\,72(43) \times 10^8$	$\text{s}^{-1} \text{T}^{-1}$	2.4×10^{-7}
	$\gamma_n/2\pi$	29.164 6933(69)	MHz T^{-1}	2.4×10^{-7}
Deuteron, d				
deuteron mass	m_d	$3.343\,583\,719(41) \times 10^{-27}$	kg	1.2×10^{-8}
		2.013 553 212 745(40)	u	2.0×10^{-11}
energy equivalent	$m_d c^2$	$3.005\,063\,183(37) \times 10^{-10}$	J	1.2×10^{-8}
		1875.612 928(12)	MeV	6.2×10^{-9}
deuteron-electron mass ratio	m_d/m_e	3670.482 967 85(13)		3.5×10^{-11}
deuteron-proton mass ratio	m_d/m_p	1.999 007 500 87(19)		9.3×10^{-11}
deuteron molar mass $N_A m_d$	$M(\text{d}), M_d$	$2.013\,553\,212\,745(40) \times 10^{-3}$	kg mol^{-1}	2.0×10^{-11}
deuteron rms charge radius	r_d	$2.1413(25) \times 10^{-15}$	m	1.2×10^{-3}
deuteron magnetic moment	μ_d	$0.433\,073\,5040(36) \times 10^{-26}$	J T^{-1}	8.3×10^{-9}
to Bohr magneton ratio	μ_d/μ_B	$0.466\,975\,4554(26) \times 10^{-3}$		5.5×10^{-9}
to nuclear magneton ratio	μ_d/μ_N	0.857 438 2311(48)		5.5×10^{-9}
deuteron g -factor μ_d/μ_N	g_d	0.857 438 2311(48)		5.5×10^{-9}
deuteron-electron magnetic moment ratio	μ_d/μ_e	$-4.664\,345\,535(26) \times 10^{-4}$		5.5×10^{-9}
deuteron-proton magnetic moment ratio	μ_d/μ_p	0.307 012 2077(15)		5.0×10^{-9}
deuteron-neutron magnetic moment ratio	μ_d/μ_n	$-0.448\,206\,52(11)$		2.4×10^{-7}

