Electromagnetics and Applications

David H. Staelin

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA

Table of Contents

Preface

The initial development of electrical science and engineering a century ago occurred almost entirely within the domain of electromagnetics. Most electrical curricula remained polarized around that theme until the mid-twentieth century when signal, device, and computational subjects became dominant. Continued expansion of the field has currently relegated undergraduate electromagnetics to perhaps a one-semester subject even though electromagnetic technology has expanded substantially and is basic to most applications. To meet the increasing educational challenge of providing both breadth and depth in electromagnetics within a brief presentation, this text uses a more physics-based approach and novel methods of explaining certain phenomena. It introduces students to electrodynamics across the entire range from statics to dynamics, and from motors to circuits, communications, optical fibers, and lasers. For example, we currently cover approximately ninety percent of the text in a one-semester subject meeting with faculty four hours per week. The text could also support undergraduate offerings over two quarters or even two semesters, and could perhaps also be used at the entry graduate level.

The main objectives of the text are to: 1) convey those big ideas essential to understanding the electromagnetic aspects of modern electrical and computer systems, 2) expose students to enough examples to make the big ideas tangible and erase most naiveté about dominant applications, 3) provide computational experience with Maxwell's equations sufficient to treat the basic examples, 4) provide the understanding and skills prerequisite to follow-on subjects, and 5) reinforce prior exposure to physics, mathematics, and electrical systems so as to help integrate student learning, including problem solving and design methods.

The first two chapters are the core of the text. They review the basic physics of electromagnetics and electromechanics and introduce the Lorentz force law, Maxwell's equations, media, boundary conditions, static field solutions, uniform plane waves, and power and energy. Although the chapters are best read sequentially, the four topical areas that follow the core can be read in any sequence and include: 1) Chapters 3, 5, and 6, which treat RLC devices and circuits; electromagnetic forces on charges, conductors, and media; and motors, 2) Chapters 4, 7, and 8, which treat quasistatics, solutions to Laplace's equation, and TEM lines, including matching, resonators, and transients, 3) Sections 4.1–4.3 plus Chapter 9, which treat field relaxation, non-uniform plane waves, reflection, waveguides, and cavity resonators, and 4) Chapters 10 and 11, which treat radiation, wire and aperture antennas, and applications such as communications systems and radar. Two "capstone" chapters then follow: Chapter 12 introduces optical waveguides, laser amplifiers, laser oscillators, and other optical devices (Chapters 9 and 11 are prerequisites), and Chapter 13 reviews most wave phenomena in an acoustic context after Chapters 7, 9, and 10 have been covered. This organization permits use of the text in a wide variety of formats, including one- and two-semester options. Most prerequisites are reviewed briefly in the Appendix or within the text. Future versions will have home problems and more examples.

Special thanks are owed to the many MIT faculty who have taught this subject and its three merged predecessors while sharing their insights with the author over the past forty years. Without such collegial participation the scope and brevity of this text would not have been

Chapter 1: Introduction to Electromagnetics and Electromagnetic Fields

1.1   Review of foundations

1.1.1   <u>Introduction</u>

Electromagnetics involves the macroscopic behavior of electric charges in vacuum and matter. This behavior can be accurately characterized by the Lorentz force law and Maxwell's equations, which were derived from experiments showing how forces on charges depend on the relative locations and motions of other charges nearby. Additional relevant laws of physics include Newton's law, photon quantization, and the conservation relations for charge, energy, power, and momentum. Electromagnetic phenomena underlie most of the "electrical" in "electrical engineering" and are basic to a sound understanding of that discipline.

Electrical engineering has delivered four "miracles" — sets of phenomena that could each be considered true magic prior to their development. The first of these to impress humanity was the electrical phenomenon of lightning, often believed to be a tool of heaven, and the less powerful magnetic force that caused lodestones to point north. The explanation and application of these invisible forces during the eighteenth and nineteenth centuries vaulted electrical engineering to the forefront of commercial interest as motors, generators, electric lights, batteries, heaters, telephones, record players, and many other devices emerged.

The second set of miracles delivered the ability to communicate instantly without wires around the world, not only dots and dashes, but also voice, images, and data. Such capabilities had been commonplace in fairy tales, but were beyond human reach until Hertz demonstrated radiowave transmission in 1888, 15 years after Maxwell's predictions. Marconi extended the technique to intercontinental distances.

Third came electronics and photonics — the ability to electrically manipulate individual electrons and atoms in vacuum and in matter so as to generate, amplify, manipulate, and detect electromagnetic signals. During the twentieth century vacuum tubes, diodes, transistors, integrated circuits, lasers, and superconductors all vastly extended the capabilities and applications of electromagnetics.

The fourth set of electrical phenomena involves cybernetics and informatics — the manipulation of electrical signals so complex that entirely new classes of functionality are obtained, such as optimum signal processing, computers, robotics, and artificial intelligence. This text focuses on the electromagnetic nature of the first three sets of phenomena and explores many of their most important applications.

Chapter 1 of this text begins with a brief review of the underlying laws of physics, followed by the Lorentz force law and the nature of electric and magnetic fields. Chapter 2 introduces electrodynamics and Maxwell's equations, leading to uniform plane waves in space and media, and definitions of power, energy, boundary conditions, and uniqueness. The next four chapters

address static and quasistatic systems beginning with Chapter 3, which explores electromagnetics in the context of RLC circuits and devices. Chapter 4 addresses the more general behavior of quasistatic electric and magnetic fields in homogeneous and inhomogeneous media. Chapter 5 introduces electromagnetic forces while Chapter 6 addresses their application to motors, generators, actuators, and sensors.

The second half of the text focuses on electrodynamics and waves, beginning with TEM transmission lines in Chapters 7 and 8, and waves in media and at boundaries in Chapter 9. Antennas and radiation are treated in Chapters 10 and 11, while optical and acoustic systems are addressed in Chapters 12 and 13, respectively. Acoustics is introduced on its own merits and as a useful way to review electromagnetic wave phenomena such as radiation and resonance in a more physical and familiar context. The appendices list natural constants and review some of the prerequisite mathematics.

The rationalized international system of units (rationalized SI units) is used, which largely avoids factors of 4π. SI units emphasize meters (m), kilograms (kg), seconds (s), Amperes (A), and Kelvins (K); most other units can be expressed in terms of these. The SI system also favors units in multiples of 10^3; for example, it favors meters and millimeters over centimeters. The algebraic convention used here is that operations within parentheses are performed before others. Within parentheses and exponents and elsewhere, exponentiation is performed first, and multiplication before division; all these operations are performed before addition and subtraction.

1.1.2   Review of basic physical concepts and definitions

The few basic concepts summarized below are central to electromagnetics. These concepts include conservation of energy, power, and charge, and the notion of a photon, which conveys one quantum of electromagnetic energy. In addition, Newton's laws characterize the kinematics of charged particles and objects influenced by electromagnetic fields. The conservation laws also follow from Maxwell's equations, which are presented in Section 2.1 and, together with the Lorentz force law, compress all macroscopic electromagnetic behavior into a few concise statements.

This text neglects relativistic issues introduced when mass approaches the velocity of light or is converted to or from energy, and therefore we have *conservation of mass*: the total mass m within a closed envelope remains constant.

Conservation of energy requires that the total energy w_T [Joules] remains constant within any system closed so that no power enters or leaves, even though the form of the internally stored energy may be changing. This total energy w_T may include electric energy w_e, magnetic energy w_m, thermal energy w_{Th}, mechanical kinetic energy w_k, mechanical potential energy w_p, and energy in chemical, atomic, or other forms w_{other}; w_{other} is neglected here. Conservation of energy means:

$$w_T = w_e + w_m + w_k + w_p + w_{Th} + w_{other} \text{ [Joules]} = \text{constant} \qquad (1.1.1)$$

In this text we generally use lower case letters to indicate totals, and upper case letters to indicate densities. Thus we represent total energy by w_T [J] and total energy density by W_T [J m^{-3}]. Similarly, f [N] denotes the total force on an object and F [N m^{-3}] denotes the force density.

Unfortunately the number of electromagnetic variables is so large that many letters are used in multiple ways, and sometimes the meaning must be extracted from the context. For example, the symbol f is used to signify both force and frequency.

Newton's law says that a one-Newton force f would cause an otherwise force-free kilogram mass to accelerate at one meter per second per second; this defines the *Newton*. One Newton is roughly the terrestrial gravitational force on a quarter-pound weight (e.g. the weight of the apple that allegedly fell on Newton's head, prompting him to conceive the law of gravity). Newton's law may be expressed as:

$$f = ma \quad \text{[Newtons]} \tag{1.1.2}$$

where m is the mass of the object [kg] and a is the induced acceleration [ms^{-2}].

The unit of energy, the *Joule*, is the total energy w_T delivered to an object when a force f of one Newton is applied to it as it moves one meter in the direction z of the force. Therefore:

$$f = \frac{dw_T}{dz} \tag{1.1.3}$$

The *kinetic energy* w_k of a mass m moving at velocity v is:

$$w_k = \frac{1}{2}mv^2 \quad \text{[J]} \tag{1.1.4}$$

which, when added to its *potential energy* w_p, equals its total energy w_T relative to a motionless reference position; i.e.:

$$w_T = w_k + w_p \tag{1.1.5}$$

It is easy to see that if w_p remains constant, then (1.1.3) and (1.1.4) are consistent with f = ma; that is, $f = dw_T/dz = dw_k/dz = mv \, dv/dz = m(dz/dt)(dv/dz) = m \, dv/dt = ma$.

Conservation of power means, for example, that the total power P_{in} [Js^{-1}] entering a closed volume must equal the rate of increase [Js^{-1}] of the total energy stored there; that is:

$$P_{in} \text{ [W]} = \frac{dw_T}{dt} \text{ [Js}^{-1}] \tag{1.1.6}$$

where dw_T/dt is the time derivative of w_T, and the units [Joules per second] are often replaced by their equivalent, Watts [W]. If $dw_T/dt = 0$, then the power flowing into a closed volume must equal the power flowing out so that power is conserved. These laws also apply to

- 13 -

electromagnetic power and energy, and their definition in terms of electromagnetic fields appears in Section 2.7.

In mechanical systems one watt is delivered to an object if it received one joule in one second. More generally the *mechanical power* P delivered to an object is P = fv [W], where f is the only force [N] acting on the object, and v [ms^{-1}] is the object's velocity in the same direction as the force vector $\overline{f}$. More generally,

$$P = \overline{f} \bullet \overline{v} \equiv fv \cos\theta \ [W] \tag{1.1.7}$$

where $\overline{v}$ is the velocity vector and θ is the angle between $\overline{f}$ and $\overline{v}$.

Conservation of momentum requires that the total momentum of a set of interacting masses m_i remains constant if the set is free from external forces. The *momentum* of any object is mv [kg ms^{-1}], so in a force-free environment:

$$d\left(\sum_i m_i v_i\right)/dt = 0 \tag{1.1.8}$$

Conservation of charge requires that the total electric charge Q inside any volume must remain constant if no net charge crosses the boundaries of that volume. This is analogous to *conservation of mass*, although nuclear and other processes can convert mass m to energy E and vice-versa (E = mc^2). Charge conservation, however, has no significant exceptions. *Electric charge* is generally quantized in positive or negative multiples of the charge e on an *electron*, where:

$$e = -1.6021 \times 10^{-19} \text{ Coulombs} \tag{1.1.9}$$

The unit of charge, one *Coulomb*, is the charge conveyed by one Ampere flowing for one second, where the *Ampere* is the unit of electric current.

Photons carry the smallest unit of energy that can be conveyed by electromagnetic waves. The energy E of a single photon is:

$$E = hf \ [J] \tag{1.1.10}$$

where h is *Planck's constant* (6.624×10^{-34} [J s]) and f is the photon frequency [Hz]. Sometimes it is more convenient to think of electromagnetic waves as continuous waves, and sometimes it is more convenient to think of them as consisting of particles (photons), each of energy E. The total power P conveyed by an electromagnetic wave at frequency f is therefore the number N of photons passing per second times the photon energy E:

$$P = N \, hf \ [W] \tag{1.1.11}$$

The frequency of a wave is simply related to its wavelength λ and the velocity of light c:

$$f = c/\lambda \qquad\qquad (1.1.12)$$

Example 1.1A

A typical fully charged 1-kilowatt-hour car battery can accelerate a perfectly efficient 1000-kg electric automobile to what maximum speed?

Solution: The battery energy w_e[J] equals 1000 watts times 3600 seconds (one kilowatt-hour). It also equals the maximum kinetic energy, $w_k = mv^2/2$, of the speeding automobile (mass = m = 1000, velocity = v) after the battery is totally drained. Therefore $w_k = 3.6 \times 10^6 \Rightarrow v = (2w_k/m)^{0.5} = (7.2 \times 10^6/1000)^{0.5} \cong 85$ m s$^{-1} \cong$ 190 mph.

Example 1.1B

A sunny day delivers ~1kw m^{-2}; to how many photons N per second per square meter does this correspond if we (incorrectly) assume they all have the same wavelength $\lambda = 5 \times 10^{-7}$ meters? (0.5 microns is in the visible band.)?

Solution: Power = Nhf = Nhc/λ = 1 kw, so N = $10^3\lambda$/hc $\cong$ $10^3 \times 5 \times 10^{-7}/(6.6 \times 10^{-34} \times 3 \times 10^8) \cong$ 2.5×10^{20} photons m^{-2}s^{-1}.

1.2   Forces and the measurement and nature of electromagnetic fields

Electric fields $\overline{E}$ and magnetic fields $\overline{H}$ are manifest only by the forces they exert on free or bound electric charges q [Coulombs]. These forces are completely characterized by the *Lorentz force law*:

$$\overline{f}[N] = q(\overline{E} + \overline{v} \times \mu_o \overline{H}) \qquad \text{(Lorentz force law)} \qquad (1.2.1)$$

Thus we can define *electric field* $\overline{E}$ (volts/meter) in terms of the observable force vector $\overline{f}$:

$$\overline{E}[v/m] = \overline{f}/q \qquad\qquad \text{(electric field)} \qquad (1.2.2)$$

for the special case of a charge q with velocity $\overline{v} = 0$.

Similarly we can define *magnetic field* $\overline{H}$ [A m^{-1}] in terms of the observed force vector $\overline{f}$ given by the Lorentz force equation when $\overline{E} = 0$; $\overline{H}$ can be sensed only by charges in motion relative to the observer. Although a single measurement of force on a motionless charge suffices to determine $\overline{E}$, measurements of two charge velocity vectors $\overline{v}$ or current directions $\overline{I}$ are required to determine $\overline{H}$. For example, the arbitrary test charge velocity vector $\hat{x}v_1$ yields $f_1 = q\mu_o v_1 H \sin\theta$, where θ is the angle between $\overline{v}$ and $\overline{H}$ in the $\hat{x}$-$\hat{y}$ plane (see Figure 1.2.1). The unit vector $\hat{y}$ is defined as being in the observed direction $\overline{f}_1 \times \overline{v}_1$ where $\overline{f}_1$ defines the $\hat{z}$ axis. A second measurement with the test charge velocity vector $\hat{y}v_2$ yields $f_2 = q\mu_o v_2 H \cos\theta$. If $v_1 =$

v_2 then the force ratio $f_1/f_2 = \tan \theta$, yielding θ within the $\hat{x}$-$\hat{y}$ plane, plus the value of H: H = $f_1/(q\mu_o v_1 \sin \theta)$. There is no other physical method for detecting or measuring static electric or magnetic fields; we can only measure the forces on charges or on charged bodies, or measure the consequences of that force, e.g., by measuring the resulting currents.
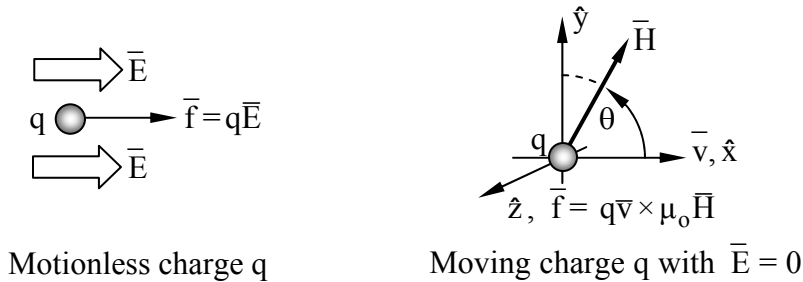
Figure 1.2.1   Measurement of electric and magnetic fields using charges.

It is helpful to have a simple physical picture of how fields behave so that their form and behavior can be guessed or approximately understood without recourse to mathematical solutions. Such physical pictures can be useful even if they are completely unrelated to reality, provided that they predict all observations in a simple way. Since the Lorentz force law plus Maxwell's equations explain essentially all non-relativistic and non-quantum electromagnetic behavior in a simple way using the fields $\bar{E}$ and $\bar{H}$, we need only to ascertain how $\bar{E}$ and $\bar{H}$ behave given a particular distribution of stationary or moving charges q.

First consider static distributions of charge. Electric field lines are parallel to $\bar{E}$, and the strength of $\bar{E}$ is proportional to the density of those field lines. Electric field lines begin on positive charges and terminate on negative ones, and the more charge there is, the more field lines there are. Field strength is proportional to lines per square meter. These lines pull on those charges to which they are attached, whether positive or negative, much as would a rubber band. Like rubber bands, they would also like to take the shortest path between two points, except that they also tend to repel their neighbors laterally, as do the charges to which they are attached.
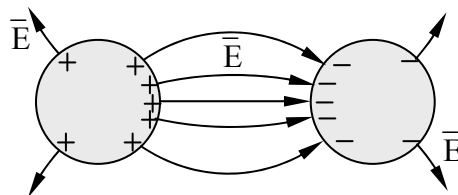
Figure 1.2.2   Electric field lines between two conducting cylinders.

Figure 1.2.2 illustrates the results of this mutual field-line repulsion, even as they pull opposite charges on conducting cylinders toward one another. Later we shall see that such electric field

lines are always perpendicular to perfectly conducting surfaces. Although these lines are illustrated as discrete, they actually are a continuum, even if only two charges are involved.

The same intuition applies to magnetic field lines $\overline{H}$. For example, Figure 1.2.2 would apply if the two cylinders corresponded instead to the north (+) and south (-) poles of a magnet, and if $\overline{E}$ became $\overline{H}$, although $\overline{H}$ need not emerge perfectly perpendicular to the magnet surface. In this case too the field lines would physically pull the two magnet poles toward one another. Both electric and magnetic motors can be driven using either the attractive force along field lines or the lateral repulsive force between lines, depending on motor design, as discussed later.

Another intuitive picture applies to time-dependent electromagnetic waves, where distributions of position-dependent electric and magnetic fields at right angles propagate as plane waves in the direction $\overline{E} \times \overline{H}$ much like a rigid body at the speed of light c, $\sim 3 \times 10^8$ m/s. Because electromagnetic waves can superimpose, it can be shown that any distribution of electric and magnetic fields can be considered merely as the superposition of such plane waves. Such plane waves are introduced in Section 2.2. If we examine such superpositions on spatial scales small compared to a wavelength, both the electric and magnetic fields behave much as they would in the static case.