Triton, t

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
triton mass	m_t	$5.007\,356\,665(62) \times 10^{-27}$	kg	1.2×10^{-8}
		$3.015\,500\,716\,32(11)$	u	3.6×10^{-11}
energy equivalent	$m_t c^2$	$4.500\,387\,735(55) \times 10^{-10}$	J	1.2×10^{-8}
		$2808.921\,112(17)$	MeV	6.2×10^{-9}
triton-electron mass ratio	m_t/m_e	$5496.921\,535\,88(26)$		4.6×10^{-11}
triton-proton mass ratio	m_t/m_p	$2.993\,717\,033\,48(22)$		7.5×10^{-11}
triton molar mass $N_A m_t$	$M(t), M_t$	$3.015\,500\,716\,32(11) \times 10^{-3}$	kg mol ⁻¹	3.6×10^{-11}
triton magnetic moment	μ_t	$1.504\,609\,503(12) \times 10^{-26}$	J T ⁻¹	7.8×10^{-9}
to Bohr magneton ratio	μ_t/μ_B	$1.622\,393\,6616(76) \times 10^{-3}$		4.7×10^{-9}
to nuclear magneton ratio	μ_t/μ_N	$2.978\,962\,460(14)$		4.7×10^{-9}
triton g -factor $2\mu_t/\mu_N$	g_t	$5.957\,924\,920(28)$		4.7×10^{-9}
Helion, h				
helion mass	m_h	$5.006\,412\,700(62) \times 10^{-27}$	kg	1.2×10^{-8}
		$3.014\,932\,246\,73(12)$	u	3.9×10^{-11}
energy equivalent	$m_h c^2$	$4.499\,539\,341(55) \times 10^{-10}$	J	1.2×10^{-8}
		$2808.391\,586(17)$	MeV	6.2×10^{-9}
helion-electron mass ratio	m_h/m_e	$5495.885\,279\,22(27)$		4.9×10^{-11}
helion-proton mass ratio	m_h/m_p	$2.993\,152\,670\,46(29)$		9.6×10^{-11}
helion molar mass $N_A m_h$	$M(h), M_h$	$3.014\,932\,246\,73(12) \times 10^{-3}$	kg mol ⁻¹	3.9×10^{-11}
helion magnetic moment	μ_h	$-1.074\,617\,522(14) \times 10^{-26}$	J T ⁻¹	1.3×10^{-8}
to Bohr magneton ratio	μ_h/μ_B	$-1.158\,740\,958(14) \times 10^{-3}$		1.2×10^{-8}
to nuclear magneton ratio	μ_h/μ_N	$-2.127\,625\,308(25)$		1.2×10^{-8}
helion g -factor $2\mu_h/\mu_N$	g_h	$-4.255\,250\,616(50)$		1.2×10^{-8}
shielded helion magnetic moment (gas, sphere, 25 °C)	μ'_h	$-1.074\,553\,080(14) \times 10^{-26}$	J T ⁻¹	1.3×10^{-8}
to Bohr magneton ratio	μ'_h/μ_B	$-1.158\,671\,471(14) \times 10^{-3}$		1.2×10^{-8}
to nuclear magneton ratio	μ'_h/μ_N	$-2.127\,497\,720(25)$		1.2×10^{-8}
shielded helion to proton magnetic moment ratio (gas, sphere, 25 °C)	μ'_h/μ_p	$-0.761\,766\,5603(92)$		1.2×10^{-8}
shielded helion to shielded proton magnetic moment ratio (gas/H ₂ O, spheres, 25 °C)	μ'_h/μ'_p	$-0.761\,786\,1313(33)$		4.3×10^{-9}
shielded helion gyromagnetic ratio $2 \mu'_h /\hbar$ (gas, sphere, 25 °C)	γ'_h	$2.037\,894\,585(27) \times 10^8$	s ⁻¹ T ⁻¹	1.3×10^{-8}
	$\gamma'_h/2\pi$	$32.434\,099\,66(43)$	MHz T ⁻¹	1.3×10^{-8}
Alpha particle, α				
alpha particle mass	m_α	$6.644\,657\,230(82) \times 10^{-27}$	kg	1.2×10^{-8}
		$4.001\,506\,179\,127(63)$	u	1.6×10^{-11}
energy equivalent	$m_\alpha c^2$	$5.971\,920\,097(73) \times 10^{-10}$	J	1.2×10^{-8}
		$3727.379\,378(23)$	MeV	6.2×10^{-9}
alpha particle to electron mass ratio	m_α/m_e	$7294.299\,541\,36(24)$		3.3×10^{-11}
alpha particle to proton mass ratio	m_α/m_p	$3.972\,599\,689\,07(36)$		9.2×10^{-11}
alpha particle molar mass $N_A m_\alpha$	$M(\alpha), M_\alpha$	$4.001\,506\,179\,127(63) \times 10^{-3}$	kg mol ⁻¹	1.6×10^{-11}
PHYSICOCHEMICAL				
Avogadro constant	N_A, L	$6.022\,140\,857(74) \times 10^{23}$	mol ⁻¹	1.2×10^{-8}
atomic mass constant				
$m_u = \frac{1}{12}m(^{12}\text{C}) = 1\text{ u}$	m_u	$1.660\,539\,040(20) \times 10^{-27}$	kg	1.2×10^{-8}
energy equivalent	$m_u c^2$	$1.492\,418\,062(18) \times 10^{-10}$	J	1.2×10^{-8}
		$931.494\,0954(57)$	MeV	6.2×10^{-9}

Fundamental Physical Constants — Extensive Listing

Quantity	Symbol	Value	Unit	Relative std. uncert. u_r
Faraday constant ⁶ $N_A e$	F	96 485.332 89(59)	C mol ⁻¹	6.2×10^{-9}
molar Planck constant	$N_A h$	$3.990\,312\,7110(18) \times 10^{-10}$	J s mol ⁻¹	4.5×10^{-10}
	$N_A h c$	0.119 626 565 582(54)	J m mol ⁻¹	4.5×10^{-10}
molar gas constant	R	8.314 4598(48)	J mol ⁻¹ K ⁻¹	5.7×10^{-7}
Boltzmann constant R/N_A	k	$1.380\,648\,52(79) \times 10^{-23}$	J K ⁻¹	5.7×10^{-7}
		$8.617\,3303(50) \times 10^{-5}$	eV K ⁻¹	5.7×10^{-7}
	k/h	$2.083\,6612(12) \times 10^{10}$	Hz K ⁻¹	5.7×10^{-7}
	k/hc	69.503 457(40)	m ⁻¹ K ⁻¹	5.7×10^{-7}
molar volume of ideal gas RT/p $T = 273.15$ K, $p = 100$ kPa	V_m	$22.710\,947(13) \times 10^{-3}$	m ³ mol ⁻¹	5.7×10^{-7}
Loschmidt constant N_A/V_m	n_0	$2.651\,6467(15) \times 10^{25}$	m ⁻³	5.7×10^{-7}
molar volume of ideal gas RT/p $T = 273.15$ K, $p = 101.325$ kPa	V_m	$22.413\,962(13) \times 10^{-3}$	m ³ mol ⁻¹	5.7×10^{-7}
Loschmidt constant N_A/V_m	n_0	$2.686\,7811(15) \times 10^{25}$	m ⁻³	5.7×10^{-7}
Sackur-Tetrode (absolute entropy) constant ⁷ $\frac{5}{2} + \ln[(2\pi m_u k T_1 / h^2)^{3/2} k T_1 / p_0]$ $T_1 = 1$ K, $p_0 = 100$ kPa	S_0/R	-1.151 7084(14)		1.2×10^{-6}
$T_1 = 1$ K, $p_0 = 101.325$ kPa		-1.164 8714(14)		1.2×10^{-6}
Stefan-Boltzmann constant $(\pi^2/60)k^4/h^3c^2$	σ	$5.670\,367(13) \times 10^{-8}$	W m ⁻² K ⁻⁴	2.3×10^{-6}
first radiation constant $2\pi h c^2$	c_1	$3.741\,771\,790(46) \times 10^{-16}$	W m ²	1.2×10^{-8}
first radiation constant for spectral radiance $2hc^2$	c_{1L}	$1.191\,042\,953(15) \times 10^{-16}$	W m ² sr ⁻¹	1.2×10^{-8}
second radiation constant hc/k	c_2	$1.438\,777\,36(83) \times 10^{-2}$	m K	5.7×10^{-7}
Wien displacement law constants				
$b = \lambda_{\max} T = c_2/4.965\,114\,231\dots$	b	$2.897\,7729(17) \times 10^{-3}$	m K	5.7×10^{-7}
$b' = \nu_{\max}/T = 2.821\,439\,372\dots c/c_2$	b'	$5.878\,9238(34) \times 10^{10}$	Hz K ⁻¹	5.7×10^{-7}

¹ See the “Adopted values” table for the conventional value adopted internationally for realizing representations of the volt using the Josephson effect.

² See the “Adopted values” table for the conventional value adopted internationally for realizing representations of the ohm using the quantum Hall effect.

³ Value recommended by the Particle Data Group (Olive *et al.*, 2014).

⁴ Based on the ratio of the masses of the W and Z bosons m_W/m_Z recommended by the Particle Data Group (Olive *et al.*, 2014). The value for $\sin^2\theta_W$ they recommend, which is based on a particular variant of the modified minimal subtraction ($\overline{\text{MS}}$) scheme, is $\sin^2\hat{\theta}_W(M_Z) = 0.231\,26(5)$.

⁵ This and all other values involving m_τ are based on the value of $m_\tau c^2$ in MeV recommended by the Particle Data Group (Olive *et al.*, 2014).

⁶ The numerical value of F to be used in coulometric chemical measurements is 96 485.3251(12) [1.2×10^{-8}] when the relevant current is measured in terms of representations of the volt and ohm based on the Josephson and quantum Hall effects and the internationally adopted conventional values of the Josephson and von Klitzing constants K_{J-90} and R_{K-90} given in the “Adopted values” table.

⁷ The entropy of an ideal monoatomic gas of relative atomic mass A_r is given by $S = S_0 + \frac{3}{2}R \ln A_r - R \ln(p/p_0) + \frac{5}{2}R \ln(T/K)$.