Example 1.2A

A typical old vacuum tube accelerates electrons in a $\sim 10^4$ v m^{-1} electric field. What is the resulting electron velocity v(t) if it starts from rest? How long (τ) does it take the electron to transit the 1-cm tube?

Solution: Force f = ma = qE, and so v = at = qEt/m $\cong 1.6 \times 10^{-19} \times 10^4$ t/$(9.1 \times 10^{-28}) \cong 1.8 \times 10^{12}$ t [m s^{-1}]. Obviously v cannot exceed the speed of light c, $\sim 3 \times 10^8$ m/s. In this text we deal only with non-relativistic electrons traveling much slower than c. Distance traveled = d = aτ^2/2 = 0.01, so the transit time τ = $(2d/a)^{0.5}$ = $(2dm/qE)^{0.5} \cong$ $[2 \times 0.01 \times 9.1 \times 10^{-28}/(1.6 \times 10^{-19} \times 10^4)]^{0.5} \cong 1.1 \times 10^{-7}$ seconds. This slow transit limited most vacuum tubes to signal frequencies below several megahertz, although smaller gaps and higher voltages have enabled simple tubes to reach 100 MHz and higher. The microscopic gaps of semiconductors can eliminate transit time as an issue for most applications below 1 GHz; other phenomena often determine the frequency range instead.

1.3   Gauss's Law and electrostatic fields and potentials

While the Lorentz force law defines how electric and magnetic fields can be observed, Maxwell's four equations explain how these fields can be created directly from charges and currents, or indirectly and equivalently from other time varying fields. One of those four equations is *Gauss's Law for charge*, which states that the total charge Q [Coulombs] within volume V equals the integral of the normal component of the *electric displacement vector* $\overline{D}$ over the surface area A of that volume:

$$\oiint_A (\overline{D} \bullet \hat{n})\, da = \iiint_V \rho\, dv = Q \qquad \text{(\textit{Gauss's Law} for charge)} \qquad (1.3.1)$$

In vacuum:

$$\overline{D} = \varepsilon_o \overline{E} \qquad (1.3.2)$$

where the permittivity of vacuum $\varepsilon_o = 8.854 \times 10^{-12}$ Farads/m. Equation (1.3.1) reveals the dimensions of $\overline{D}$: Coulombs/m^2, often abbreviated here as [C/m^2].

A few simple examples illustrate typical electric fields for common charge distributions, and how Gauss's law can be used to compute those fields. First consider a sphere of radius R uniformly filled with charge of density ρ_o [C/m^3], as illustrated in Figure 1.3.1(a).

(a)   (b)
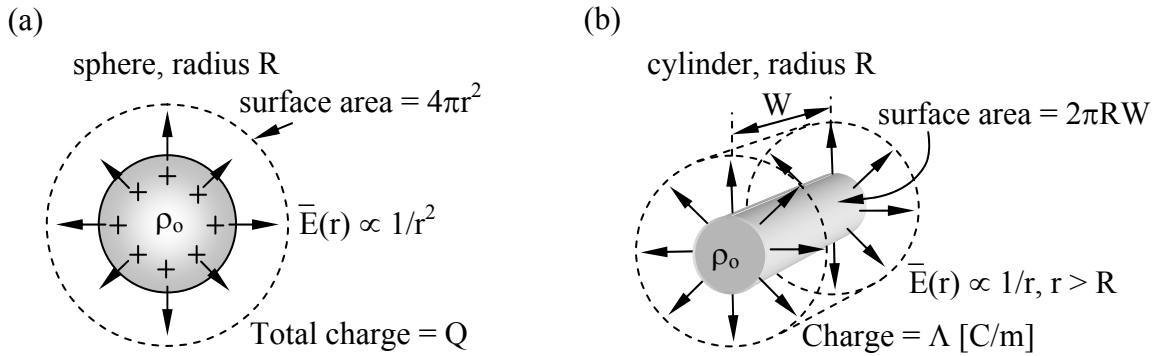
Figure 1.3.1   Electric fields $\overline{E}(r)$ produced by uniformly charged spheres and cylinders.

The symmetry of the solution must match the spherical symmetry of the problem, so $\overline{E}$ must be independent of θ and ϕ, although it can depend on radius r. This symmetry requires that $\overline{E}$ be radial and, more particularly:

$$\overline{E}(r,\theta,\phi) = \hat{r}E(r) \quad [V/m] \qquad (1.3.3)$$

We can find $\overline{E}(r)$ by substituting (1.3.3) into (1.3.1). First consider r > R, for which (1.3.1) becomes:

$$4\pi r^2 \varepsilon_o E(r) = (4/3)\pi R^3 \rho_o = Q \qquad (1.3.4)$$

$$\overline{E}(r) = \hat{r}\frac{Q}{4\pi\varepsilon_o r^2} \quad (r > R) \qquad (1.3.5)$$

Inside the sphere the same substitution into (1.3.1) yields:

$$4\pi r^2 \varepsilon_o E(r) = (4/3)\pi r^3 \rho_o \qquad (1.3.6)$$

$$\overline{E}(r) = \hat{r} \, \rho_o r / 3\varepsilon_o \quad [V/m] \quad (r < R) \tag{1.3.7}$$

It is interesting to compare this dependence of $\overline{E}$ on r with that for cylindrical geometries, which are also illustrated in Figure 1.3.1. We assume a uniform charge density of ρ_o within radius R, corresponding to Λ coulombs/meter. Substitution of (1.3.4) into (1.3.1) yields:

$$2\pi r W \varepsilon_o E(r) = \pi R^2 \rho_o W = \Lambda W \quad [C] \quad (r > R) \tag{1.3.8}$$

$$\overline{E}(r) = \hat{r} \frac{\Lambda}{2\pi\varepsilon_o r} = \hat{r} \frac{R^2 \rho_o}{2\varepsilon_o r} \quad [V/m] \quad (r > R) \tag{1.3.9}$$

Inside the cylinder ($r < R$) the right-hand-side of (1.3.9) still applies, but with R^2 replaced with r^2, so $\overline{E}(r) = \hat{r} \, r\rho_o/2\varepsilon_o$ instead.

To find the voltage difference, often called the difference in *electrical potential* Φ or the *potential difference*, between two points in space [V], we can simply integrate the static electric field $\overline{E} \bullet \hat{r}$ [V/m] along the field line $\overline{E}$ connecting them. Thus in the spherical case the voltage difference $\Phi(r_1) - \Phi(r_2)$ between points at r_1 and at $r_2 > r_1$ is:

$$\Phi(r_1) - \Phi(r_2) = \int_{r_1}^{r_2} \overline{E} \bullet d\overline{r} = \frac{Q}{4\pi\varepsilon_o} \int_{r_1}^{r_2} \frac{1}{r^2} \hat{r} \bullet d\overline{r} = -\frac{Q}{4\pi\varepsilon_o r} \Big|_{r_1}^{r_2} = \frac{Q}{4\pi\varepsilon_o} \left(\frac{1}{r_1} - \frac{1}{r_2} \right) [V] \tag{1.3.10}$$

If we want to assign an absolute value to electrical potential or voltage V at a given location, we usually define the potential Φ to be zero at $r_2 = \infty$, so a spherical charge Q produces an electric potential $\Phi(r)$ for $r > R$ which is:

$$\Phi(r) = Q/4\pi\varepsilon_o r \quad [V] \tag{1.3.11}$$

The same computation for the cylindrical charge of Figure 1.3.1 and the field of (1.3.9) yields:

$$\Phi(r_1) - \Phi(r_2) = \int_{r_1}^{r_2} \overline{E} \bullet d\overline{r} = \frac{\Lambda}{2\pi\varepsilon_o} \int_{r_1}^{r_2} \frac{1}{r} \hat{r} \bullet d\overline{r} = \frac{\Lambda \ln r}{2\pi\varepsilon_o} \Big|_{r_1}^{r_2} = \frac{\Lambda}{2\pi\varepsilon_o} \ln(r_2/r_1) \tag{1.3.12}$$

A third simple geometry is that of charged infinite parallel conducting plates separated by distance d, where the inner-facing surfaces of the upper and lower plates have surface charge density $+\rho_s$ and $-\rho_s$ [C/m^2], respectively, as illustrated in Figure 1.3.2 for finite plates. The uniformity of infinite plates with respect to x, y, and ϕ requires that the solution $\overline{E}$ also be independent of x, y, and ϕ. The symmetry with respect to ϕ requires that $\overline{E}$ point in the $\pm z$ direction. Gauss's law (1.3.1) then requires that $\overline{E}$ be independent of z because the integrals of $\overline{D}$ over the top and bottom surfaces of any rectangular volume located between the plates must cancel since there is no charge within such a volume and no $\overline{D}$ passing through its sides.
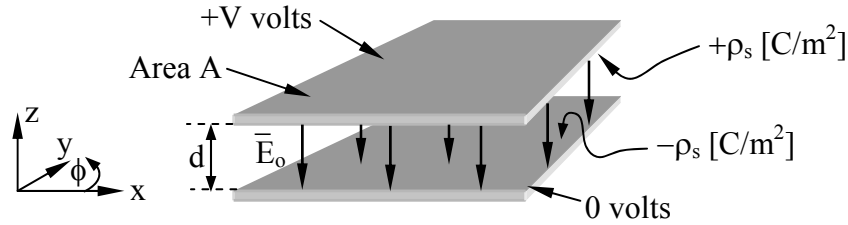
Figure 1.3.2   Electric field between parallel plates.

This solution for $\overline{E}$ is consistent with the *rubber-band model* for field lines, which suggests that the excess positive and negative charges will be mutually attracted, and therefore will be pulled to the inner surfaces of the two plates, particularly if the gap d between the plates is small compared to their width.   Gauss's Law (1.3.1) also tells us that the displacement vector $\overline{D}$ integrated over a surface enclosing the entire structure must be zero because the integrated charge within that surface is zero; that is, the integrated positive charge, $\rho_s A$, balances the integrated negative charge, $- \rho_s A$ and $\overline{D}$ external to the device can be zero everywhere.   The electric potential difference V between the two plates can be found by integrating $\overline{E}$ between the two plates.   That is, $V = E_o d$ volts for any path of integration, where $E_o = \rho_s/\varepsilon_o$ by Gauss's law.

Although the voltage difference between equipotentials can be computed by integrating along the electric field lines themselves, as done above, it is easy to show that the result does not depend on the path of integration.   Assume there are two different paths of integration P_1 and P_2 between any two points of interest, and that the two resulting voltage differences are V_1 and V_2. Now consider the closed contour C of integration that is along path P_1 in the positive direction and along P_2 in the reverse direction so as to make a closed loop.   Since this contour integral must yield zero, as shown below in (1.3.13) using Faraday's law for the static case where $\partial/\partial t = 0$, it follows that $V_1 = V_2$ and that all paths of integration yield the same voltage difference.

$$V_1 - V_2 = \int_{P_1} \overline{E} \cdot d\overline{s} \ - \int_{P_2} \overline{E} \cdot d\overline{s} \ = \ \oint_c \overline{E} \cdot d\overline{s} \ = \ -\frac{\partial}{\partial t} \iint_A \overline{B} \cdot d\overline{a} \ = \ 0 \qquad (1.3.13)$$

In summary, electric fields decay as $1/r^2$ from spherical charge concentrations, as $1/r$ from cylindrical ones, and are uniform in planar geometries.   The corresponding electric potentials decay as $1/r$, $-\ln r$, and x, respectively, as a result of integration over distance.   The potential Φ for the cylindrical case becomes infinite as $r \rightarrow \infty$ because the cylinder is infinitely long; the expression for the potential difference between concentric cylinders of finite radius is valid, however.   Within both uniform spherical and cylindrical charge distributions the electric field increases from zero linearly with radius r.   In each case the electric field distribution is explained by the rubber-band model in which the rubber bands (field lines) repel each other laterally while being pulled on by opposite electric charges.

It is extremely useful to note that Maxwell's equations are linear, so that superposition applies.   That is, the total electric field $\overline{E}$ equals that due to the sum of all charges present, where the contribution to $\overline{E}$ from each charge Q is given by (1.3.5).   Electric potentials Φ also superimpose, where the contribution from each charge Q is given by (1.3.11).

1.4   Ampere's Law and magnetostatic fields

The relevant Maxwell's equation for static current densities $\overline{J}$ [A/m^2] is Ampere's law, which says that for time-invariant cases the integral of magnetic field $\overline{H}$ around any closed contour in a right-hand sense equals the area integral of current density $\overline{J}$ [A/m^2] flowing through that contour:

$$\oint_c \overline{H} \cdot d\overline{s} = \iint_A \overline{J} \cdot d\overline{a} \tag{1.4.1}$$

Figure 1.4.1 illustrates a simple cylindrical geometry for which we can readily compute $\overline{H}$ produced by current I; the radius of the cylinder is R and the uniform current density flowing through it is J_o [A/m^2]. The cylinder is infinitely long.
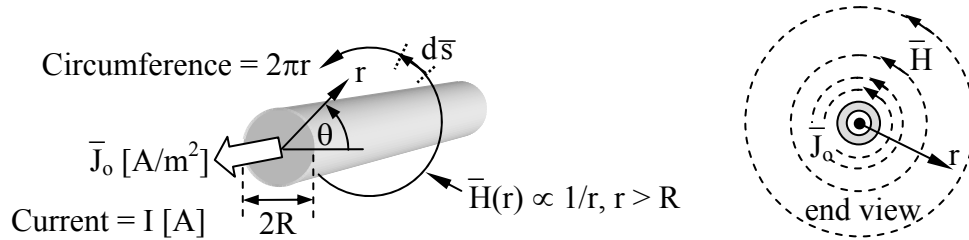
Figure 1.4.1   Magnetic field produced by a uniform cylindrical current.

Because the problem is cylindrically symmetric (not a function of θ), and uniform with respect to the cylindrical axis z, so is the solution. Thus $\overline{H}$ depends only upon radius r. Substitution of $\overline{H}(r)$ into (1.4.1) yields:

$$\int_0^{2\pi} \overline{H}(r) \cdot \hat{\theta} \, r d\theta = \int_0^{2\pi} \int_0^R J_o \, r \, dr \, d\theta = J_o \pi R^2 = I \, [A] \tag{1.4.2}$$

where the total current I is simply the uniform current density J_o times the area πR^2 of the cylinder. The left-hand-side of (1.4.2) simply equals H(r) times the circumference of a circle of radius r, so (1.4.2) becomes:

$$\overline{H}(r) = \hat{\theta} \, \frac{I}{2\pi r} = \hat{\theta} \, \frac{J_o \pi R^2}{2\pi r} \quad [A/m] \quad (r > R) \tag{1.4.3}$$

Within the cylindrical wire where r < R, (1.4.2) becomes:

$$H(r)2\pi r = \int_0^{2\pi} \int_0^r J_o \, r \, dr \, d\theta = J_o \pi r^2 \tag{1.4.4}$$

$$\overline{H}(r) = \hat{\theta} \, J_o r/2 \quad [A/m] \quad (r < R) \tag{1.4.5}$$

- 21 -

Therefore H(r) increases linearly with r within the wire and current distribution, and is continuous at r = R, where both (1.4.3) and (1.4.5) agree that H(r) = J_oR/2.

Another simple geometry involves parallel plates. Assume equal and opposite current densities, J_s [A/m], flow in infinite parallel plates separated by distance d, as illustrated in Figure 1.4.2 for finite plates. The integral of Ampere's law (1.4.1) around any contour C_1 circling both plates is zero because the net current through that contour is zero. A non-zero integral would require an external source of field, which we assume does not exist here. Thus $\overline{H}$ above and below the plates is zero. Since the integral of (1.4.1) around any contour C_2 that circles the upper plate yields H_xW = J_sW, where the x component of the magnetic field anywhere between the plates is H_x = J_s [A/m]; thus the magnetic field $\overline{H}$ between the plates is uniform. An integral around any contour in any y-z plane would circle no net current, so H_z = 0, and a similar argument applies to H_y, which is also zero. This configuration is discussed further in Section 3.2.1.
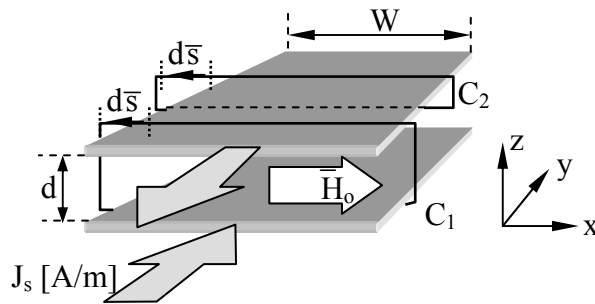
Figure 1.4.2   Static magnetic field between parallel plates.

More generally, because Maxwell's equations are linear, the total magnetic field $\overline{H}$ at any location is the integral of contributions made by current densities $\overline{J}$ nearby. Section 10.1 proves the Biot-Savart law (1.4.6), which defines how a current distribution $\overline{J}'$ at position $\overline{r}'$ within volume V' contributes to $\overline{H}$ at position $\overline{r}$:

$$\overline{H}(\overline{r}, t) = \iiint_{V'} \frac{\overline{J}' \times (\overline{r} - \overline{r}')}{4\pi |\overline{r} - \overline{r}'|^3} \, dv' \qquad \text{(Biot-Savart law)} \qquad (1.4.6)$$

To summarize, electric and magnetic fields are simple fictions that explain all electromagnetic behavior as characterized by Maxwell's equations and the Lorentz force law, which are examined further in Chapter 2. A simple physical model for the static behavior of electric fields is that of rubber bands that tend to pull opposite electric charges toward one another, but that tend to repel neighboring field lines laterally. Static magnetic fields behave similarly, except that the role of magnetic charges (which have not been shown to exist) is replaced by current loops acting as magnetic dipoles in ways that are discussed later.

Chapter 2: Introduction to Electrodynamics

2.1   Maxwell's differential equations in the time domain

Whereas the Lorentz force law characterizes the observable effects of electric and magnetic fields on charges, Maxwell's equations characterize the origins of those fields and their relationships to each other. The simplest representation of Maxwell's equations is in differential form, which leads directly to waves; the alternate integral form is presented in Section 2.4.3.

The differential form uses the vector *del operator* ∇:

$$\nabla \equiv \hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z} \tag{2.1.1}$$

where $\hat{x}$, $\hat{y}$, and $\hat{z}$ are defined as unit vectors in cartesian coordinates. Relations involving ∇ are summarized in Appendix D. Here we use the conventional vector *dot product*[1] and *cross product*[2] of ∇ with the electric and magnetic field vectors where, for example:

$$\overline{E} = \hat{x}E_x + \hat{y}E_y + \hat{z}E_z \tag{2.1.2}$$

$$\nabla \bullet \overline{E} \equiv \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \tag{2.1.3}$$

We call $\nabla \bullet \overline{E}$ the *divergence* of $\overline{E}$ because it is a measure of the degree to which the vector field $\overline{E}$ diverges or flows outward from any position. The cross product is defined as:

$$\begin{aligned}
\nabla \times \overline{E} \quad &\equiv \hat{x}\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right) + \hat{y}\left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right) + \hat{z}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) \\
&= \det\begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ E_x & E_y & E_z \end{vmatrix}
\end{aligned} \tag{2.1.4}$$

which is often called the *curl* of $\overline{E}$. Figure 2.1.1 illustrates when the divergence and curl are zero or non-zero for five representative field distributions.