Appendix Z

Periodic Table of the Elements

IUPAC Periodic Table of the Elements

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 H hydrogen [1.0078, 1.0082]	2 He helium 4.0026	3 Li lithium 6.941 [6.938, 6.977]	4 Be beryllium 9.0122	5 B boron 10.81 [10.806, 10.821]	6 C carbon 12.011 [12.009, 12.012]	7 N nitrogen 14.007 [14.006, 14.008]	8 O oxygen 15.999 [15.998, 16.000]	9 F fluorine 18.998	10 Ne neon 20.180	11 Na sodium 22.990 [22.984, 22.997]	12 Mg magnesium 24.305 [24.304, 24.307]	13 Al aluminum 26.982	14 Si silicon 28.086 [28.084, 28.088]	15 P phosphorus 30.974 [30.969, 30.979]	16 S sulfur 32.06 [32.059, 32.076]	17 Cl chlorine 35.45 [35.446, 35.457]	18 Ar argon 39.948
19 K potassium 39.098	20 Ca calcium 40.078(4)	21 Sc scandium 44.956	22 Ti titanium 47.887	23 V vanadium 50.942	24 Cr chromium 51.996	25 Mn manganese 54.938	26 Fe iron 55.845(2)	27 Co cobalt 58.933	28 Ni nickel 58.693	29 Cu copper 63.546(3)	30 Zn zinc 65.38(2)	31 Ga gallium 69.723	32 Ge germanium 72.630(8)	33 As arsenic 74.922	34 Se selenium 78.971(8)	35 Br bromine 79.904 [79.901, 79.907]	36 Kr krypton 83.798(2)
37 Rb rubidium 85.468	38 Sr strontium 87.62	39 Y yttrium 88.906	40 Zr zirconium 91.224(2)	41 Nb niobium 92.906	42 Mo molybdenum 95.95	43 Tc technetium	44 Ru ruthenium 101.07(2)	45 Rh rhodium 102.91	46 Pd palladium 106.42	47 Ag silver 107.87	48 Cd cadmium 112.41	49 In indium 114.82	50 Sn tin 118.71	51 Sb antimony 121.76	52 Te tellurium 127.60(3)	53 I iodine 126.90	54 Xe xenon 131.29
55 Cs caesium 132.91	56 Ba barium 137.33	57-71 lanthanoids	72 Hf hafnium 178.49(2)	73 Ta tantalum 180.95	74 W tungsten 183.84	75 Re rhenium 186.21	76 Os osmium 190.23(3)	77 Ir iridium 192.22	78 Pt platinum 195.08	79 Au gold 196.97	80 Hg mercury 200.59	81 Tl thallium 204.38 [204.36, 204.39]	82 Pb lead 207.2	83 Bi bismuth 208.98	84 Po polonium	85 At astatine	86 Rn radon
87 Fr francium	88 Ra radium	89-103 actinoids	104 Rf rutherfordium	105 Db dubnium	106 Sg seaborgium	107 Bh bohrium	108 Hs hassium	109 Mt meitnerium	110 Ds darmstadtium	111 Rg roentgenium	112 Cn copernicium	113 Nh nihonium	114 Fl flerovium	115 Mc moscovium	116 Lv livermorium	117 Ts tennessine	118 Og oganeson



INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

For notes and updates to this table, see www.iupac.org. This version is dated 28 November 2016.
Copyright © 2016 IUPAC, the International Union of Pure and Applied Chemistry.

References

- [1] Elroy M. Avery, *School Physics* (Sheldon and Co., New York, 1895).
- [2] Albert F. Blaisdell, *Our Bodies and How We Live*, Boston: Ginn, 190, 259.
- [3] *Encyclopædia Britannica* (11th ed., vol. 18), New York, NY: The Encyclopaedia Britannica Company, 1910.
- [4] Ellsworth D. Foster (ed.) *The American Educator* (vol. 2), Chicago, IL: Ralph Durham Company, 1921.
- [5] Elroy M. Avery, *School Physics* (New York: 1895).
- [6] R. Fay, *Hearing in Vertebrates. A Psychophysics Databook*. Hill-Fay Associates, Winnetka, Illinois, 1988.
- [7] A.E. Fitzgerald and David E. Higginbotham, *Basic Electrical Engineering*. McGraw-Hill, 1957.
- [8] R. Fitzpatrick, *Oscillations and Waves* (Web site at the University of Texas, Austin):
<http://farside.ph.utexas.edu/teaching/315/Waves/Waves.html>
- [9] E. Gelin, *Éléments de Trigonométrie plane et sphérique à l'usage des élèves des Cours professionnels des candidats aux Écoles spéciales des Universités et à l'École militaire de Bruxelles* (1888).
- [10] Douglas C. Giancoli, *Physics* (6th ed., Pearson Prentice-Hall, 2005).
- [11] Illustration taken from <http://commons.wikimedia.org>, and created using program Vector-FieldPlot. A collection of images is available at:
http://commons.wikimedia.org/wiki/Category:Created_with_VectorFieldPlot
- [12] “A New Geomagnetic Polarity Time Scale for the Late Cretaceous and Cenozoic”, Cande and Kent, *J. Geophys. Res.*, **97**, 13,917 (1992).
- [13] S.M. Sze, *Physics of Semiconductor Devices*. John Wiley & Sons, Inc. New York, New York, 1981.
- [14] John Leslie, *The Philosophy of Arithmetic; Exhibiting a Progressive View of the Theory and Practice of Calculation, with Tables for the Multiplication of Numbers as Far as One Thousand*. Abernethy & Walker, Edinburgh, 1820.
- [15] William Dwight Whitney, *The Century Dictionary, an Encyclopedic Lexicon of the English Language* (New York: The Century Co., 1902).
- [16] “Multiple rainbows from single drops of water and other liquids”, *Am. J. Phys.*, May 1976, 421433.
- [17] Jerry D. Wilson and Anthony J. Buffa, *College Physics*, Prentice-Hall, 2003.

Index

- Aberration, 200
 - chromatic, 200
 - spherical, 200
- Accidentals, 82
- Accommodation, 204
- Acoustics, 12
- Active optics, 193
- Aerial, 181
- Agonic line, 147
- Alexander's dark band, 238
- Alnico, 157
- Alternating current (AC), 173
- AM radio, 178
- Amateur radio, 132
- Amber, 90
- American Radio Relay League (ARRL), 133
- American Wire Gauge (AWG), 107, 109
- Ammeter, 120
- Ampère's law, 159, 161
- Ampere, 16
- Amplitude, 39, 55, 58
- Amplitude modulation, 178
- Analyzer, 226
- Angular frequency, 39, 58
- Antenna, 181
- Antimatter, 177, 252
- Antineutron, 252, 308
- Antinode, 66
- Antiproton, 252, 308
- Arduino, 131
- Arithmetic-geometric mean, 296
- Astigmatism, 200
- Astronomical unit, 280
- Astrophysics, 12
- Atomic mass unit (amu), 17
- Atomic physics, 12
- Aurora, 149
- Auroral oval, 153
- Baryon, 251, 308
- Base units, 14
- Battery, 101
 - internal resistance, 102, 108
 - parallel, 102
 - series, 101
- Bel, 79
- Bessel function, 69
- Beta decay, 252
- Bifocal lenses, 204
- Big Bang, 78
- Binary prefixes, 22
- Binoculars, 207
- Biophysics, 12
- Biot-Savart law, 139, 159, 164
- Birefringence, 227
- Biv, Roy G., 177, 228
- Blueshift, 78
- Boson, 308
- Bow shock, 151
- Brewster's law, 227
- Bulk modulus, 59, 71
- C major scale, 81, 83
- Calcite, 204, 227
- Calculus, 25
 - differential, 26
 - fundamental theorem of, 30
 - integral, 28
- Camera, 206
 - pinhole, 206
- Candela, 16, 212
- Capacitance, 122
- Capacitor, 122
- Capillary, 248
- Cat's whisker, 183
- Cauchy dispersion formula, 224
- Cellular telephone, 179
- Center of curvature, 189
- Chemical physics, 12
- Chromatic aberration, 200, 224
- Chromatic scale, 81
- Chromaticity diagram, 231