---

[1] The dot product of $\overline{A}$ and $\overline{B}$ can be defined as $\overline{A} \bullet \overline{B} = A_x B_x + A_y B_y + A_z B_z = |A||B|\cos\theta$, where θ is the angle between the two vectors.

[2] The cross product of $\overline{A}$ and $\overline{B}$ can be defined as $\overline{A} \times \overline{B} = \underline{\hat{x}}\left(A_y B_z - A_z B_y\right) + \hat{y}\left(A_z B_x - A_x B_z\right) + \hat{z}\left(A_x B_y - A_y B_x\right)$; its magnitude is $|\overline{A}| \bullet |\overline{B}|\sin\theta$. Alternatively, $\overline{A} \times \overline{B} = \det|[A_x, A_y, A_z], [B_x, B_y, B_z], [\hat{x}, \hat{y}, \hat{z}]|$.
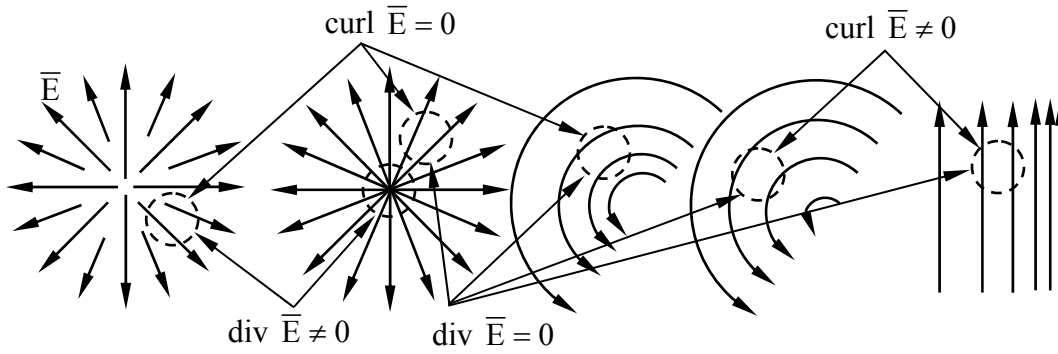
Figure 2.1.1   Fields with zero or non-zero divergence or curl.

The differential form of *Maxwell's equations* in the time domain are:

$$\nabla \times \overline{E} = -\frac{\partial \overline{B}}{\partial t} \qquad\qquad\qquad\qquad \textit{Faraday's Law} \qquad (2.1.5)$$

$$\nabla \times \overline{H} = \overline{J} + \frac{\partial \overline{D}}{\partial t} \qquad\qquad\qquad\qquad \textit{Ampere's Law} \qquad (2.1.6)$$

$$\nabla \bullet \overline{D} = \rho \qquad\qquad\qquad\qquad\qquad \textit{Gauss's Law} \qquad (2.1.7)$$

$$\nabla \bullet \overline{B} = 0 \qquad\qquad\qquad\qquad\qquad \textit{Gauss's Law} \qquad (2.1.8)$$

The field variables are defined as:

$\overline{E}$ *electric field* $\qquad\qquad\qquad\qquad$ [volts/meter; Vm^{-1}]   (2.1.9)

$\overline{H}$ *magnetic field* $\qquad\qquad\qquad\qquad$ [amperes/meter; Am^{-1}]   (2.1.10)

$\overline{B}$ *magnetic flux density* $\qquad\qquad\qquad\qquad$ [Tesla; T]   (2.1.11)

$\overline{D}$ *electric displacement* $\qquad\qquad$ [coulombs/m^2; Cm^{-2}]   (2.1.12)

$\overline{J}$ *electric current density* $\qquad\qquad$ [amperes/m^2; Am^{-2}]   (2.1.13)

ρ *electric charge density* $\qquad\qquad$ [coulombs/m^3; Cm^{-3}]   (2.1.14)

These four Maxwell equations invoke one scalar and five vector quantities comprising 16 variables.   Some variables only characterize how matter alters field behavior, as discussed later in Section 2.5. In vacuum we can eliminate three vectors (9 variables) by noting:

$$\overline{D} = \varepsilon_o \overline{E} \qquad\qquad\qquad \text{(constitutive relation for } \overline{D}) \quad (2.1.15)$$

$$\overline{B} = \mu_o \overline{H} \qquad\qquad\qquad \text{(constitutive relation for } \overline{B}) \quad (2.1.16)$$

$$\overline{J} = \rho \overline{v} = \sigma \overline{E} \qquad\qquad\qquad \text{(constitutive relation for } \overline{J}) \quad (2.1.17)$$

where $\varepsilon_o = 8.8542 \times 10^{-12}$ [farads m^{-1}] is the *permittivity* of vacuum, $\mu_o = 4\pi \times 10^{-7}$ [henries m^{-1}] is the *permeability* of vacuum[3], $\overline{v}$ is the velocity of the local net charge density ρ, and σ is the *conductivity* of a medium [Siemens m^{-1}]. If we regard the electrical sources ρ and $\overline{J}$ as given, then the equations can be solved for all remaining unknowns. Specifically, we can then find $\overline{E}$ and $\overline{H}$, and thus compute the forces on all charges present. Except for special cases we shall avoid solving problems where the electromagnetic fields and the motions of ρ are interdependent.

The constitutive relations for vacuum, $\overline{D} = \varepsilon_0 \overline{E}$ and $\overline{B} = \mu_0 \overline{H}$, can be generalized to $\overline{D} = \varepsilon \overline{E}$, $\overline{B} = \mu \overline{H}$, and $\overline{J} = \sigma \overline{E}$ for simple media. Media are discussed further in Section 2.5.

Maxwell's equations require conservation of charge. By taking the divergence of Ampere's law (2.1.6) and noting the vector identity $\nabla \bullet (\nabla \times \overline{A}) = 0$, we find:

$$\nabla \bullet (\nabla \times \overline{H}) = 0 = \nabla \bullet \frac{\partial \overline{D}}{\partial t} + \nabla \bullet \overline{J} \qquad\qquad\qquad (2.1.18)$$

Then, by reversing the sequence of the derivatives in (2.1.18) and substituting Gauss's law $\nabla \bullet \overline{D} = \rho$ (2.1.7), we obtain the differential expression for *conservation of charge*:

$$\nabla \bullet \overline{J} = -\frac{\partial \rho}{\partial t} \qquad\qquad\qquad \text{(conservation of charge)} \quad (2.1.19)$$

The integral expression can be derived from the differential expression by using *Gauss's divergence theorem*, which relates the integral of $\nabla \bullet \overline{G}$ over any volume V to the integral of $\overline{G} \bullet \hat{n}$ over the surface area A of that volume, where the surface normal unit vector $\hat{n}$ points outward:

$$\iiint_V \nabla \bullet \overline{G} \, dv = \oiint_A \overline{G} \bullet \hat{n} \, da \qquad\qquad \text{(Gauss's divergence theorem)} \quad (2.1.20)$$

Thus the integral expression for conservation of charge is:

$$\frac{d}{dt} \iiint_V \rho \, dv = -\oiint_A \overline{J} \bullet \hat{n} \, da \qquad\qquad \text{(conservation of charge)} \quad (2.1.21)$$

---

[3] The constant $4\pi \times 10^{-7}$ is exact and enters into the definition of an ampere.

which says that if no net current $\overline{J}$ flows through the walls A of a volume V, then the total charge inside must remain constant.

Example 2.1A

If the electric field in vacuum is $\overline{E} = \hat{x}E_o \cos(\omega t - ky)$, what is $\overline{H}$?

Solution: From Faraday's law (2.1.5): $\mu_o(\partial\overline{H}/\partial t) = -(\nabla\times\overline{E}) = \hat{z}\,\partial E_x/\partial y = \hat{z}\,kE_o \sin(\omega t-ky)$, using (2.1.4) for the curl operator. Integration of this equation with respect to time yields: $\overline{H} = -\hat{z}(kE_o/\mu_o\omega)\cos(\omega t - ky)$.

Example 2.1B

Does the electric field in vacuum $\overline{E} = \hat{x}E_o \cos(\omega t - kx)$ satisfy Maxwell's equations? Under what circumstances would this $\overline{E}$ satisfy the equations?

Solution: This electric field does not satisfy Gauss's law for vacuum, which requires $\nabla\bullet\overline{D} = \rho = 0$. It satisfies Gauss's law only for non-zero charge density: $\rho = \nabla\bullet\overline{D} = \varepsilon_o\partial E_x/\partial x = \partial[\varepsilon_o E_o \cos(\omega t - kx)]/\partial x = k\varepsilon_o E_o \sin(\omega t - kx) \neq 0$. To satisfy the remaining Maxwell equations and conservation of charge (2.1.19) there must also be a current $\overline{J} \neq 0$ corresponding to $\rho: \overline{J} = \sigma\overline{E} = \hat{x}\sigma E_o \cos(\omega t - kx)$, where (2.1.17) simplified the computation.

2.2   *Electromagnetic waves in the time domain*

Perhaps the greatest triumph of Maxwell's equations was their ability to predict in a simple way the existence and velocity of electromagnetic waves based on simple laboratory measurements of the permittivity and permeability of vacuum. In vacuum the charge density $\rho = \overline{J} = 0$, and so Maxwell's equations become:

$$\nabla\times\overline{E} = -\mu_o \frac{\partial\overline{H}}{\partial t} \qquad\qquad \text{(Faraday's law in vacuum)} \qquad (2.2.1)$$

$$\nabla\times\overline{H} = \varepsilon_o \frac{\partial\overline{E}}{\partial t} \qquad\qquad \text{(Ampere's law in vacuum)} \qquad (2.2.2)$$

$$\nabla\bullet\overline{E} = 0 \qquad\qquad\qquad \text{(Gauss's law in vacuum)} \qquad (2.2.3)$$

$$\nabla\bullet\overline{H} = 0 \qquad\qquad\qquad \text{(Gauss's law in vacuum)} \qquad (2.2.4)$$

We can eliminate $\overline{H}$ from these equations by computing the curl of Faraday's law, which introduces $\nabla\times\overline{H}$ on its right-hand side so Ampere's law can be substituted:

$$\nabla \times (\nabla \times \overline{E}) = -\mu_o \frac{\partial (\nabla \times \overline{H})}{\partial t} = -\mu_o \varepsilon_o \frac{\partial \overline{E}^2}{\partial t^2} \tag{2.2.5}$$

Using the well known vector identity (see Appendix D):

$$\nabla \times (\nabla \times \overline{A}) = \nabla (\nabla \bullet \overline{A}) - \nabla^2 \overline{A} \qquad \text{("well-known vector identity")} \tag{2.2.6}$$

and then using (2.2.3) to eliminate $\nabla \bullet \overline{E}$, (2.2.5) becomes the electromagnetic *wave equation*, often called the *Helmholtz wave equation*:

$$\nabla^2 \overline{E} - \mu_o \varepsilon_o \frac{\partial^2 \overline{E}}{\partial t^2} = 0 \qquad \text{(Helmholtz wave equation)} \tag{2.2.7}$$

where:

$$\nabla^2 \overline{E} \equiv \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \left(\hat{x} E_x + \hat{y} E_y + \hat{z} E_z \right) \tag{2.2.8}$$

The solutions to this wave equation (2.2.7) are any fields $\overline{E}(\overline{r}, t)$ for which the second spatial derivative $(\nabla^2 \overline{E})$ equals a constant times the second time derivative $(\partial^2 \overline{E}/\partial t^2)$. The *position vector* $\overline{r} \equiv \hat{x}x + \hat{y}y + \hat{z}z$. The wave equation is therefore satisfied by any arbitrary $\overline{E}(\overline{r}, t)$ having identical dependence on space and time within a constant multiplier. For example, arbitrary functions of the arguments (z - ct), (z + ct), or (t ± z/c) have such an identical dependence and are among the valid solutions to (2.2.7), where c is some constant to be determined. One such solution is:

$$\overline{E}(\overline{r}, t) = \overline{E}(z - ct) = \hat{x} E_x (z - ct) \tag{2.2.9}$$

where the arbitrary function $E_x(z - ct)$ might be that illustrated in Figure 2.2.1 at time t = 0 and again at some later time t. Note that as time advances within the argument (z - ct), z must advance with ct in order for that argument, or $\overline{E}$ at any point of interest on the waveform, to remain constant.
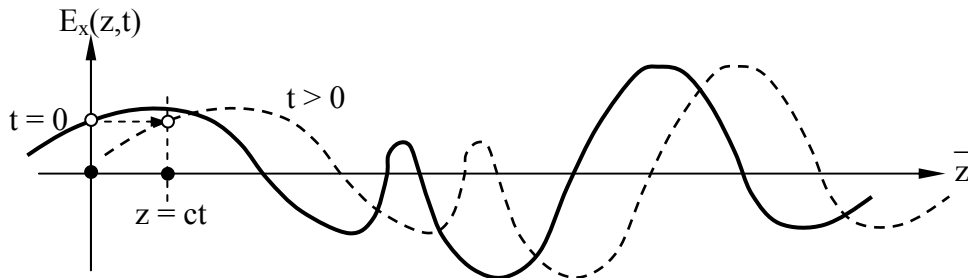
Figure 2.2.1   Arbitrary electromagnetic wave propagating in the +z direction.

We can test this candidate solution (2.2.9) by substituting it into the wave equation (2.2.7), yielding:

$$\nabla^2 \overline{E}(z-ct) = \frac{\partial^2 \left[\overline{E}(z-ct)\right]}{\partial z^2} \equiv \overline{E}''(z-ct)$$
$$= \mu_o \varepsilon_o \frac{\partial^2 \left[\overline{E}(z-ct)\right]}{\partial t^2} = \mu_o \varepsilon_o (-c)^2 \overline{E}''(z-ct)$$

(2.2.10)

where we define $\overline{A}'(q)$ as the first derivative of $\overline{A}$ with respect to its argument q and $\overline{A}''(q)$ as its second derivative. Equation (2.2.10) is satisfied if:

$$c = \frac{1}{\sqrt{(\mu_o \varepsilon_o}} \quad [\text{m/s}]$$

(2.2.11)

where we define c as the *velocity of light* in vacuum:

$$c = 2.998 \times 10^8 \ [\text{m s}^{-1}]$$

(velocity of light)   (2.2.12)

Figure 2.2.1 illustrates how an arbitrary $\overline{E}(z,t)$ can propagate by translating at velocity c. However, some caution is warranted when $\overline{E}(z,t)$ is defined. Although our trial solution (2.2.9) satisfies the wave equation (2.2.7), it may not satisfy Gauss's laws. For example, consider the case where:

$$\overline{E}(z,t) = \hat{z}E_z(z-ct)$$

(2.2.13)

Then Gauss's law $\nabla \bullet \overline{E} = 0$ is not satisfied:

$$\nabla \bullet \overline{E} = \frac{\partial \overline{E}_z}{\partial z} \neq 0 \text{ for arbitrary } \overline{E}(z)$$

(2.2.14)

In contrast, if $\overline{E}(z,t)$ is oriented perpendicular to the direction of propagation (in the $\hat{x}$ and/or $\hat{y}$ directions for z-directed propagation), then all Maxwell's equations are satisfied and the solution is valid. In the case $\overline{E}(z,t) = \hat{y}E_y(z-ct)$, independent of x and y, we have a *uniform plane wave* because the fields are uniform with respect to two of the coordinates (x,y) so that $\partial\overline{E}/\partial x = \partial\overline{E}/\partial y = 0$. Since this electric field is in the y direction, it is said to be y-polarized; by convention, *polarization* of a wave refers to the direction of its electric vector. Polarization is discussed further in Section 2.3.4.

Knowing $\overline{E}(z,t) = \hat{y}E_y(z-ct)$ for this example, we can now find $\overline{H}(z,t)$ using Faraday's law (2.2.1):

$$\frac{\partial \overline{H}}{\partial t} = -\frac{(\nabla \times \overline{E})}{\mu_o} \tag{2.2.15}$$

We can evaluate the curl of $\overline{E}$ using (2.1.4) and knowing $E_x = E_z = \frac{\partial}{\partial x} = \frac{\partial}{\partial y} = 0$:

$$\nabla \times \overline{E} = \hat{x}\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right) + \hat{y}\left(\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x}\right) + \hat{z}\left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right) = -\hat{x}\frac{\partial E_y}{\partial z} \tag{2.2.16}$$

Then, by integrating (2.2.15) over time it becomes:

$$\overline{H}(z,t) = -\int_{-\infty}^{t} \frac{(\nabla \times \overline{E})}{\mu_o}\,dt = \hat{x}\frac{1}{\mu_o}\int_{-\infty}^{t} \frac{\partial E_y(z-ct)}{\partial z}\,dt$$

$$= -\hat{x}\frac{1}{c\mu_o}E_y(z-ct) = -\hat{x}\sqrt{\frac{\varepsilon_o}{\mu_o}}E_y(z-ct) \tag{2.2.17}$$

$$\overline{H}(z,t) = \sqrt{\frac{\varepsilon_o}{\mu_o}}\hat{z} \times \overline{E}(z,t) = \hat{z} \times \frac{\overline{E}(z,t)}{\eta_o} \tag{2.2.18}$$

where we used the velocity of light $c = 1/\sqrt{\varepsilon_o\mu_o}$, and defined $\eta_o = \sqrt{\mu_o/\varepsilon_o}$.

Thus $\overline{E}$ and $\overline{H}$ in a uniform plane wave are very simply related. Their directions are orthogonal to each other and to the direction of propagation, and the magnitude of the electric field is $(\mu_o/\varepsilon_o)^{0.5}$ times that of the magnetic field; this factor $\eta_o = \sqrt{\mu_o/\varepsilon_o}$ is known as the *characteristic impedance of free space* and equals ~377 ohms. That is, for a single uniform plane wave in free space,

$$|\overline{E}|/|\overline{H}| = \eta_o = \sqrt{\mu_o/\varepsilon_o} \cong 377 \ \ [\text{ohms}] \tag{2.2.19}$$

Electromagnetic waves can propagate in any arbitrary direction in space with arbitrary time behavior. That is, we are free to define $\hat{x}$, $\hat{y}$, and $\hat{z}$ in this example as being in any three orthogonal directions in space. Because Maxwell's equations are linear in field strength, superposition applies and any number of plane waves propagating in arbitrary directions with arbitrary polarizations can be superimposed to yield valid electromagnetic solutions. Exactly which superposition is the valid solution in any particular case depends on the boundary conditions and the initial conditions for that case, as discussed later in Chapter 9 for a variety of geometries.

Example 2.2A

Show that $\overline{E} = \hat{y}E_o(t + z/c)$ satisfies the wave equation (2.2.7). In which direction does this wave propagate?

Solution: $\left(\nabla^2 - \dfrac{\partial^2}{c^2 \partial t^2}\right)\overline{E} = \hat{y}\dfrac{1}{c^2}\left[E_o''(t + z/c) - E_o''(t + z/c)\right] = 0$; Q.E.D[4]. Since the argument

(t + z/c) remains constant as t increases only if z/c decreases correspondingly, the wave is propagating in the -z direction.