- CIE
 - chromaticity diagram, 231, 299
 - color matching functions, 299
- Circuit, 111
- Classical mechanics, 12
- Clef sign, 84
- Coercivity, 157
- Coherent light, 216
- Color
 - blue, 228
 - complement, 229
 - constancy, 228
 - cyan, 229
 - green, 228
 - magenta, 229
 - orange, 230
 - primary, 228
 - purple, 231
 - red, 228
 - secondary, 229
 - spectral, 230
 - violet, 231
 - white, 229
 - yellow, 228
- Coma, 200
- Compression, 57
- Computer glasses, 204
- Condenser, 122
- Conductor, 92
- Cones, 202, 228
- Converging
 - lens, 196
 - mirror, 189
- Cornea, 202
- Coulomb, 91
- Coulomb's law, 91
 - magnetic, 135
- Cross product, 288
- Cross-disciplinary physics, 12
- Crystal oscillator, 184
- Crystal radio, 181
- CubeSat, 133
- Curie temperature, 157
- Currency exchange rates, 22
- Current divider, 116
- Cyclotron frequency, 143
- Cyclotron radius, 143
 - overdamped, 47
 - underdamped, 47
- Daraf, 122
- Decibel, 79
- Declination, magnetic, 147
- Degree, 17, 285
 - square, 286
- Diamagnetism, 156
- Dielectric, 92
- Dielectric breakdown, 97
- Dielectric constant, 124
- Diffraction, 218
- Dimensional analysis, 19
- Diode, 181
- Dipole
 - electric, 95
 - magnetic, 138, 147
 - moment, electric, 95
 - moment, magnetic, 138
- Direct current (DC), 173
- Dispersion, 200, 224
- Displacement current, 176
- Diverging
 - lens, 196
 - mirror, 189
- Doppler effect, 75
 - relativistic, 76
- Dot product, 288
- Drift velocity, 103
- Earth, 280
- Earthquake, 65
- Eddy current, 158
- Elastance, 122
- Electric current, 103
- Electric field, 94
- Electric field lines, 94
- Electric flux, 95
- Electric generator, 161
- Electric motor, 161
- Electricity, 90
- Electricity and magnetism, 12
- Electrode, 101
- Electrolyte, 101
- Electromagnet, 136
- Electromagnetic force, 252
- Electromagnetic units, 19
- Electromagnetic wave, 176
- Electromagnetism, 135
- Electromotive force, 108, 161

- Electron, 308
- Electron volt, 100
- Electrostatic units, 19
- Electroweak theory, 252
- Elementary charge (e), 90
- Emmetropia, 204
- Epicenter, 65
- Equilibrium position, 39
- Equipotential surface, 99
- Eye
 - compound, 204
 - human, 202
 - trilobite, 204
- Farad, 122
- Faraday's law, 161
- Farsightedness, 204
- Fermat's principle, 194
- Fermion, 308
- Ferromagnetism, 156
- Field-programmable gate array (FPGA), 132
- Flat, 82
- FM radio, 178
- Focal length
 - lens, 196
 - mirror, 189
- Focus, 189, 196
- Foot, 15
- Foot-candle, 213
- Forced oscillations, 49
- Fountain effect, 248
- Fourth, 285
- FPGA, *see* Field programmable gate array
- Franklin, Benjamin, 90
- Frequency, 43
- Frequency modulation, 178
- Fresnel lens, 199
- Gadget Factory, 132
- Galena, 183
- Gamma rays, 177
- Gauss, 137
- Gauss's law, 96, 122, 161
 - for magnetism, 139, 161
- Gaussian units, 19
- Geophysics, 12
- Grad, 17, 285
- Grand Unified Theory, 252
- Grave (*f.n.*), 18
- Gravitational force, 252
- Graviton, 252
- Ground, 99, 111
- Gyrofrequency, 143
- Gyroradius, 143
- HackerBoxes, 132
- Hadron, 251, 308
- Half step, 81
- Hall effect, 144
- Hall emf, 145
- Heaviside-Lorentz units, 19
- Helium, 72
- Helium II, 248
- Henry, 164
- Higgs
 - boson, 253
 - field, 253
- Higgs boson, 308
- Hooke's law, 39, 54, 110
- Hubble constant, 78
- Hubble's law, 78
- Hyperopia, 204
- Hysteresis, 156, 157
- Iceland spar, 227
- Illuminance, 212
- IMAGE, 303
- Image distance, 189, 196
- Image height, 189, 196
- Impact parameter, 237
- Impedance, 173
- Imperial units, 14
- Incidence, angle of, 188
- Index of refraction, 194, 224
- Inductance, 164
- Infinitesimal numbers, 25
- Infrared light, 177
- Infrasound, 74
- Insulator, 92
- Integral, 29
 - double, 34
- Integrand, 29
- Interference, 61
 - constructive, 62
 - destructive, 62
- International Prototype Kilogram (IPK), 15
- Inverted image, 191, 197
- Ionosphere, 180
- Isogonic chart, 147
- Isotropic, 64, 212

- Jacobi elliptic function, 295
Jupiter, 280
- K20, 16
Kaleidoscope, 210
Kelvin, 16
Kirchhoff plot, 111
Kirchhoff's rules, 117
- L wave, 65
Lagrangian mechanics, 56
Lambda point, 248
Land effect, 204, 228
Larmor radius, 143
Law of reflection, 188
LC circuit, 169
LCR circuit, 172
Left-hand rule, 142
Lens, 196
 - double concave, 196
 - double convex, 196
 - meniscus, 196
 - of human eye, 202
 - plano-concave, 196
 - plano-convex, 196

Lens maker's equation, 196
Lenz's law, 162
Lepton, 251, 308
Light
 - white, 228

Lightning, 71, 97, 135
Line of purples, 231
Lodestone, 135
Logic probe, 121
Lorentz force, 142
Love wave, 65
LR circuit, 167
Lumen, 211
Luminance, 299
Luminous efficiency curve, 211
Luminous flux, 211
Luminous intensity, 212
Lux, 213

Magnet
 - alnico, 157
 - ferrite, 157
 - neodymium, 157
 - permanent, 157
 - rare-earth, 157
 - samarium-cobalt, 157

Magnetic declination, 147
Magnetic domain, 156
Magnetic field, 137
Magnetic field lines, 137
Magnetic flux, 138
Magnetic inclination, 147
Magnetic monopole, 136, 139
Magnetic reconnection, 149
Magnetic susceptibility, 166
Magnetism, 135
Magnetite, 135
Magnetopause, 151
Magnetosheath, 151
Magnetosphere, 149, 303
Magnetotail, 149
Magnification, 189, 196
Magnification equation, 192, 198
Magnifier, 202
Magnifying glass, 202
Magnitude, 214
Major scales, 83
Maker Shed, 132
Malus's law, 226
Mars, 280
Mathematical physics, 12
Maxwell's equations, 96, 135, 161, 176
Memristance, 175
Memristor, 175
Mercury, 280
Meson, 251, 308
Metallic hydrogen, 92
Meter, 15
Metric ton, 16
Metric units, 14
Mho, 107
Microcontrollers, 131
Micron, 22
Microscope, 206
Microwaves, 177
Middle C, 82
Minor scales, 84
Mirror, 189
 - concave, 189
 - convex, 189

Mirror equation, 192, 198
Modulus of rigidity, 45
Mole, 16
Moment magnitude scale, 65

- Moment of inertia, 292
- Monochromatic light, 216
- Monocular, 207
- Motional emf, 162
- Multimeter, 120
- Music, 81
 - instruments, 87
 - measure, 86
 - notes, 81, 85
 - rests, 85
 - scale, 81, 82
 - tempo, 86
 - time signature, 86
- Myopia, 204
- Natural units, 14
- Nearsightedness, 204
- Neper, 80
- Neptune, 280
- NerdKits, 132
- Neutrino, 308
- Neutron, 251, 308
- Newton, 16
- Newton's laws of motion, 12
- Newton-Laplace equation, 71
- Node, 66
- Nought, 47
- Nuclear physics, 12
- Object distance, 189, 196
- Object height, 189, 196
- Obliquity of the ecliptic, 280
- Octave, 81
- Ohm, 104
- Ohm's law, 110
- Ohmmeter, 120
- Optic nerve, 202
- Optics, 12
- Oscilloscope, 120
- Oval of Descartes, 204
- P wave, 65
- Papilio FPGA board, 132
- Parallel
 - batteries, 102
 - capacitors, 123
 - inductors, 165
 - resistors, 106
 - springs, 44
- Parallel axis theorem, 292
- Paramagnetism, 156
- Parsec, 78
- Particle physics, 12
- Pascal, 141
- Pendulum
 - ballistic, 56
 - conical, 54
 - double, 56
 - Foucault, 56
 - nonlinear, 294
 - physical, 54
 - simple plane, 51, 294
 - spherical, 52
 - torsional, 54
- Pentatonic scale, 84
- Period, 43
- Permeability, 166
 - of free space (μ_0), 135
 - relative, 166
- Permittivity, 124
 - of free space (ϵ_0), 91
- Phase constant, 39, 58
- Photometry, 211
- Photon, 216
- Physics, 12
- Piezoelectric effect, 184
- Pigments, 230
- Pitch, 81
- Plasma, 142
- Plasma physics, 12
- Pluto, 280
- Poisson ratio, 45
- Polar wandering, 147
- Polarization angle, 227
- Polarized light, 226
- Polarizer, 226
- Pole strength, 135
- Positron, 252, 308
- Potential, 99
- Pound, 19
 - force, 19
 - mass, 16, 19
- Presbyopia, 204
- Pressure
 - magnetic, 141
- Primary colors, 228
- Proton, 251, 308
- Purples, line of, 231
- Quantum electrodynamics, 135