2.3   Maxwell's equations, waves, and polarization in the frequency domain

2.3.1   Sinusoidal waves

Linear systems are easily characterized by the magnitude and phase of each output as a function of the frequency at which the input is sinusoidally stimulated. This simple characterization is sufficient because sinusoids of different frequencies can be superimposed to construct any arbitrary input waveform[5], and the output of a linear system is the superposition of its responses to each superimposed input. Systems with multiple inputs and outputs can be characterized in the same way. Nonlinear systems are more difficult to characterize because their output frequencies generally include harmonics of their inputs.

Fortunately free space is a linear system, and therefore it is fully characterized by its response to sinusoidal plane waves. For example, the arbitrary z-propagating x-polarized uniform plane wave of (2.2.9) and Figure 2.2.1 could be sinusoidal and represented by:

$$\overline{E}(\overline{r}, t) = \hat{x}E_o \cos\left[k(z - ct)\right] \tag{2.3.1}$$

$$\overline{H}(\overline{r}, t) = \hat{y}\sqrt{\varepsilon_o/\mu_o}\, E_o \cos\left[k(z - ct)\right] \tag{2.3.2}$$

where the *wave amplitude* E_o is a constant and the factor k is related to frequency, as shown below.

It is more common to represent sinusoidal waves using the argument (ωt - kz) so that their frequency and spatial dependences are more evident. The *angular frequency* ω is simply related to frequency f [Hz]:

$$\omega = 2\pi f \ \ [\text{radians s}^{-1}] \qquad\qquad \text{(angular frequency)} \tag{2.3.3}$$

---

[4] Q.E.D. is the abbreviation for the Latin phrase "*quod erat demonstratum*" or "that which was to be demonstrated."

[5] The Fourier transform pair (10.4.17) and (10.4.18) relate arbitrary pulse waveforms h(t) to their corresponding spectra $\underline{H}$(f), where each frequency f has its own magnitude and phase represented by $\underline{H}$(f).

and the *spatial frequency* k, often called the *wavenumber*, is simply related to ω and wavelength λ [m], which is the length of one period in space:

$$k = 2\pi/\lambda = \omega/c \quad [\text{radians m}^{-1}] \qquad \text{(wave number)} \qquad (2.3.4)$$

The significance and dimensions of ω and k are directly analogous; they are radians s^{-1} and radians m^{-1}, respectively.

Therefore we can alternatively represent the wave of (2.3.1) and (2.3.2) as:

$$\overline{E}(\overline{r}, t) = \hat{x} E_o \cos(\omega t - kz) \quad [\text{v m}^{-1}] \qquad (2.3.5)$$

$$\overline{H}(z, t) = \hat{y} \sqrt{\varepsilon_o/\mu_o} E_o \cos(\omega t - kz) \quad [\text{A m}^{-1}] \qquad (2.3.6)$$

Figure 2.3.1 suggests the form of this wave. Its *wavelength* is λ, the length of one cycle, where:

$$\lambda = c/f \quad [\text{m}] \qquad \text{(wavelength)} \qquad (2.3.7)$$

The figure illustrates how these electric and magnetic fields are in phase but orthogonal to each other and to the direction of propagation. When the argument $(\omega t - kz)$ equals zero, the fields are maximum, consistent with $\cos(\omega t - kz)$.
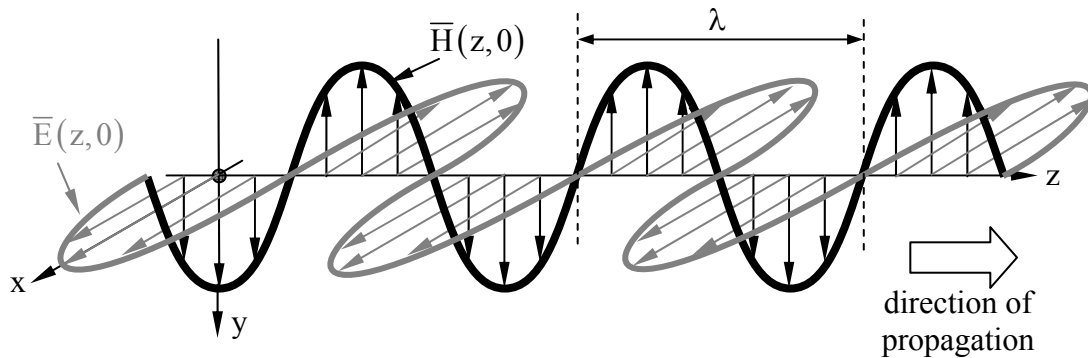
Figure 2.3.1   +z propagating y-polarized uniform plane wave of wavelength λ.

This notation makes it easy to characterize uniform plane waves propagating in other directions as well. For example:

$$\overline{E}(\overline{r}, t) = \hat{x} E_o \cos(\omega t + kz) \qquad \text{(x-polarized wave in -z direction)} \qquad (2.3.8)$$

$$\overline{E}(\overline{r},t) = \hat{y}E_o \cos(\omega t - kz) \qquad \text{(y-polarized wave in +z direction)} \qquad (2.3.9)$$

$$\overline{E}(\overline{r},t) = \hat{y}E_o \cos(\omega t - kx) \qquad \text{(y-polarized wave in +x direction)} \qquad (2.3.10)$$

$$\overline{E}(\overline{r},t) = \hat{z}E_o \cos(\omega t + kx) \qquad \text{(z-polarized wave in -x direction)} \qquad (2.3.11)$$

2.3.2   Maxwell's equations in the complex-frequency domain

Electromagnetic fields are commonly characterized in the frequency domain in terms of their magnitudes and phases as a function of position $\overline{r}$ for frequency f. For example, the $\hat{x}$ component of a general sinusoidally varying $\overline{E}$ might be:

$$\overline{E}(\overline{r},t) = \hat{x}E_x(\overline{r})\cos\left[\omega t + \phi(\overline{r})\right] \qquad (2.3.12)$$

This might become $\overline{E}(\overline{r},t) = \hat{x}E_x \cos(\omega t - kz)$ for a uniform plane wave propagating in the +z direction.

It is generally more convenient to express phase using *complex notation* (see Appendix B). The x-component of the wave of (2.3.12) can also be represented as:

$$E_x(\overline{r},t) = R_e\left\{E_x(\overline{r})e^{j(\omega t + \phi_x(\overline{r}))}\right\} = \hat{x}R_e\left\{\underline{E}_x(\overline{r})e^{j\omega t}\right\} \qquad (2.3.13)$$

where the spatial and frequency parts of $E_x(\overline{r},t)$ have been separated and $\underline{E}_x(\overline{r}) = \left|\underline{E}_x(\overline{r})\right|e^{j\phi_x(\overline{r})}$ is called a *phasor*. The simplicity will arise later when we omit $R_e\left\{[\ ]e^{j\omega t}\right\}$ from our expressions as "understood", so only the phasors remain. The underbar under $\underline{E}_x$ indicates $\underline{E}_x$ is not a function of time, but rather is a complex quantity with a real part and an imaginary part, where:

$$\underline{E}_x(\overline{r}) = R_e\left\{\underline{E}_x(\overline{r})\right\} + jI_m\left\{\underline{E}_x(\overline{r})\right\} = \left|\underline{E}_x(\overline{r})\right|e^{j\phi_x(\overline{r})} \qquad (2.3.14)$$

and $\phi_x(\overline{r}) = \tan^{-1}\left(Im\{\underline{E}_x(\overline{r})\}/Re\{\underline{E}_x(\overline{r})\}\right)$. A general vector can also be a phasor, e.g., $\overline{\underline{E}}(\overline{r}) = \hat{x}\underline{E}_x(\overline{r}) + \hat{y}\underline{E}_y(\overline{r}) + \hat{z}\underline{E}_z(\overline{r})$, where $\overline{E}(\overline{r},t) = R_e\left\{\overline{\underline{E}}(\overline{r})e^{j\omega t}\right\}$.

We can use such phasors to simplify Maxwell's equations. For example, we can express Faraday's law (2.2.1) as:

$$\nabla \times R_e\left\{\overline{\underline{E}}(\overline{r})e^{j\omega t}\right\} = -\partial R_e\left\{\overline{\underline{B}}(\overline{r})e^{j\omega t}\right\}\big/\partial t = R_e\left\{\nabla \times \overline{E}(\overline{r})e^{j\omega t}\right\} = R_e\left\{-j\omega\overline{\underline{B}}(\overline{r})e^{j\omega t}\right\} \qquad (2.3.15)$$

The other Maxwell equations can be similarly transformed, which suggests that the notation $R_e\{[\ ]e^{j\omega t}\}$ can be omitted and treated as understood. For example, removing this redundant notation from (2.3.15) results in: $\nabla\times\underline{E} = -j\omega\underline{B}$. Any problem solution expressed as a phasor, e.g. $\underline{E}(\bar{r})$, can be converted back into a time-domain expression by the operator $R_e\{[\ ]e^{j\omega t}\}$. These omissions of the understood notation result in the complex or *time-harmonic Maxwell equations*:

$$\nabla\times\underline{E} = -j\omega\underline{B} \qquad\qquad \text{(Faraday's law)} \quad (2.3.16)$$

$$\nabla\times\underline{H} = \underline{J} + j\omega\underline{D} \qquad\qquad \text{(Ampere's law)} \quad (2.3.17)$$

$$\nabla\bullet\underline{D} = \rho \qquad\qquad \text{(Gauss's law)} \quad (2.3.18)$$

$$\nabla\bullet\underline{B} = 0 \qquad\qquad \text{(Gauss's law)} \quad (2.3.19)$$

Note that these equations are the same as before [i.e., (2.2.1–4)], except that we have simply replaced the operator $\partial/\partial t$ with $j\omega$ and placed an underbar under all variables, signifying that they are now phasors.

We can immediately derive the time-harmonic equation for conservation of charge (2.1.19) by computing the divergence of (2.3.17), noting that $\nabla\bullet(\nabla\times\overline{A}) = 0$ for any $\overline{A}$, and substituting $\nabla\bullet\underline{D} = \rho$ (2.3.18):

$$\nabla\bullet\underline{J} + j\omega\underline{\rho} = 0 \qquad\qquad (2.3.20)$$

Example 2.3A

Convert the following expressions into their time-domain equivalents: $j\omega\nabla\times\overline{Q} = \overline{R}j$, $\overline{R}e^{-jkz}$, and $\underline{E} = \hat{x}3 + \hat{y}j4$.

Solution: $-\omega(\nabla\times\overline{Q})\sin(\omega t) = -\overline{R}\sin\omega t$, $\overline{R}\cos(\omega t - kz)$, and $3\hat{x}\cos\omega t - 4\hat{y}\sin\omega t$.

Example 2.3B

Convert the following expressions into their complex frequency-domain equivalents: $A\cos(\omega t + kz)$, and $B\sin(\omega t + \phi)$.

Solution: Ae^{+jkz}, and $-jBe^{j\phi} = -jB\cos\phi + B\sin\phi$.

2.3.3   Sinusoidal uniform plane waves

We can readily derive from Maxwell's equations the time-harmonic Helmholtz wave equation for vacuum (2.2.7) by substituting $j\omega$ for $\partial/\partial t$ or, as we did earlier, by taking the curl of Faraday's law, using the well known vector identity (2.2.6) and Gauss's law, replacing $\overline{B}$ by $\mu_o\overline{H}$, and using Ampere's law to replace $\nabla\times\overline{H}$.   In both cases the Helmholtz wave equation becomes:

$$\left(\nabla^2 + \omega^2\mu_o\varepsilon_o\right)\overline{E} = 0 \qquad\qquad \text{(wave equation)} \qquad (2.3.21)$$

As before, the solution $\overline{E}(\overline{r})$ to the wave equation can be any arbitrary function of space $(\overline{r})$ such that its second spatial derivative $\left(\nabla^2\overline{E}\right)$ equals a constant $\left(-\omega^2\varepsilon_o\mu_o\right)$ times that same function $\overline{E}(\overline{r})$.   One solution with these properties is the time-harmonic version of the time-domain expression $\overline{E}(\overline{r},t) = \hat{y}E_o\cos(\omega t - kz)$:

$$\overline{E}(\overline{r}) = \hat{y}E_oe^{-jkz} \quad \left[\text{v m}^{-1}\right] \qquad\qquad (2.3.22)$$

Substituting (2.3.22) into the wave equation (2.3.21) yields:

$$\left(\left[\partial^2/\partial z^2\right] + \omega^2\mu_o\varepsilon_o\right)\overline{E} = \left(\left[-jk\right]^2 + \omega^2\mu_o\varepsilon_o\right)\overline{E} = 0 \qquad\qquad (2.3.23)$$

which is satisfied if the wavenumber k is:

$$k = \omega\sqrt{\mu_o\varepsilon_o} = \frac{\omega}{c} = \frac{2\pi f}{c} = \frac{2\pi}{\lambda} \left[\text{radians m}^{-1}\right] \qquad\qquad (2.3.24)$$

It is now an easy matter to find the magnetic field that corresponds to (2.3.22) by using Faraday's law (2.3.16), $\overline{B} = \mu_o\overline{H}$, and the definition of the "$\nabla\times$" operator (2.1.1):

$$\begin{aligned}\overline{H}(\overline{r}) &= -\frac{\left(\nabla\times\overline{E}\right)}{j\omega\mu_o} = \frac{1}{j\omega\mu_o}\frac{\hat{x}\partial E_y}{\partial z} = -\frac{\hat{x}kE_oe^{-jkz}}{\omega\mu_o} \\ &= -\hat{x}\frac{1}{\eta_o}E_oe^{-jkz} \left[\text{Am}^{-1}\right]\end{aligned} \qquad\qquad (2.3.25)$$

As before, $\overline{E}$ and $\overline{H}$ are orthogonal to each other and to the direction of propagation, and $\left|\overline{E}\right| = \eta_o\left|\overline{H}\right|$.

As another example, consider a z-polarized wave propagating in the -x direction; then:

$$\overline{E}(\overline{r}) = \hat{z}E_oe^{+jkx} \ , \quad \overline{H}(\overline{r}) = \hat{y}E_oe^{jkx}\big/\eta_o \tag{2.3.26}$$

It is easy to convert phasor expressions such as (2.3.26) into time-domain expressions. We simply divide the phasor expressions into their real and imaginary parts, and note that the real part varies as cos(ωt - kz) and the imaginary part varies as sin(ωt - kz). Thus the fields in (2.3.22) could be written instead as a real time-domain expression:

$$\overline{E}(\overline{r},t) = \hat{y}E_o \cos(\omega t - kz) \tag{2.3.27}$$

Had the electric field solution been instead the phasor $\hat{y}jE_oe^{-jkz}$, the time domain expression $R_e\{\overline{E}(\overline{r})e^{j\omega t}\}$ would then be:

$$\overline{E}(\overline{r},t) = -\hat{y}E_o \sin(\omega t - kz) \tag{2.3.28}$$

The conversion of complex phasors to time-domain expressions, and vice-versa, is discussed further in Appendix B.

2.3.4   Wave polarization

Complex notation simplifies the representation of wave *polarization*, which characterizes the behavior of the sinusoidally varying electric field vector as a function of time. It is quite distinct from the polarization of media discussed in Section 2.5.3. Previously we have seen waves for which the time-varying electric vector points only in the ±x, ±y, or ±z directions, corresponding to x, y, or z polarization, respectively. By superimposing such waves at the same frequency and propagating in the same direction we can obtain any other desired time-harmonic polarization. Linear polarization results when the oscillating electric vector points only along a single direction in the plane perpendicular to the direction of propagation, while elliptical polarization results when the x and y components of the electric vector are out of phase so that the tip of the electric vector traces an ellipse in the same plane. Circular polarization results only when the phase difference between x and y is 90 degrees and the two amplitudes are equal. These various polarizations for $+\hat{z}$ propagation are represented below at z = 0 in the time domain and as phasors, and in Figure 2.3.2.

$$\overline{E}(t) = \hat{y}E_o \cos \omega t \qquad\qquad \underline{\overline{E}} = \hat{y}E_o \qquad\qquad \text{(y-polarized)} \quad (2.3.29)$$

$$\overline{E}(t) = \hat{x}E_o \cos \omega t \qquad\qquad \underline{\overline{E}} = \hat{x}E_o \qquad\qquad \text{(x-polarized)} \quad (2.3.30)$$

$$\overline{E}(t) = (\hat{x} + \hat{y})E_o \cos \omega t \qquad\qquad \underline{\overline{E}} = (\hat{x} + \hat{y})E_o \qquad\qquad \text{(45°-polarized)} \quad (2.3.31)$$

$$\overline{E}(t) = E_o(\hat{x} \cos \omega t + \hat{y} \sin \omega t) \qquad \underline{\overline{E}} = (\hat{x} - j\hat{y})E_o \qquad\qquad \text{(right-circular)} \quad (2.3.32)$$

$$\overline{E}(t) = E_o(\hat{x} \cos \omega t + 1.5\hat{y} \sin \omega t) \qquad \underline{\overline{E}} = (\hat{x} - 1.5j\hat{y})E_o \qquad \text{(right-elliptical)} \quad (2.3.33)$$

$$\overline{E}(t) = E_o \left[\hat{x} \cos \omega t + \hat{y} \cos (\omega t + 20°) \right] \qquad \overline{\underline{E}} = \left(\hat{x} + e^{0.35j} \hat{y} \right) E_o \qquad \text{(left-elliptical)} \qquad (2.3.34)$$
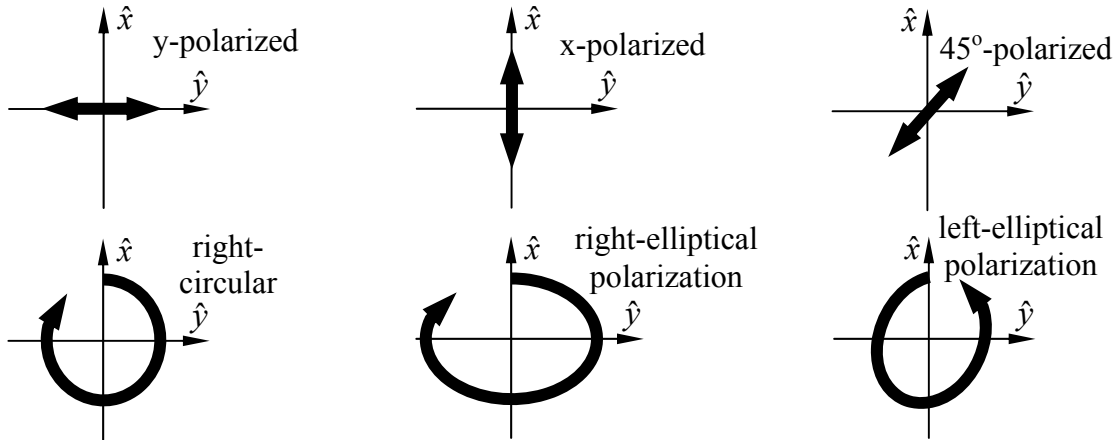
Figure 2.3.2   Polarization ellipses for +z-propagating plane waves (into the page).

The Institute of Electrical and Electronics Engineers (IEEE) has defined polarization as right-handed if the electric vector traces a right-handed ellipse in the x-y plane for a wave propagating in the +z direction, as suggested in Figure 2.3.3.   That is, for *right-handed polarization* the fingers of the right hand circle in the direction taken by the electric vector while the thumb points in the direction of propagation. This legal definition is opposite that commonly used in physics, where that alternative definition is consistent with the handedness of the "screw" formed by the instantaneous three-dimensional loci of the tips of the electric vectors comprising a wave.
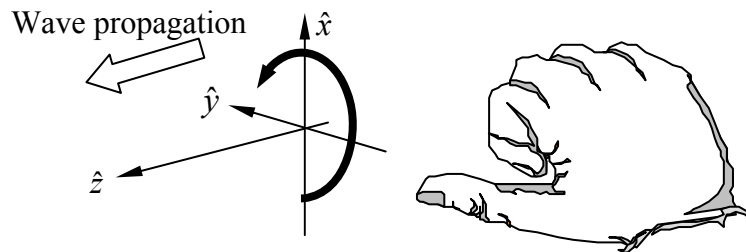
Figure 2.3.3   IEEE definition of right-handed polarization.

___

Example 2.3C

If $\overline{E} = \overline{\underline{E}}_o e^{-jkz}$, what polarizations correspond to: $\overline{\underline{E}}_o = \hat{y}$, $\overline{\underline{E}}_o = \hat{x} + 2\hat{y}$, and $\overline{\underline{E}}_o = \hat{x} - j\hat{y}$?

Solution: y polarization, linear polarization at angle $\tan^{-1}2$ relative to the x-z plane, and right-circular polarization.

___

2.4 *Relation between integral and differential forms of Maxwell's equations*

2.4.1   Gauss's divergence theorem

Two theorems are very useful in relating the differential and integral forms of Maxwell's equations: Gauss's divergence theorem and Stokes theorem.   Gauss's divergence theorem (2.1.20) states that the integral of the normal component of an arbitrary analytic vector field $\overline{A}$ over a surface S that bounds the volume V equals the volume integral of $\nabla \bullet \overline{A}$ over V.   The theorem can be derived quickly by recalling (2.1.3):

$$\nabla \bullet \overline{A} \equiv \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z} \tag{2.4.1}$$

Therefore $\nabla \bullet \overline{A}$ at the position x_o, y_o, z_o can be found using (2.4.1) in the limit where Δx, Δy, and Δz approach zero:

$$
\begin{aligned}
\nabla \bullet \overline{A} = \lim_{\Delta i \to 0} \Big\{ & \left[A_x \left(x_o + \Delta x/2 \right) - A_x \left(x_o - \Delta x/2 \right) \right] \big/ \Delta x \\
& + \left[A_y \left(y_o + \Delta y/2 \right) - A_y \left(y_o - \Delta y/2 \right) \right] \big/ \Delta y \\
& + \left[A_z \left(z_o + \Delta z/2 \right) - A_z \left(z_o - \Delta z/2 \right) \right] \big/ \Delta z \Big\}
\end{aligned}
\tag{2.4.2}
$$

$$
\begin{aligned}
= \lim_{\Delta i \to 0} \Big\{ & \Delta y \Delta z \left[A_x \left(x_o + \Delta x/2 \right) - A_x \left(x_o - \Delta x/2 \right) \right] \\
& + \Delta x \Delta z \left[A_y \left(y_o + \Delta y/2 \right) - A_y \left(y_o - \Delta y/2 \right) \right] \\
& + \Delta x \Delta y \left[A_z \left(z_o + \Delta z/2 \right) - A_z \left(z_o - \Delta z/2 \right) \right] \Big\} \big/ \Delta x \Delta y \Delta z
\end{aligned}
\tag{2.4.3}
$$

$$= \lim_{\Delta v \to 0} \left\{ \oiint_{S_c} \overline{A} \bullet \hat{n} \ da / \Delta v \right\} \tag{2.4.4}$$

where $\hat{n}$ is the unit normal vector for an incremental cube of dimensions Δx, Δy, Δz; da is its differential surface area; S_c is its surface area; and Δv is its volume, as suggested in Figure 2.4.1(a).

We may now stack an arbitrary number of such infinitesimal cubes to form a volume V such as that shown in Figure 2.4.1(b).   Then we can sum (2.4.4) over all these cubes to obtain:

$$\lim_{\Delta v \to 0} \sum_i \left(\nabla \bullet \overline{A} \right) \Delta v_i = \lim_{\Delta v \to 0} \sum_i \left\{ \oiint_{S_c} \overline{A} \bullet \hat{n} \ da_i \right\} \tag{2.4.5}$$
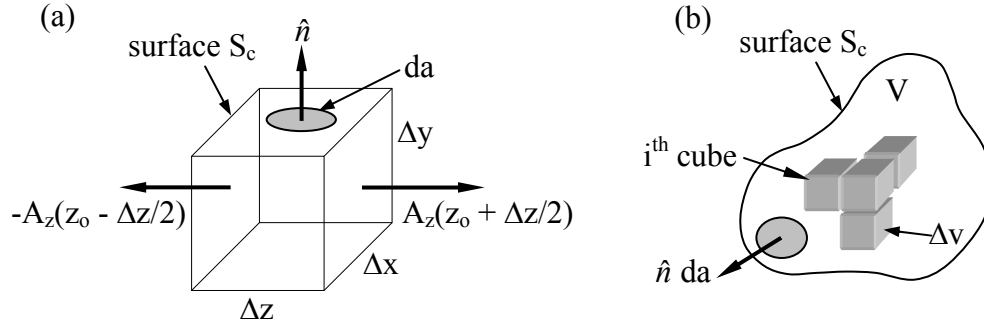
Figure 2.4.1   Derivation of Gauss's divergence theorem.

Since all contributions to $\sum_i \left\{ \oiint_S \overline{A} \bullet \hat{n} \, da_i \right\}$ from interior-facing adjacent cube faces cancel, the only remaining contributions from the right-hand side of (2.4.5) are from the outer surface of the volume V. Proceeding to the limit, we obtain *Gauss's divergence theorem*:

$$\iiint_V (\nabla \bullet \overline{A}) \, dv = \oiint_S (\overline{A} \bullet \hat{n}) \, da \qquad (2.4.6)$$

2.4.2   Stokes' theorem

Stokes' theorem states that the integral of the curl of a vector field over a bounded surface equals the line integral of that vector field along the contour C bounding that surface. Its derivation is similar to that for Gauss's divergence theorem (Section 2.4.1), starting with the definition of the z component of the curl operator [from Equation (2.1.4)]:

$$(\nabla \times \overline{A})_z \equiv \hat{z} \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) \qquad (2.4.7)$$

$$= \hat{z} \lim_{\Delta x, \Delta y \to 0} \left\{ \left[A_y \left(x_o + \Delta x/2 \right) - A_y \left(x_o - \Delta x/2 \right) \right] / \Delta x \right.$$
$$\left. - \left[A_x \left(y_o + \Delta y/2 \right) - A_x \left(y_o - \Delta y/2 \right) \right] / \Delta y \right\} \qquad (2.4.8)$$

$$= \hat{z} \lim_{\Delta x, \Delta y \to 0} \left\{ \Delta y \left[A_y \left(x_o + \Delta x/2 \right) - A_y \left(x_o - \Delta x/2 \right) \right] / \Delta x \Delta y \right.$$
$$\left. - \Delta x \left[A_x \left(y_o + \Delta y/2 \right) - A_x \left(y_o - \Delta y/2 \right) \right] / \Delta x \Delta y \right\} \qquad (2.4.9)$$

Consider a surface in the x-y plane, perpendicular to $\hat{z}$ and $\hat{n}$, the local surface normal, as illustrated in Figure 2.4.2(a).
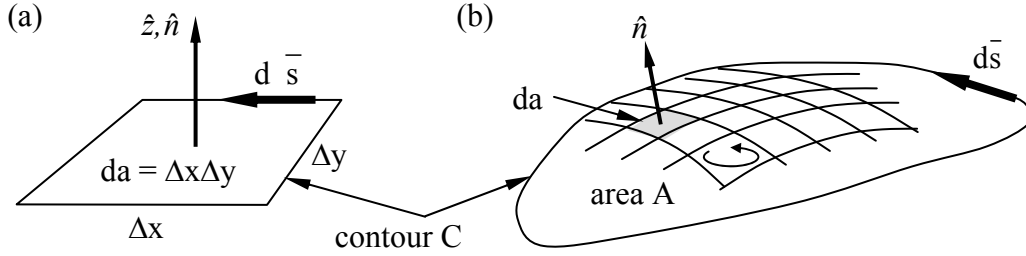
Figure 2.4.2   Derivation of Stokes' theorem.

Then (2.4.9) applied to $\Delta x \Delta y$ becomes:

$$\Delta x \Delta y (\nabla \times \overline{A}) \bullet \hat{n} = \oint_C \overline{A} \bullet d\overline{s} \tag{2.4.10}$$

where $d\overline{s}$ is a vector differential length [m] along the contour C bounding the incremental area defined by $\Delta x \Delta y = da$. The contour C is transversed in a right-hand sense relative to $\hat{n}$. We can assemble such infinitesimal areas to form surfaces of arbitrary shapes and area A, as suggested in Figure 2.4.2(b). When we sum (2.4.10) over all these infinitesimal areas da, we find that all contributions to the right-hand side interior to the area A cancel, leaving only the contributions from contour C along the border of A. Thus (2.4.10) becomes *Stokes' theorem*:

$$\iint_A (\nabla \times \overline{A}) \bullet \hat{n} \; da = \oint_C \overline{A} \bullet d\overline{s} \tag{2.4.11}$$

where the relation between the direction of integration around the loop and the orientation of $\hat{n}$ obey the right-hand rule (if the right-hand fingers curl in the direction of $d\overline{s}$, then the thumb points in the direction $\hat{n}$).

2.4.3   Maxwell's equations in integral form

The differential form of Maxwell's equations (2.1.5–8) can be converted to integral form using Gauss's divergence theorem and Stokes' theorem. Faraday's law (2.1.5) is:

$$\nabla \times \overline{E} = -\frac{\partial \overline{B}}{\partial t} \tag{2.4.12}$$

Applying Stokes' theorem (2.4.11) to the curved surface A bounded by the contour C, we obtain:

$$\iint_A (\nabla \times \overline{E}) \bullet \hat{n} \; da = \oint_C \overline{E} \bullet d\overline{s} = -\iint_A \frac{\partial \overline{B}}{\partial t} \bullet \hat{n} \; da \tag{2.4.13}$$

This becomes the integral form of Faraday's law:

$$\oint_C \overline{E} \bullet d\overline{s} = -\frac{\partial}{\partial t} \iint_A \overline{B} \bullet \hat{n} \; da \qquad\qquad (\textit{Faraday's Law}) \quad (2.4.14)$$

A similar application of Stokes' theorem to the differential form of Ampere's law yields its integral form:

$$\oint_C \overline{H} \bullet d\overline{s} = \iint_A \left[\overline{J} + \frac{\partial \overline{D}}{\partial t} \right] \bullet \hat{n} \; da \qquad\qquad (\textit{Ampere's Law}) \quad (2.4.15)$$

Gauss's divergence theorem (2.1.20) can be similarly applied to Gauss's laws to yield their integral form:

$$\iiint_V (\nabla \bullet \overline{D}) \, dv = \iiint_V \rho \; dv = \oiint_A (\overline{D} \bullet \hat{n}) \; da \qquad\qquad (2.4.16)$$

This conversion procedure thus yields the integral forms of Gauss's laws. That is, we can integrate $\overline{D} \bullet \hat{n}$ and $\overline{B} \bullet \hat{n}$ in the differential equations over the surface A that bounds the volume V:

$$\oiint_A (\overline{D} \bullet \hat{n}) \, da = \iiint_V \rho \; dv \qquad\qquad (\textit{Gauss's Law} \text{ for charge}) \quad (2.4.17)$$

$$\oiint_A (\overline{B} \bullet \hat{n}) \, da = 0 \qquad\qquad (\textit{Gauss's Law} \text{ for } \overline{B}) \quad (2.4.18)$$

Finally, conservation of charge (1.3.19) can be converted to integral form as were Gauss's laws:

$$\oiint_A (\overline{J} \bullet \hat{n}) \, da = -\iiint_V \frac{\partial \rho}{\partial t} dv \qquad\qquad (\textit{conservation of charge}) \quad (2.4.19)$$
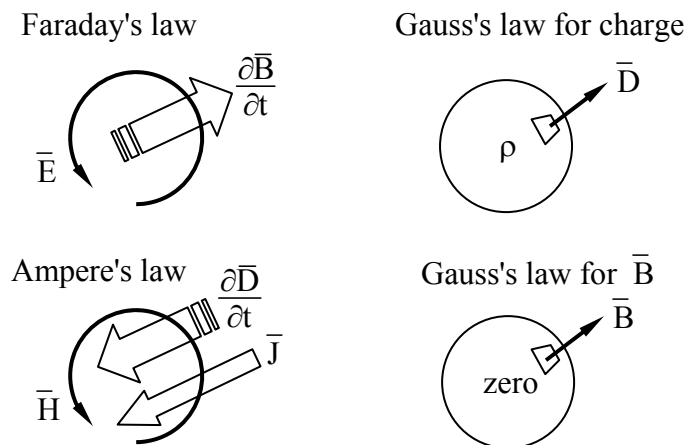
Figure 2.4.3   Maxwell's equations in sketch form.

The four sketches of Maxwell's equations presented in Figure 2.4.3 may facilitate memorization; they can be interpreted in either differential or integral form because they capture the underlying physics.

Example 2.4A

Using Gauss's law, find $\overline{E}$ at distance r from a point charge q.

Solution: The spherical symmetry of the problem requires $\overline{E}$ to be radial, and Gauss's law requires $\int_A \varepsilon_o \overline{E} \bullet \hat{r} \, dA = \int_V \rho \, dv = q = 4\pi r^2 \varepsilon_o E_r$, so $\overline{E} = \hat{r} E_r = \hat{r} q / 4\pi \varepsilon_o r^2$.

Example 2.4B

What is $\overline{H}$ at r = 1 cm from a line current $\overline{I} = \hat{z}$ [amperes] positioned at r = 0?

Solution: Because the geometry of this problem is cylindrically symmetric, so is the solution. Using the integral form of Ampere's law (2.4.15) and integrating in a right-hand sense around a circle of radius r centered on the current and in a plane orthogonal to it, we obtain $2\pi r H = I$, so $\overline{H} = \hat{\theta} \, 100/2\pi$ [A m^{-1}].

2.5 *Electric and magnetic fields in media*

2.5.1 Maxwell's equations and media

The great success of Maxwell's equations lies partly in their simple prediction of electromagnetic waves and their simple characterization of materials in terms of conductivity σ [Siemens m^{-1}], permittivity ε [Farads m^{-1}], and permeability μ [Henries m^{-1}]. In vacuum we find $\sigma = 0$, $\varepsilon = \varepsilon_o$, and $\mu = \mu_o$, where $\varepsilon_o = 8.8542 \times 10^{-12}$ and $\mu_o = 4\pi \times 10^{-7}$. For reference, Maxwell's equations are:

$$\nabla \times \overline{E} = -\frac{\partial \overline{B}}{\partial t} \tag{2.5.1}$$

$$\nabla \times \overline{H} = \overline{J} + \frac{\partial \overline{D}}{\partial t} \tag{2.5.2}$$

$$\nabla \bullet \overline{D} = \rho \tag{2.5.3}$$

$$\nabla \bullet \overline{B} = 0 \tag{2.5.4}$$

The electromagnetic properties of most media can be characterized by the *constitutive relations*:

$$\overline{D} = \varepsilon \overline{E} \tag{2.5.5}$$

$$\overline{B} = \mu\overline{H} \tag{2.5.6}$$

$$\overline{J} = \sigma\overline{E} \tag{2.5.7}$$

In contrast, the nano-structure of media can be quite complex and requires quantum mechanics for its full explanation. Fortunately, simple classical approximations to atoms and molecules suffice to understand the origins of σ, ε, and μ, as discussed below in that sequence.

2.5.2   Conductivity

Conduction in metals and *n-type semiconductors*[6] involves free electrons moving many atomic diameters before they lose momentum by interacting with atoms or other particles. Acceleration induced by the small applied electric field inside the conductor restores electron velocities to produce an equilibrium current. The total current density $\overline{J}$ [A m^{-2}] is proportional to the product of the average electron velocity $\overline{v}$ [m s^{-1}] and the number density n [m^{-3}] of free electrons. A related conduction process occurs in ionic liquids, where both negative and positive ions can carry charge long distances.

In *metals* there is approximately one free electron per atom, and in warm n-type semiconductors there is approximately one free electron per donor atom, where the sparse donor atoms are easily ionized thermally. Since, for non-obvious reasons, the average electron velocity $\langle\overline{v}\rangle$ is proportional to $\overline{E}$, therefore $\overline{J} = -en_3\langle\overline{v}\rangle = \sigma\overline{E}$, as stated in (2.5.7). As the conductivity σ approaches infinity the electric field inside a conductor approaches zero for any given current density $\overline{J}$.

Warm donor atoms in n-type semiconductors can be easily ionized and contribute electrons to the conduction band where they move freely. Only certain types of impurity atoms function as donors--those that are most easily ionized. As the density of donor atoms approaches zero and as temperature declines, the number of free electrons and the conductivity approach very low values that depend on temperature and any alternative ionization mechanisms that are present.

In *p-type semiconductors* the added impurity atoms readily trap nearby free electrons to produce a negative ion; this results in a corresponding number of positively ionized semiconductor atoms that act as "*holes*". As a result any free electrons typically move only short distances before they are trapped by one of these holes. Moreover, the threshold energy required to move an electron from a neutral atom to an adjacent positive ion is usually less than the available thermal energy, so such transfers occur rapidly, causing the hole to move quickly from

---

[6] n-type semiconductors (e.g., silicon, germanium, gallium arsenide, indium phosphide, and others) are doped with a tiny percentage of donor atoms that ionize easily at temperatures of interest, releasing mobile electrons into the conduction band of the semiconductor where they can travel freely across the material until they recombine with another ionized donor atom. The conduction band is not a place; it refers instead to an energy and wave state of electrons that enables them to move freely. The conductivity of semiconductors therefore increases with temperature; they become relatively insulating at low temperatures, depending on the ionization potentials of the impurity atoms.

place to place over long distances. Thus holes are the *dominant charge carriers* in p-type semiconductors, whereas electrons dominate in n-type semiconductors.

More broadly, *semiconductors* have a *conduction band* in which free electrons can propagate long distances; this band is separated by an energy of one or a few electron volts from the *valence band* in which electrons cannot move. The conduction band is not a location, it is a family of possible electron wave states. When electrons are excited from the valence band to the conduction band by some energetic process, they become free to move in response to electric fields. Semiconductor conductivity is approximately proportional to the number of free electrons or holes produced by the scarce impurity atoms, and therefore to the doping density of those impurity atoms. Easily ionized impurity atoms are the principal mechanism by which electrons enter the conduction band, and impurities that readily trap adjacent electrons are the principal mechanism by which holes enter and move in the valence band. Semiconductors are discussed further in Section 8.2.4. The current leakage processes in insulators vaguely resemble electron and hole conduction in semiconductors, and can include weak surface currents as well as bulk conduction; microscopic flaws can also increase conductivity. The conductivities of typical materials are listed in Table 2.5.1.