- Quantum mechanics, 12, 216, 248, 250
Quark, 251, 308
- Radian, 17
radian, 285
Radio waves, 177
Radiometry, 211
Radius of curvature
 mirror, 189
Ragchewing, 179
Rainbow, 224, 234
 angle, 237
 primary, 234
 secondary, 234
Rarefaction, 57
Raspberry Pi, 131
Rayleigh criterion, 218
Rayleigh wave, 65
RC circuit, 126
Real image, 191, 196
Rectangular rule, 32
Redshift, 78
Reflected wave, 59
Reflection
 angle of, 188
 coefficient of, 59
Refraction, 194
 law of, 194
Relativity, 12
 general, 12
 special, 12
Remanence, 157
Resistance, 104
Resistivity, 104, 105
 temperature coefficient of, 104, 105
Resistor, 104
Resonance, 49
Retina, 202
Return stroke, 97
Richter scale, 65
Right-hand rule, 139, 140, 142, 288
Robotics, 133
Rods, 202, 228
Rollin film, 248
- S wave, 65
Saturn, 280
Scattering, 227
Scattering angle, 237
Schematic diagram, 111
- Second (of time), 16
Second sound, 250
Secondary colors, 229
Seismic waves, 65
Selective absorption, 226
Selectivity, 184
Sellmeier dispersion formula, 224
Semiconductor, 92
Series
 batteries, 101
 capacitors, 123
 inductors, 165
 resistors, 106
 springs, 44
Sharp, 82
Shortwave radio, 178
SI units, 15
Siemens, 107
Simple harmonic motion, 39, 141, 169
 kinetic energy, 41
 potential energy, 41
 total energy, 41
Single-slit diffraction, 218
Slug, 19
Snell's law, 194
Solar wind, 149
Solenoid, 140, 164
Solid angle, 285
Solid-state physics, 12
Sound, 71
 audible, 74
 infrasonic, 74
 loudness, 79
 speed, 71
 ultrasonic, 74
Sound level, 79
Space physics, 149
SparkFun, 132
Spectral power distribution, 299
Spectrum, 230
Spherical aberration, 189, 196, 200
Spring
 vertical, 43
Spring constant, 39, 43, 44
Spyglass, 207
Staff, 84
Standard Model, 251
Standing waves, 66
Statistical mechanics, 12

- Steinhart-Hart equation, 106
- Stepped leader, 97
- Steradian, 285
- String theory, 253
- Strong nuclear force, 252
- Submillimeter waves, 177
- Sulfur hexafluoride, 72
- Superconductor, 92
- Superflow, 248
- Superfluid, 92, 248
- Superposition, 61
- Susceptibility, *see* Magnetic susceptibility

- Telescope, 210
- Telescope, 207, 208
 - Cassegrain, 208
 - Newtonian, 208
 - reflecting, 207
 - refracting, 207
- Television, 179
- Terminal voltage, 108
- Tesla, 137
- Thermistor, 106
- Thermodynamics, 12
- Third, 285
- Threshold of hearing, 79
- Threshold of pain, 80
- Time constant, 126, 167
- Time travel, 245
- Total internal reflection, 195
- Transmission, coefficient of, 59
- Transmitted wave, 59
- Transmitter, 184, 186
- Transverse wave, 57
- Trifocal lenses, 204
- Trilobite, 204
- Tristimulus values, 299
- Two-fluid model, 248

- Ultrasound, 74
- Ultraviolet light, 177
- Unit vector, 92, 287
- Upright image, 191, 197
- Uranus, 280

- Vector, 287
 - polar form, 288
 - rectangular form, 288
- Vector boson, 252, 308
- Venus, 280

- Verilog, 132
- VHDL, 132
- Virtual image, 191, 196
- Visible light, 177
- Voltage divider, 116
- Voltmeter, 120

- W boson, 252
- Wave, 57
 - cylindrical, 64
 - energy, 62
 - intensity, 64
 - longitudinal, 57
 - ocean, 64
 - plane, 64
 - power, 64
 - speed, 58, 59
 - spherical, 64
 - standing, 66
 - string, 59
 - transverse, 57
 - tsunami, 65
- Wave equation, 176
- Wave number, 58
- Wavelength, 58
- Weak nuclear force, 252
- Weber, 139
- Weight, 16
- White light, 224
- Whole step, 81
- Whole tone scale, 84
- Wire, 107

- X-rays, 177

- Young's experiment, 216
- Young's modulus, 45, 59, 71

- Z boson, 252