Table 2.5.1   Nominal conductivities σ of common materials [Siemens m^{-1}].

paraffin	$10^{-14} - 10^{-16}$	sea water	3-5
glass	10^{-12}	iron	10^7
dry earth	$10^{-4} - 10^{-5}$	copper	5.8×10^7
distilled water	2×10^{-4}	silver	6.14×10^7

In some exotic materials the conductivity is a function of direction and can be represented by the 3×3 matrix $\overline{\overline{\sigma}}$; such materials are not addressed here, but Section 2.5.3 addresses similar issues in the context of permittivity ε.

Some materials exhibit *superconductivity*, or infinite conductivity. In these materials pairs of electrons become loosely bound magnetically and move as a unit called a *Cooper pair*. Quantum mechanics prevents these pairs from colliding with the lattice and losing energy. Because the magnetic binding energy for these pairs involves electron spins, it is quite small. Normal conductivity returns above a threshold *critical temperature* at which the pairs are shaken apart, and it also returns above some threshold *critical magnetic field* at which the magnetic bonds coupling the electrons break as the electron spins all start to point in the same direction. Materials having critical temperatures above 77K (readily obtained in cryogenic refrigerators) are difficult to achieve. The finite number of such pairs at any temperature and magnetic field limits the current to some maximum value; moreover that current itself produces a magnetic field that can disrupt pairs. Even a few pairs can move so as to reduce electric fields to zero by short-circuiting the normal electrons until the maximum current carrying capacity of those pairs is exceeded. If the applied fields have frequency f > 0, then the Cooper pairs behave much like collisionless electrons in a plasma and therefore the applied electric field can penetrate that plasma to its skin depth, as discussed in Section 9.8. Those penetrating electric fields interact

with a small number of normal electrons to produce tiny losses in superconductors that increase with frequency.

2.5.3   Permittivity

The *permittivity* ε_o of free space is 8.854×10^{-12} farads/meter, where $\overline{D} = \varepsilon_o \overline{E}$. The permittivity ε of any material deviates from ε_o for free space if applied electric fields induce *electric dipoles* in the medium; such dipoles alter the applied electric field seen by neighboring atoms. Electric fields generally distort atoms because $\overline{E}$ pulls on positively charged nuclei ($f = q\overline{E}$ [N]) and repels the surrounding negatively charged electron clouds. The resulting small offset $\overline{d}$ of each atomic nucleus of charge $+q$ relative to the center of its associated electron cloud produces a tiny electric dipole in each atom, as suggested in Figure 2.5.1(a). In addition, most asymmetric molecules are permanently polarized, such as H_2O or NH_3, and can rotate within fluids or gases to align with an applied field. Whether the dipole moments are induced, or permanent and free to rotate, the result is a complete or partial alignment of dipole moments as suggested in Figure 2.5.1(b).

These polarization charges generally cancel inside the medium, as suggested in Figure 2.5.1(b), but the immobile atomic dipoles on the outside surfaces of the medium are not fully cancelled and therefore contribute the *surface polarization charge* ρ_{sp}.
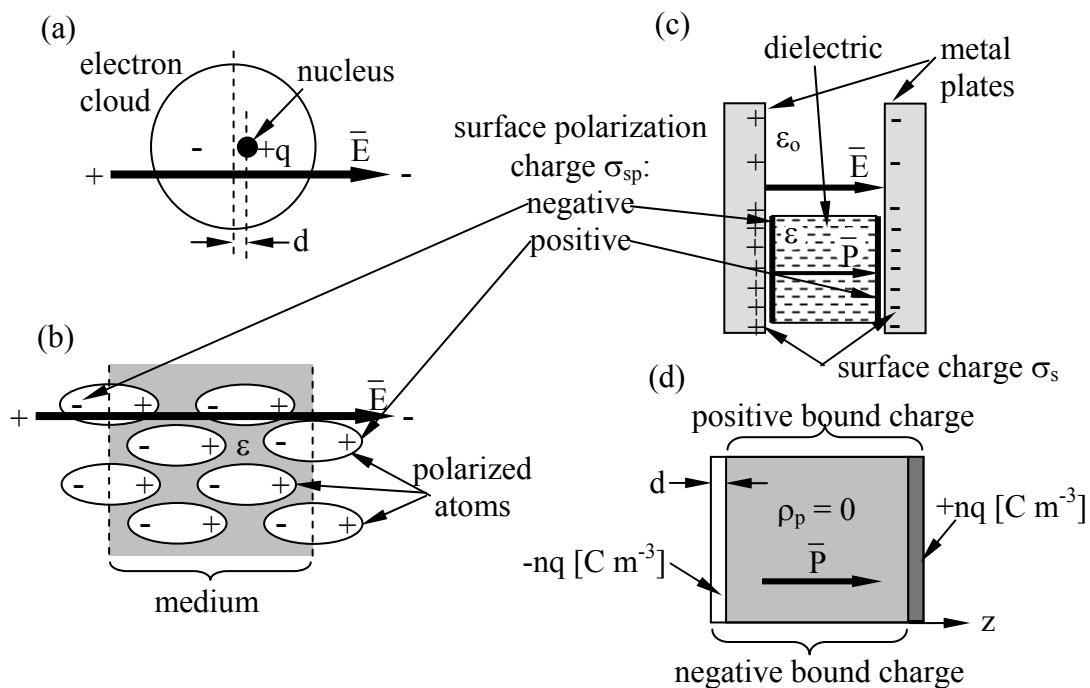
Figure 2.5.1   Polarized media.

Figure 2.5.1(c) suggests how two charged plates might provide an electric field $\overline{E}$ that polarizes a dielectric slab having permittivity $\varepsilon > \varepsilon_o$. [The electric field $\overline{E}$ is the same in vacuum

as it is inside the dielectric (assuming no air gaps) because the path integral of $\overline{E} \bullet d\overline{s}$ from plate to plate equals their voltage difference V in both cases. The electric displacement vector $\overline{D}_e = \varepsilon \overline{E}$ and therefore differs.] We associate the difference between $\overline{D}_o = \varepsilon_o \overline{E}$ (vacuum) and $\overline{D}_\varepsilon = \varepsilon \overline{E}$ (dielectric) with the *electric polarization vector* $\overline{P}$, where:

$$\overline{D} = \varepsilon \overline{E} = \varepsilon_o \overline{E} + \overline{P} = \varepsilon_o \overline{E}(1 + \chi) \tag{2.5.8}$$

The *polarization vector* $\overline{P}$ is defined by (2.5.8) and is normally parallel to $\overline{E}$ in the same direction, as shown in Figure 2.5.1(c); it points from the negative surface polarization charge to the positive surface polarization charge (unlike $\overline{E}$, which points from positive charges to negative ones). As suggested in (2.5.8), $\overline{P} = \overline{E}\varepsilon_o \chi$, where χ is defined as the dimensionless *susceptibility* of the dielectric. Because nuclei are bound rather tightly to their electron clouds, χ is generally less than 3 for most atoms, although some molecules and crystals, particularly in fluid form, can exhibit much higher values. It is shown later in (2.5.13) that $\overline{P}$ simply equals the product of the number density n of these dipoles and the average vector *electric dipole moment* of each atom or molecule, $\overline{p} = q\overline{d}$, where $\overline{d}$ is the offset (meters) of the positive charge relative to the negative charge:

$$\overline{P} = nq\overline{d} \quad \left[C\ m^{-2} \right] \tag{2.5.9}$$

Gauss's law relates $\overline{D}$ to *charge density* ρ [C m^{-2}], but we now have two types of density: that of free charges ρ_f, including ions and surface charges on conductors, and that of any locally un-neutralized polarization charge ρ_p bound within charge-neutral atoms or molecules. Gauss's law says:

$$\nabla \bullet \overline{D} = \rho_f \tag{2.5.10}$$

where ρ_f is the free charge density [C m^{-3}].

We can derive a relation similar to (2.5.10) for $\overline{P}$ by treating materials as distributed bound positive and negative charges with vacuum between them; the net bound charge density is designated the *polarization charge density* ρ_p [C m^{-3}]. Then in the vacuum between the charges $\overline{D} = \varepsilon_o \overline{E}$ and (2.5.10) becomes:

$$\varepsilon_o \nabla \bullet \overline{E} = \rho_f + \rho_p \tag{2.5.11}$$

From $\nabla \bullet$ (2.5.8), we obtain $\nabla \bullet \overline{D} = \varepsilon_o \nabla \bullet \overline{E} + \nabla \bullet \overline{P} = \rho_f$. Combining this with (2.5.11) yields:

$$\nabla \bullet \overline{P} = -\rho_p \tag{2.5.12}$$

The negative sign in (2.5.12) is consistent with the fact that $\overline{P}$, unlike $\overline{E}$, is directed from negative to positive polarization charge, as noted earlier.

Outside a polarized dielectric the polarization $\overline{P}$ is zero, as suggested by Figure 2.5.1(d). Note that the net polarization charge density is ±nq for only an atomic-scale distance d at the boundaries of the dielectric, where we model the positive and negative charge distributions within the medium as continuous uniform rectilinear slabs of density ± nq [C m^{-3}]. These two slabs are offset by the distance d. If $\overline{P}$ is in the z direction and arises from n dipole moments $\overline{p} = q\overline{d}$ per cubic meter, where $\overline{d}$ is the offset [m] of the positive charge relative to the negative charge, then (2.5.12) can be integrated over a volume V that encloses a unit area of the left-hand face of a polarized dielectric [see Figure 2.5.1(d)] to yield the polarization vector $\overline{P}$ inside the dielectric:

$$\overline{P} = \int_V \nabla \bullet \overline{P} \, dv = -\int_V \rho_p \, dv = nq\overline{d} \qquad (2.5.13)$$

The first equality of (2.5.13) involving $\overline{P}$ follows from Gauss's divergence theorem: $\int_V \nabla \bullet \overline{P} \, dv = \int_A \overline{P} \bullet \hat{z} \, da = P_z$ if A = 1. Therefore, $\overline{P} = nq\overline{d}$, proving (2.5.9).

When the electric displacement $\overline{D}$ varies with time it becomes *displacement current*, $\partial \overline{D}/\partial t$ [A/m^2], analogous to $\overline{J}$, as suggested by Ampere's law: $\nabla \times \overline{H} = \overline{J} + \partial \overline{D}/\partial t$. For reference, Table 2.5.2 presents the *dielectric constant*s $\varepsilon/\varepsilon_o$ for some common materials near 1 MHz, after Von Hippel (1954).

Table 2.5.2   Dielectric constants $\varepsilon/\varepsilon_o$ of common materials near 1 MHz.

vacuum	1.0	fused quartz	3.78
fir wood	1.8 – 2.0	ice	4.15
Teflon, petroleum	2.1	pyrex glass	5.1
vaseline	2.16	aluminum oxide	8.8
paper	2 – 3	ethyl alcohol	24.5
polystyrene	2.55	water	81.0
sandy soil	2.59	titaniumdioxide	100.0

Most dielectric materials are lossy because oscillatory electric fields dither the directions and magnitudes of the induced electric dipoles, and some of this motion heats the dielectric. This effect can be represented by a frequency-dependent complex permittivity $\underline{\varepsilon}$, as discussed further in Section 9.5. Some dielectrics have direction-dependent permittivities that can be represented by $\overline{\overline{\varepsilon}}$; such anisotropic materials are discussed in Section 9.6. Lossy anisotropic materials can be characterized by $\underline{\overline{\overline{\varepsilon}}}$.

Some special dielectric media are spontaneously polarized, even in the absence of an externally applied $\overline{E}$. This occurs for highly polarizable media where orientation of one electric dipole in the media can motivate its neighbors to orient themselves similarly, forming domains

of atoms, molecules, or crystal unit cells that are all polarized the same. Such spontaneously polarized domains are illustrated for magnetic materials in Figure 2.5.2. As in the case of ferromagnetic domains, in the absence of externally applied fields, domain size is limited by the buildup of stored field energy external to the domain; adjacent domains are oriented so as to largely cancel each other. Such ferroelectric materials have large effective values of ε, although $\overline{D}$ saturates if $\overline{E}$ is sufficient to produce ~100% alignment of polarization. They can also exhibit hysteresis as do the ferromagnetic materials discussed in Section 2.5.4.

Example 2.5A

What are the free and polarization charge densities ρ_f and ρ_p in a medium having conductivity $\sigma = \sigma_o/(1+z)$, permittivity $\varepsilon = 3\varepsilon_o$, and current density $\overline{J} = \hat{z}J_o$?

Solution: $\overline{J} = \sigma\overline{E}$, so $\overline{E} = \hat{z}J_o(1+z)/\sigma_o = \varepsilon_o\overline{E} + \overline{P}$.

From (2.5.10) $\rho_f = \nabla \bullet \overline{D} = (\partial/\partial z)\left[3\varepsilon_o J_o(1+z)/\sigma_o\right] = 3\varepsilon_o J_o/\sigma_o \; \left[\text{C m}^{-3}\right]$.

From (2.5.12) $\rho_p = -\nabla \bullet \overline{P} = -\nabla \bullet (\varepsilon - \varepsilon_o)\overline{E} = -(\partial/\partial z)2\varepsilon_o J_o(1+z)/\sigma_o$

$= 2\varepsilon_o J_o/\sigma_o \; \left[\text{C m}^{-3}\right]$.

2.5.4   Permeability

The *permeability* μ_o of free space is $4\pi10^{-7}$ Henries/meter by definition, where $\overline{B} = \mu\overline{H}$. The permeability μ of matter includes the additional contributions to $\overline{B}$ from atomic magnetic dipoles aligning with the applied $\overline{H}$. These magnetic dipoles and their associated magnetic fields arise either from electrons orbiting about atomic nuclei or from spinning charged particles, where such moving charge is current. All electrons and protons have spin $\pm1/2$ in addition to any orbital spin about the nucleus, and the net spin of an atom or nucleus can be non-zero. Their magnetic fields are linked to their equivalent currents by Ampere's law, $\nabla \times \overline{H} = \overline{J} + \partial\overline{D}/\partial t$. Quantum theory describes how these magnetic moments are quantized at discrete levels, and for some devices quantum descriptions are necessary. In this text we average over large numbers of atoms, so that $\overline{B} = \mu\overline{H}$ accurately characterizes media, and quantum effects can be ignored.

In any medium the cumulative contribution of induced magnetic dipoles to $\overline{B}$ is characterized by the *magnetization* $\overline{M}$, which is defined by:

$$\overline{B} = \mu\overline{H} = \mu_o(\overline{H} + \overline{M}) = \mu_o\overline{H}(1 + \chi_m) \tag{2.5.14}$$

where χ_m is the *magnetic susceptibility* of the medium. Because of quantum effects χ_m for *diamagnetic* materials is slightly negative so that $\mu < \mu_o$; examples include silver, copper, and water, as listed in Table 2.5.3. The table also lists representative *paramagnetic* materials, which have slightly positive magnetic susceptibilities, and ferromagnetic materials, which have very large susceptibilities (e.g., cobalt, etc.).

Table 2.5.3   Approximate relative permeabilities μ/μ_o of common materials.

bismuth	0.99983	aluminum	1.00002
silver	0.99998	cobalt	250
lead	0.999983	nickel	600
copper	0.999991	mild steel	2000
water	0.999991	iron	5000
vacuum	1.000000	mumeltal	100,000
air	1.0000004	supermalloy	1,000,000

The sharp difference between normal materials with $\mu \cong \mu_o$ and *ferromagnetic* materials having $\mu \gg \mu_o$ is due to the spontaneous alignment of atomic magnetic dipoles in the same direction so as to increase the applied field, reorienting the remaining dipoles. That is, if the susceptibility of a material is above some threshold, then the atomic magnetic dipoles spontaneously align over regions of size limited by grain structure or energy considerations, as suggested in Figure 2.5.2(a and b). These regions of nearly perfect alignment are called *magnetic domains*. These domains are normally quite small (perhaps micron-size) so as to minimize the stored magnetic energy μH^2. In this regime, if only energy considerations control domain size, then the sizes of those domains oriented in the general direction of the applied magnetic field grow as that field increases, while other domains shrink, as suggested in Figure 2.5.2(c). Since domain walls cannot easily move across grain walls, the granular structure of the material can be engineered to control magnetic properties. If domain walls move easily, the magnetic susceptability χ_m is large.
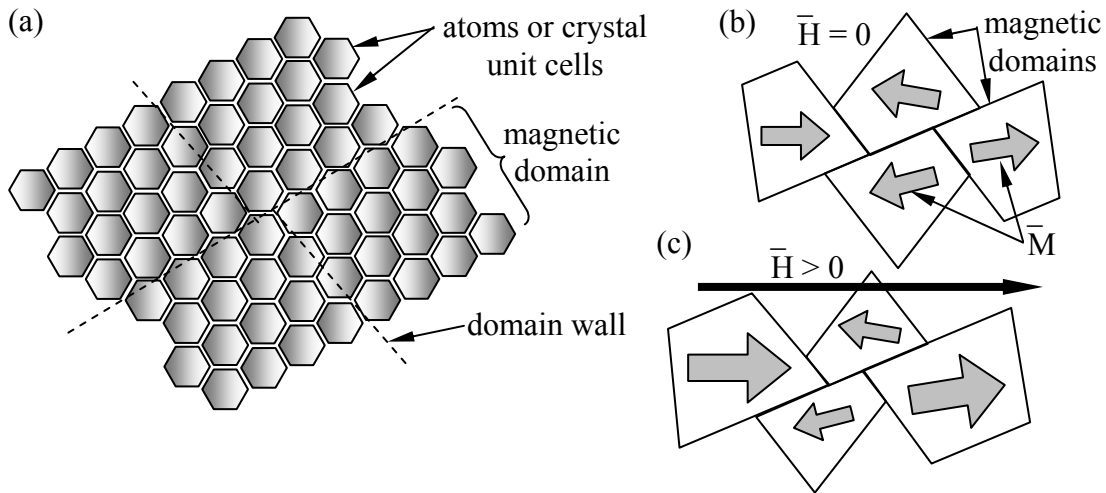
Figure 2.5.2   Magnetic domains in ferromagnetic materials.

At sufficiently high magnetic fields all domains will expand to their maximum size and/or rotate in the direction of $\overline{H}$. This corresponds to the maximum value of $\overline{M}$ and *magnetic saturation*. The resulting typical non-linear behavior of the *magnetization curve* relating B and H for ferromagnetic materials is suggested in Figure 2.5.3(a). The slope of the B vs. H curve is

~μ near the origin and ~μ_o beyond saturation. If the domains resist change and dissipate energy when doing so, the *hysteresis curve* of Figure 2.5.3(b) results. It can be shown that the area enclosed by the figure is the energy dissipated per cycle as the applied $\overline{H}$ oscillates.
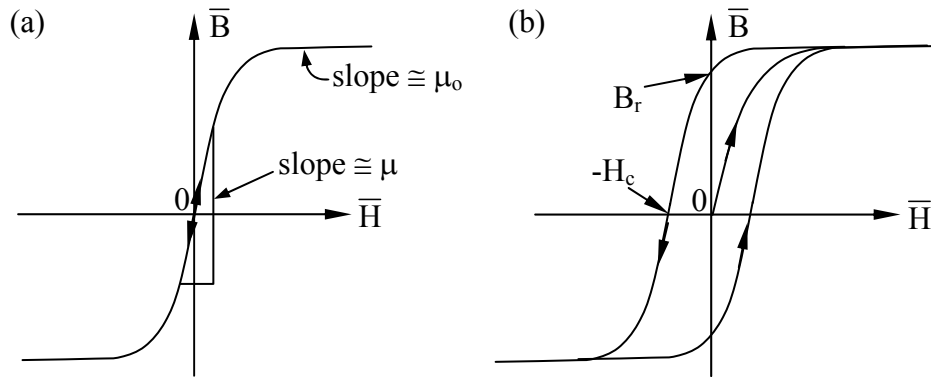
Figure 2.5.3   Magnetization curve and hysterisis loop for a ferromagnetic material.

Hard magnetic materials have large values of *residual flux density* B_r and *magnetic coercive force* or *coercivity* H_c, as illustrated. B_r corresponds to the magnetic strength B of this *permanent magnet* with no applied H. The magnetic energy density $W_m = \overline{B} \bullet \overline{H}/2 \cong 0$ inside permanent magnets because $\overline{H} = 0$, while $W_m = \mu_o H^2/2$ [J m^{-3}] outside. To *demagnetize* a permanent magnet we can apply a magnetic field H of magnitude H_c, which is the field strength necessary to drive B to zero.

If we represent the magnetic dipole moment of an atom by m, where $\overline{m}$ for a current loop of magnitude I and area $\hat{n}A$ is $\hat{n}IA$ [A m]^7, then it can be shown that the total magnetization $\overline{M}$ of a medium is $n\overline{m}$ [A m^{-1}] via the same approach used to derive $\overline{P} = n\overline{p}$ (2.5.13) for the polarization of dielectrics; n is the number of dipoles per m^3.

Example 2.5B
Show how the power dissipated in a hysteretic magnetic material is related to the area circled in Figure 2.5.3(b) as H oscillates. For simplicity, approximate the loop in the figure by a rectangle bounded by $\pm H_o$ and $\pm B_o$.

Solution: We seek the energy dissipated in the material by one traverse of this loop as H goes from $+H_o$ to $-H_o$ and back to $+H_o$. The energy density $W_m = BH/2$ when B = 0 at t = 0 is $W_m = 0$; $W_m \rightarrow B_o H_o/2$ [J m^{-3}] as B$\rightarrow B_o$. As H returns to 0 while B = B_o, this energy is dissipated and cannot be recovered by an external circuit because any voltage induced in that circuit would be $\propto \partial B/\partial t = 0$. As H$\rightarrow -H_o$, $W_m \rightarrow B_o H_o/2$; this energy can be recovered by an external circuit later as B$\rightarrow 0$ because $\partial B/\partial t \neq 0$. As B$\rightarrow -B_o$, $W_m \rightarrow B_o H_o/2$, which is lost later as H$\rightarrow 0$ with $\partial B/\partial t = 0$. The energy stored

---

^{7} $\hat{n}$ is the unit vector normal to the tiny area A enclosed by the current I, using the right-hand-rule.

as $H \rightarrow H_o$ with $B = -B_o$ is again recoverable as $B \rightarrow 0$ with $H = H_o$. Thus the minimum energy dissipated during one loop traverse is $vB_oH_o[J]$, where v is material volume. If the drive circuits do not recapture the available energy but dissipate it, the total energy dissipated per cycle is doubled.

2.6 Boundary conditions for electromagnetic fields

2.6.1 Introduction

Maxwell's equations characterize macroscopic matter by means of its permittivity ε, permeability μ, and conductivity σ, where these properties are usually represented by scalars and can vary among media. Section 2.5 discussed media for which ε, μ, and σ are represented by matrices, complex quantities, or other means. This Section 2.6 discusses how Maxwell's equations strongly constrain the behavior of electromagnetic fields at boundaries between two media having different properties, where these constraint equations are called *boundary conditions*. Section 2.6.2 discusses the boundary conditions governing field components perpendicular to the boundary, and Section 2.6.3 discusses the conditions governing the parallel field components. Section 2.6.4 then treats the special case of fields adjacent to perfect conductors.

One result of these boundary conditions is that waves at boundaries are generally partially transmitted and partially reflected with directions and amplitudes that depend on the two media and the incident angles and polarizations. Static fields also generally have different amplitudes and directions on the two sides of a boundary. Some boundaries in both static and dynamic situations also possess surface charge or carry surface currents that further affect the adjacent fields.

2.6.2 Boundary conditions for perpendicular field components

The boundary conditions governing the perpendicular components of $\overline{E}$ and $\overline{H}$ follow from the integral forms of Gauss's laws:

$$\oiint_S (\overline{D} \bullet \hat{n}) \, da = \iiint_V \rho \, dv \qquad\qquad (\textit{Gauss's Law for } \overline{D}) \qquad (2.6.1)$$

$$\oiint_S (\overline{B} \bullet \hat{n}) \, da = 0 \qquad\qquad (\textit{Gauss's Law for } \overline{B}) \qquad (2.6.2)$$

We may integrate these equations over the surface S and volume V of the thin infinitesimal pillbox illustrated in Figure 2.6.1. The pillbox is parallel to the surface and straddles it, half being on each side of the boundary. The thickness δ of the pillbox approaches zero faster than does its surface area S, where S is approximately twice the area A of the top surface of the box.
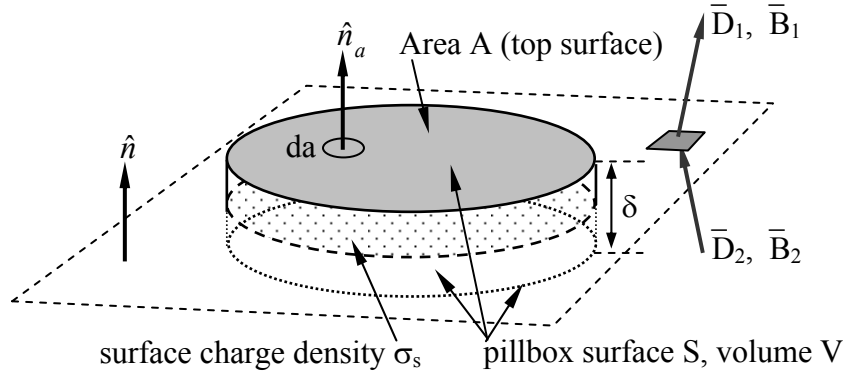
Figure 2.6.1   Elemental volume for deriving boundary conditions
for perpendicular field components.

Beginning with the boundary condition for the perpendicular component $D_\perp$, we integrate Gauss's law (2.6.1) over the pillbox to obtain:

$$\oiint_S \left(\overline{D} \bullet \hat{n}_a\right) da \cong \left(D_{1\perp} - D_{2\perp}\right) A = \iiint_V \rho \, dv = \rho_s A \qquad (2.6.3)$$

where ρ_s is the surface charge density [Coulombs m^{-2}]. The subscript s for surface charge ρ_s distinguishes it from the volume charge density ρ [C m^{-3}]. The pillbox is so thin ($\delta \to 0$) that: 1) the contribution to the surface integral of the sides of the pillbox vanishes in comparison to the rest of the integral, and 2) only a surface charge q can be contained within it, where $\rho_s = q/A =$ lim $\rho\delta$ as the charge density $\rho \to \infty$ and $\delta \to 0$. Thus (2.6.3) becomes $D_{1\perp}$ - $D_{2\perp} = \rho_s$, which can be written as:

$$\hat{n} \bullet \left(\overline{D}_1 - \overline{D}_2\right) = \rho_s \qquad \text{(boundary condition for } \overline{D}_\perp) \qquad (2.6.4)$$

where $\hat{n}$ is the unit vector normal to the boundary and points into medium 1. Thus the perpendicular component of the electric displacement vector $\overline{D}$ changes value at a boundary in accord with the surface charge density ρ_s.

Because Gauss's laws are the same for electric and magnetic fields, except that there are no magnetic charges, the same analysis for the magnetic flux density $\overline{B}$ in (2.6.2) yields a similar boundary condition:

$$\hat{n} \bullet \left(\overline{B}_1 - \overline{B}_2\right) = 0 \qquad \text{(boundary condition for } \overline{B}_\perp) \qquad (2.6.5)$$

Thus the perpendicular component of $\overline{B}$ must be continuous across any boundary.

2.6.3 Boundary conditions for parallel field components

The boundary conditions governing the parallel components of $\overline{E}$ and $\overline{H}$ follow from Faraday's and Ampere's laws:

$$\oint_C \overline{E} \bullet d\overline{s} = -\frac{\partial}{\partial t} \iint_A \overline{B} \bullet \hat{n} \, da \qquad\qquad (Faraday's\ Law) \qquad (2.6.6)$$

$$\oint_C \overline{H} \bullet d\overline{s} = \iint_A \left[\overline{J} + \frac{\partial \overline{D}}{\partial t} \right] \bullet \hat{n} \, da \qquad\qquad (Ampere's\ Law) \qquad (2.6.7)$$

We can integrate these equations around the elongated rectangular contour C that straddles the boundary and has infinitesimal area A, as illustrated in Figure 2.6.2. We assume the total height δ of the rectangle is much less than its length W, and circle C in a right-hand sense relative to the surface normal $\hat{n}_a$.
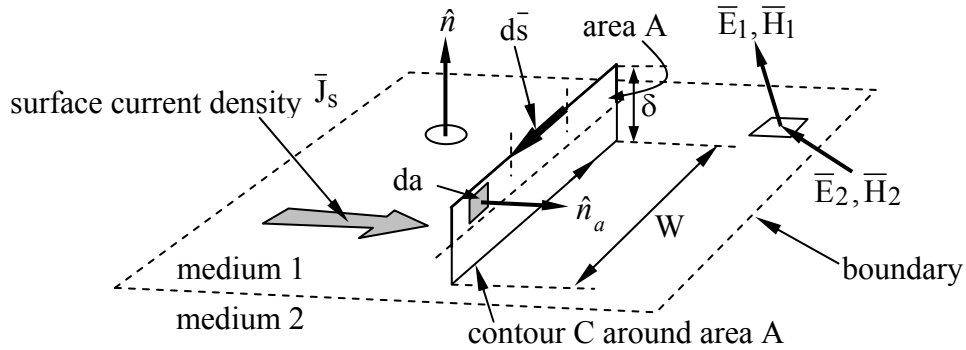
Figure 2.6.2 Elemental contour for deriving boundary conditions for parallel field components.

Beginning with Faraday's law, (2.6.6), we find:

$$\oint_C \overline{E} \bullet d\overline{s} \cong \left(\overline{E}_{1//} - \overline{E}_{2//} \right) W = -\frac{\partial}{\partial t} \iint_A \overline{B} \bullet \hat{n}_a da \rightarrow 0 \qquad (2.6.8)$$

where the integral of $\overline{B}$ over area A approaches zero in the limit where δ approaches zero too; there can be no impulses in $\overline{B}$. Since $W \neq 0$, it follows from (2.6.8) that $E_{1//} - E_{2//} = 0$, or more generally:

$$\hat{n} \times \left(\overline{E}_1 - \overline{E}_2 \right) = 0 \qquad\qquad (boundary\ condition\ for\ \overline{E}_{//}) \qquad (2.6.9)$$

Thus the parallel component of $\overline{E}$ must be continuous across any boundary.

A similar integration of Ampere's law, (2.6.7), under the assumption that the contour C is chosen to lie in a plane perpendicular to the surface current $\overline{J}_s$ and is traversed in the right-hand sense, yields:

$$\oint_C \overline{H} \bullet d\overline{s} = \left(\overline{H}_{1//} - \overline{H}_{2//}\right) W$$

$$= \iint_A \left[\overline{J} + \frac{\partial \overline{D}}{\partial t}\right] \bullet \hat{n} \; da \Rightarrow \iint_A \overline{J} \bullet \hat{n}_a \; da = \overline{J}_s W \tag{2.6.10}$$

where we note that the area integral of $\partial\overline{D}/\partial t$ approaches zero as $\delta \to 0$, but not the integral over the surface current $\overline{J}_s$, which occupies a surface layer thin compared to δ. Thus $\overline{H}_{1//} - \overline{H}_{2//} = \overline{J}_s$, or more generally:

$$\hat{n} \times \left(\overline{H}_1 - \overline{H}_2\right) = \overline{J}_s \qquad \text{(boundary condition for } \overline{H}_{//}) \tag{2.6.11}$$

where $\hat{n}$ is defined as pointing from medium 2 into medium 1. If the medium is non-conducting, $\overline{J}_s = 0$.

A simple static example illustrates how these boundary conditions generally result in fields on two sides of a boundary pointing in different directions. Consider the magnetic fields $\overline{H}_1$ and $\overline{H}_2$ illustrated in Figure 2.6.3, where $\mu_2 \neq \mu_1$, and both media are insulators so the surface current must be zero. If we are given $\overline{H}_1$, then the magnitude and angle of $\overline{H}_2$ are determined because $\overline{H}_{//}$ and $\overline{B}_\perp$ are continuous across the boundary, where $\overline{B}_i = \mu_i \overline{H}_i$. More specifically, $\overline{H}_{2//} = \overline{H}_{1//}$, and:

$$H_{2\perp} = B_{2\perp}/\mu_2 = B_{1\perp}/\mu_2 = \mu_1 H_{1\perp}/\mu_2 \tag{2.6.12}$$
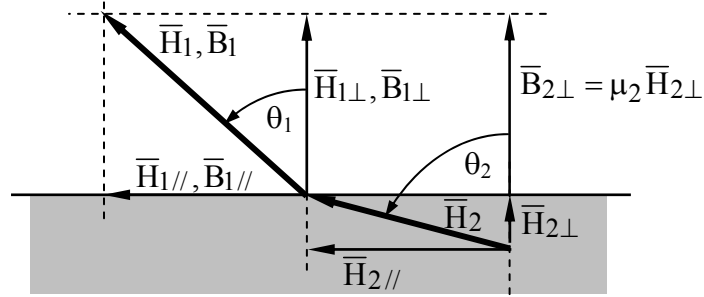
Figure 2.6.3   Static magnetic field directions at a boundary.

It follows that:

$$\theta_2 = \tan^{-1}\left(\left|\overline{H}_{2//}\right|/H_{2\perp}\right) = \tan^{-1}\left(\mu_2\left|\overline{H}_{1//}\right|/\mu_1 H_{1\perp}\right) = \tan^{-1}\left[(\mu_2/\mu_1)\tan\theta_1\right] \tag{2.6.13}$$

Thus θ_2 approaches 90 degrees when $\mu_2 \gg \mu_1$, almost regardless of θ_1, so the magnetic flux inside high permeability materials is nearly parallel to the walls and trapped inside, even when the field orientation outside the medium is nearly perpendicular to the interface. The flux escapes high-μ material best when $\theta_1 \cong 90°$. This phenomenon is critical to the design of motors or other systems incorporating iron or nickel.

If a static surface current $\overline{J}_s$ flows at the boundary, then the relations between $\overline{B}_1$ and $\overline{B}_2$ are altered along with those for $\overline{H}_1$ and $\overline{H}_2$. Similar considerations and methods apply to static electric fields at a boundary, where any static surface charge on the boundary alters the relationship between $\overline{D}_1$ and $\overline{D}_2$. Surface currents normally arise only in non-static or "dynamic" cases.

Example 2.6A

Two insulating planar dielectric slabs having ε_1 and ε_2 are bonded together. Slab 1 has $\overline{E}_1$ at angle θ_1 to the surface normal. What are $\overline{E}_2$ and θ_2 if we assume the surface charge at the boundary $\rho_s = 0$? What are the components of $\overline{E}_2$ if $\rho_s \neq 0$?

Solution: $\overline{E}_{//}$ is continuous across any boundary, and if $\rho_s = 0$, then $\overline{D}_\perp = \varepsilon_i \overline{E}_\perp$ is continuous too, which implies $\overline{E}_{2\perp} = (\varepsilon_1/\varepsilon_2)\overline{E}_{1\perp}$. Also, $\theta_1 = \tan^{-1}(E_{//}/E_{1\perp})$, and $\theta_2 = \tan^{-1}(E_{//}/E_{2\perp})$. It follows that $\theta_2 = \tan^{-1}[(\varepsilon_2/\varepsilon_1)\tan\theta_1]$. If $\rho_s \neq 0$ then $\overline{E}_{//}$ is unaffected and $\overline{D}_{2\perp} = \overline{D}_{1\perp} + \hat{n}\rho_s$ so that $\overline{E}_{2\perp} = \overline{D}_{2\perp}/\varepsilon_2 = (\varepsilon_1/\varepsilon_2)\overline{E}_{1\perp} + \hat{n}\rho_s/\varepsilon_2$.

2.6.4   Boundary conditions adjacent to perfect conductors

The four boundary conditions (2.6.4), (2.6.5), (2.6.9), and (2.6.11) are simplified when one medium is a perfect conductor ($\sigma = \infty$) because electric and magnetic fields must be zero inside it. The electric field is zero because otherwise it would produce enormous $\overline{J} = \sigma\overline{E}$ so as to redistribute the charges and to neutralize that $\overline{E}$ almost instantly, with a time constant $\tau = \varepsilon/\sigma$ seconds, as shown in Equation (4.3.3).

It can also be easily shown that $\overline{B}$ is zero inside perfect conductors. Faraday's law says $\nabla \times \overline{E} = -\partial\overline{B}/\partial t$, so if $\overline{E} = 0$, then $\partial\overline{B}/\partial t = 0$. If the perfect conductor were created in the absence of $\overline{B}$ then $\overline{B}$ would always remain zero inside. It has further been observed that when certain special materials become superconducting at low temperatures, as discussed in Section 2.5.2, any pre-existing $\overline{B}$ is thrust outside.

The boundary conditions for perfect conductors are also relevant for normal conductors because most metals have sufficient conductivity σ to enable $\overline{J}$ and ρ_s to cancel the incident electric field, although not instantly. As discussed in Section 4.3.1, this relaxation process by which charges move to cancel $\overline{E}$ is sufficiently fast for most metallic conductors that they

largely obey the perfect-conductor boundary conditions for most wavelengths of interest, from DC to beyond the infrared region. This relaxation time constant is $\tau = \varepsilon/\sigma$ seconds. One consequence of finite conductivity is that any surface current penetrates metals to some depth $\delta = \sqrt{2/\omega\mu\sigma}$, called the skin depth, as discussed in Section 9.2. At sufficiently low frequencies, even sea water with its limited conductivity largely obeys the perfect-conductor boundary condition.

The four boundary conditions for fields adjacent to perfect conductors are presented below together with the more general boundary condition from which they follow when all fields in medium 2 are zero:

$$\hat{n} \bullet \overline{B} = 0 \qquad\qquad \left[\text{from } \hat{n} \bullet (\overline{B}_1 - \overline{B}_2) = 0 \right] \qquad (2.6.14)$$

$$\hat{n} \bullet \overline{D} = \rho_s \qquad\qquad \left[\text{from } \hat{n} \bullet (\overline{D}_1 - \overline{D}_2) = \rho_s \right] \qquad (2.6.15)$$

$$\hat{n} \times \overline{E} = 0 \qquad\qquad \left[\text{from } \hat{n} \times (\overline{E}_1 - \overline{E}_2) = 0 \right] \qquad (2.6.16)$$

$$\hat{n} \times \overline{H} = \overline{J}_s \qquad\qquad \left[\text{from } \hat{n} \times (\overline{H}_1 - \overline{H}_2) = \overline{J}_s \right] \qquad (2.6.17)$$

These four boundary conditions state that magnetic fields can only be parallel to perfect conductors, while electric fields can only be perpendicular. Moreover, the magnetic fields are always associated with surface currents flowing in an orthogonal direction; these currents have a numerical value equal to $\overline{H}$. The perpendicular electric fields are always associated with a surface charge ρ_s numerically equal to $\overline{D}$; the sign of σ is positive when $\overline{D}$ points away from the conductor.

Example 2.6B
What boundary conditions apply when $\mu \rightarrow \infty$, $\sigma = 0$, and $\varepsilon = \varepsilon_o$?

Solution: Inside this medium $\overline{H} = 0$ and $\overline{J} = 0$ because otherwise infinite energy densities, $\mu|H|^2/2$, are required; static $\overline{E}$ and $\overline{B}$ are unconstrained, however. Since $\nabla \times \overline{H} = 0 = \overline{J} + \partial\overline{D}/\partial t$ inside, dynamic $\overline{E}$ and $\overline{D} = 0$ there too. Since $\overline{H}_{//}$ and $\overline{B}_{\perp}$ are continuous across the boundary, $\overline{H}_{//} = 0$ and $\overline{H}_{\perp}$ can be anything at the boundary. Since $\overline{E}_{//}$ and $\overline{D}_{\perp}$ are continuous (let's assume $\rho_s = 0$ if $\overline{J} = 0$), static $\overline{E}$ and $\overline{D}$ are unconstrained at the boundary while dynamic $\overline{E} = \overline{D} = 0$ there because there is no dynamic electric field inside and no dynamic surface charge. Since only $\overline{H}_{\perp} \neq 0$ at the boundary, this is non-physical and such media don't exist. For example, there is no way to match boundary conditions for an incoming plane wave. This impasse would be avoided if $\sigma \neq 0$, for then dynamic $\overline{H}_{//}$ and $\overline{E}_{\perp}$ could be non-zero.

2.7   Power and energy in the time and frequency domains, Poynting theorem

2.7.1   Poynting theorem and definition of power and energy in the time domain

To derive the Poynting theorem we can manipulate Maxwell's equations to produce products of variables that have the dimensions and character of power or energy. For example, the power dissipated in a resistor is the product of its voltage and current, so the product $\overline{E} \bullet \overline{J}$ [W m^{-3}] would be of interest. The dimensions of $\overline{E}$ and $\overline{J}$ are volts per meter and amperes per square meter, respectively. Faraday's and Ampere's laws are:

$$\nabla \times \overline{E} = -\frac{\partial \overline{B}}{\partial t} \qquad\qquad\qquad \text{(Faraday's law)} \qquad (2.7.1)$$

$$\nabla \times \overline{H} = \overline{J} + \frac{\partial \overline{D}}{\partial t} \qquad\qquad\qquad \text{(Ampere's law)} \qquad (2.7.2)$$

We can produce the product $\overline{E} \bullet \overline{J}$ and preserve symmetry in the resulting equation by taking the dot product of $\overline{E}$ with Ampere's law and subtracting from it the dot product of $\overline{H}$ with Faraday's law, yielding:

$$\overline{E} \bullet (\nabla \times \overline{H}) - \overline{H} \bullet (\nabla \times \overline{E}) = \overline{E} \bullet \overline{J} + \overline{E} \bullet \frac{\partial \overline{D}}{\partial t} + \overline{H} \bullet \frac{\partial \overline{B}}{\partial t} \qquad (2.7.3)$$

$$= -\nabla \bullet (\overline{E} \times \overline{H}) \qquad (2.7.4)$$

where (2.7.4) is a vector identity. Equations (2.7.3) and (2.7.4) can be combined to form the *Poynting theorem*:

$$\nabla \bullet (\overline{E} \times \overline{H}) + \overline{E} \bullet \overline{J} + \overline{E} \bullet \frac{\partial \overline{D}}{\partial t} + \overline{H} \bullet \frac{\partial \overline{B}}{\partial t} = 0 \ \left[Wm^{-3} \right] \qquad (2.7.5)$$

The dimension of $\overline{E} \bullet \overline{J}$ and every other term in this equation is W m^{-3}. If $\overline{D} = \varepsilon \overline{E}$ and $\overline{B} = \mu \overline{H}$, then $\overline{E} \bullet \partial \overline{D}/\partial t = \partial \left[\varepsilon |\overline{E}|^2 /2 \right] / \partial t$ and $\overline{H} \bullet \partial \overline{B}/\partial t = \partial \left[\mu |\overline{H}|^2 /2 \right] / \partial t$. The factor of one-half arises because we are now differentiating with respect to a squared quantity rather than a single quantity, as in (2.7.5). It follows that $\varepsilon |\overline{E}|^2 /2$ and $\mu |\overline{H}|^2 /2$ have the dimension of J m^{-3} and represent electric and magnetic energy density, respectively, denoted by W_e and W_m. The product $\overline{E} \bullet \overline{J}$ can represent either power dissipation or a power source, both denoted by P_d. If $\overline{J} = \sigma \overline{E}$, then $P_d = \sigma |\overline{E}|^2$ [W m^{-3}], where σ is the conductivity of the medium, as discussed later in Section 3.1.2. To summarize:

$$P_d = \overline{J} \bullet \overline{E} \ \left[W \, m^{-3} \right] \qquad\qquad (\textit{power dissipation density}) \qquad (2.7.6)$$

$$W_e = \frac{1}{2}\varepsilon|\overline{E}|^2 \quad \left[\text{J m}^{-3}\right] \qquad\qquad (\textit{electric energy density}) \qquad (2.7.7)$$

$$W_m = \frac{1}{2}\mu|\overline{H}|^2 \quad \left[\text{J m}^{-3}\right] \qquad\qquad (\textit{magnetic energy density}) \qquad (2.7.8)$$

Thus we can write Poynting's theorem in a simpler form:

$$\nabla \bullet (\overline{E} \times \overline{H}) + \overline{E} \bullet \overline{J} + \frac{\partial}{\partial t}\left(W_e + W_m\right) = 0 \quad \left[\text{Wm}^{-3}\right] \qquad (\text{Poynting theorem}) \qquad (2.7.9)$$

suggesting that the sum of the divergence of electromagnetic power associated with $\overline{E} \times \overline{H}$, the density of power dissipated, and the rate of increase of energy storage density must equal zero.

The physical interpretation of $\nabla \bullet (\overline{E} \times \overline{H})$ is best seen by applying Gauss's divergence theorem to yield the integral form of the Poynting theorem:

$$\oiint_A (\overline{E} \times \overline{H}) \bullet \hat{n}\, da + \iiint_V \overline{E} \bullet \overline{J}\, dv + (\partial/\partial t)\iiint_V \frac{1}{2}(\overline{E} \bullet \overline{D} + \overline{H} \bullet \overline{B})\, dv = 0 \text{ [W]} \qquad (2.7.10)$$

which can also be represented as:

$$\oiint_A (\overline{E} \times \overline{H}) \bullet \hat{n}\, da + p_d + \frac{\partial}{\partial t}\left(w_e + w_m\right) = 0 \text{ [W]} \qquad (\text{Poynting theorem}) \qquad (2.7.11)$$

Based on (2.7.8) and conservation of power (1.1.6) it is natural to associate $\overline{E} \times \overline{H}$ [W m^{-2}] with the instantaneous power density of an electromagnetic wave characterized by $\overline{E}$ [V m^{-1}] and $\overline{H}$ [A m^{-1}]. This product is defined as the *Poynting vector*:

$$\overline{S} \equiv \overline{E} \times \overline{H} \quad \left[\text{Wm}^{-2}\right] \qquad\qquad (\text{Poynting vector}) \qquad (2.7.12)$$

The instantaneous *electromagnetic wave intensity* of a uniform plane wave. Thus Poynting's theorem says that the integral of the inward component of the Poynting vector over the surface of any volume V equals the sum of the power dissipated and the rate of energy storage increase inside that volume.

Example 2.7A

Find $\overline{S}(t)$ and $\langle \overline{S}(t) \rangle$ for a uniform plane wave having $\overline{E} = \hat{x}E_o \cos(\omega t - kz)$. Find the electric and magnetic energy densities $W_e(t,z)$ and $W_m(t,z)$ for the same wave; how are they related?

Solution: $\overline{H} = \hat{z} \times \overline{E}/\eta_o = \hat{y}E_o \cos(\omega t - kz)/\eta_o$ where $\eta_o = (\mu_o/\varepsilon_o)^{0.5}$. $\overline{S} = \overline{E} \times \overline{H} = \hat{z}\, E_o^2 \cos^2(\omega t - kz)/\eta_o$ and $\langle \overline{S}(t) \rangle = \hat{z}\, E_o^2/2\eta_o$ [Wm^{-2}]. $W_e(t,z) = \varepsilon_o E_o^2 \cos^2(\omega t-kz)/2$ [Jm^{-3}] and $W_m(t,z) = \mu_o E_o^2 \cos^2(\omega t - kz)/2\eta_o^2 = \varepsilon_o E_o^2 \cos^2(\omega t - kz)/2 = W_e(t,z)$; the electric and

magnetic energy densities vary together in space and time and are equal for a single uniform plane wave.

2.7.2   Complex Poynting theorem and definition of complex power and energy

Unfortunately we cannot blindly apply to power and energy our standard conversion protocol between frequency-domain and time-domain representations because we no longer have only a single frequency present. Time-harmonic power and energy involve the products of sinusoids and therefore exhibit sum and difference frequencies. More explicitly, we cannot simply represent the Poynting vector $\overline{S}(t)$ for a field at frequency f [Hz] by $\mathrm{Re}\{\underline{\overline{S}}\,e^{j\omega t}\}$ because power has components at both f = 0 and 2f Hz, where $\omega = 2\pi f$. Nonetheless we can use the convenience of the time-harmonic notation by restricting it to fields, voltages, and currents while representing their products, i.e. power and energy, only by their complex averages.

To understand the definitions of complex power and energy, consider the product of two sinusoids, a(t) and b(t), where:

$$a(t) = R_e\left\{\underline{A}e^{j\omega t}\right\} = R_e\left\{Ae^{j\alpha}e^{j\omega t}\right\} = A\cos(\omega t + \alpha), \quad b(t) = B\cos(\omega t + \beta) \qquad (2.7.13)$$

$$a(t)b(t) = AB\cos(\omega t + \alpha)\cos(\omega t + \beta) = (AB/2)\left[\cos(\alpha - \beta) + \cos(2\omega t + \alpha + \beta)\right] \quad (2.7.14)$$

where we used the identity $\cos\gamma\cos\theta = \left[\cos(\gamma - \theta) + \cos(\gamma + \theta)\right]/2$. If we time average (2.7.14) over a full cycle, represented by the operator $\langle\bullet\rangle$, then the last term becomes zero, leaving:

$$\langle a(t)b(t)\rangle = \frac{1}{2}AB\cos(\alpha - \beta) = \frac{1}{2}R_e\left\{Ae^{j\alpha}Be^{-j\beta}\right\} = \frac{1}{2}R_e\left\{\underline{A}\,\underline{B}^*\right\} \qquad (2.7.15)$$

By treating each of the x, y, and z components separately, we can readily show that (2.7.15) can be extended to vectors:

$$\left\langle \overline{A}(t)\bullet\overline{B}(t)\right\rangle = \frac{1}{2}R_e\left\{\underline{\overline{A}}\bullet\underline{\overline{B}}^*\right\} \qquad (2.7.16)$$

The time average of the Poynting vector $\overline{S}(t) = \overline{E}(t)\times\overline{H}(t)$ is:

$$\left\langle \overline{S}(t)\right\rangle = \frac{1}{2}R_e\left\{\underline{\overline{E}}\times\underline{\overline{H}}^*\right\} = \frac{1}{2}\mathrm{Re}\{\underline{\overline{S}}\}\ \left[W/m^2\right] \quad \text{(Poynting average power density)} \qquad (2.7.17)$$

where we define the complex Poynting vector as:

$$\underline{\overline{S}} = \underline{\overline{E}}\times\underline{\overline{H}}^*\ \left[W\,m^{-2}\right] \qquad\qquad (\textit{complex Poynting vector}) \qquad (2.7.18)$$

Note that $\underline{\overline{S}}$ is complex and can be purely imaginary. Its average power density is given by (2.7.17).

We can re-derive Poynting's theorem to infer the physical significance of this complex vector, starting from the complex Maxwell equations:

$$\nabla \times \underline{\overline{E}} = -j\omega\underline{\overline{B}} \qquad \text{(Faraday's law)} \qquad (2.7.19)$$

$$\nabla \times \underline{\overline{H}} = \underline{\overline{J}} + j\omega\underline{\overline{D}} \qquad \text{(Ampere's law)} \qquad (2.7.20)$$

To see how time-average dissipated power, $\underline{\overline{E}} \bullet \underline{\overline{J}}^* / 2$ is related to other terms, we compute the dot product of $\underline{\overline{E}}$ and the complex conjugate of Ampere's law, and subtract it from the dot product of $\underline{\overline{H}}^*$ and Faraday's law to yield:

$$\underline{\overline{H}}^* \bullet \left(\nabla \times \underline{\overline{E}}\right) - \underline{\overline{E}} \bullet \left(\nabla \times \underline{\overline{H}}^*\right) = -j\omega\underline{\overline{H}}^* \bullet \underline{\overline{B}} - \underline{\overline{E}} \bullet \underline{\overline{J}}^* + j\omega\underline{\overline{E}} \bullet \underline{\overline{D}}^* \qquad (2.7.21)$$

Using the vector identity in (2.7.4), we obtain from (2.7.21) the differential form of the complex Poynting theorem:

$$\nabla \bullet \left(\underline{\overline{E}} \times \underline{\overline{H}}^*\right) = -\underline{\overline{E}} \bullet \underline{\overline{J}}^* - j\omega\left(\underline{\overline{H}}^* \bullet \underline{\overline{B}} - \underline{\overline{E}} \bullet \underline{\overline{D}}^*\right) \qquad \text{(\textit{complex Poynting theorem})} \qquad (2.7.22)$$

The integral form of the complex Poynting theorem follows from the complex differential form as it did in the time domain, by using Gauss's divergence theorem. The integral form of (2.7.22), analogous to (2.7.10), therefore is:

$$\oiint_A \left(\underline{\overline{E}} \times \underline{\overline{H}}^*\right) \bullet \hat{n} \, da + \iiint_V \left\{\underline{\overline{E}} \bullet \underline{\overline{J}}^* + j\omega\left(\underline{\overline{H}}^* \bullet \underline{\overline{B}} - \underline{\overline{E}} \bullet \underline{\overline{D}}^*\right)\right\} dv = 0 \qquad (2.7.23)$$

We can interpret the complex Poynting theorem in terms of physical quantities using (2.7.17) and by expressing the integral form of the complex Poynting theorem as:

$$\frac{1}{2}\oiint_A \underline{\overline{S}} \bullet \hat{n} \, da + \frac{1}{2}\iiint_V \left[\underline{\overline{E}} \bullet \underline{\overline{J}}^* + 2j\omega\left(\underline{W}_m - \underline{W}_e\right)\right] dv = 0 \qquad (2.7.24)$$

where the complex energy densities and time-average power density dissipated P_d are:

$$\underline{W}_m = \frac{1}{2}\underline{\overline{H}}^* \bullet \underline{\overline{B}} = \frac{1}{2}\underline{\mu}\left|\underline{\overline{H}}\right|^2 \quad \left[\text{J/m}^3\right] \qquad (2.7.25)$$

$$\underline{W}_e = \frac{1}{2}\underline{\overline{E}} \bullet \underline{\overline{D}}^* = \frac{1}{2}\underline{\varepsilon}\left|\underline{\overline{E}}\right|^2 \quad \left[\text{J/m}^3\right] \qquad (2.7.26)$$

$$P_d = \frac{1}{2}\sigma|\underline{\overline{E}}|^2 + \iiint_V 2\omega I_m \left[\underline{W}_e - \underline{W}_m\right] dv \equiv \left[W/m^3\right]$$ (2.7.27)

We recall that the instantaneous *magnetic energy density* $W_m(t)$ is $\mu|\overline{H}|^2\big/2$ [J/m³] from (2.7.8), and that its time average is $\mu|\overline{H}|^2\big/4$ because its peak density is $\mu|\overline{H}|^2\big/2$ and it varies sinusoidally at 2f Hz. If $\overline{B} = \underline{\mu}\overline{H} = (\mu_r + j\mu_i)\overline{H}$ and $\overline{D} = \underline{\varepsilon}\overline{E} = (\varepsilon_r + j\varepsilon_i)\overline{E}$, then (2.7.25)-(2.5.27) become:

$$\langle W_m(t)\rangle = \frac{1}{2}R_e\{\underline{W}_m\} = \frac{1}{4}\mu_r|\overline{H}|^2 \left[J/m^3\right]$$ (magnetic energy density)   (2.7.28)

$$\langle W_e(t)\rangle = \frac{1}{2}R_e\{\underline{W}_e\} = \frac{1}{4}\varepsilon_r|\overline{E}|^2 \left[J/m^3\right]$$ (electric energy density)   (2.7.29)

$$P_d = \frac{1}{2}\sigma|\underline{\overline{E}}|^2 + \iiint_V 2\omega I_m \left[\underline{W}_e - \underline{W}_m\right] dv \equiv \left[W/m^3\right]$$ (power dissipation density) (2.7.30)

We can now interpret the physical significance of the complex Poynting vector $\overline{\underline{S}}$ by restating the real part of (2.7.24) as the time-average quantity:

$$p_r + p_d = 0 \quad [W]$$ (2.7.31)

where the time-average total *power radiated* outward across the surface area A is:

$$p_r = \frac{1}{2}R_e \oiint_A \{\overline{\underline{S}} \bullet \hat{n}\}\, da \quad [W]$$ (2.7.32)

as also given by (2.7.17), and p_d is the time-average power dissipated [W] within the volume V, as given by (2.7.27). The *flux density* or time-average radiated power intensity [W/m²] is therefore $P_r = 0.5 R_e[\overline{\underline{S}}]$. Note that the dissipated power p_d can be negative if there is an external or internal source (e.g., a battery) supplying power to the volume; it is represented by negative contribution to $\overline{\underline{E}} \bullet \overline{\underline{J}}^*$. The imaginary part of the radiation (2.7.24) becomes:

$$\oiint_A I_m\left[\overline{\underline{S}} \bullet \hat{n}\right] da + \iiint_V \left(I_m\left\{\overline{\underline{E}} \bullet \overline{\underline{J}}^*\right\} + 4\omega\left[\langle w_m(t)\rangle - \langle w_e(t)\rangle\right]\right) dv = 0 \quad [W]$$ (2.7.33)

The surface integral over A of the imaginary part of the Poynting vector is the reactive power, which is simply related by (2.7.33) to the difference between the average magnetic and electric energies stored in the volume V and to any reactance associated with $\overline{J}$.

2.7.3   Power and energy in uniform plane waves

Consider the +z-propagating uniform time-harmonic plane wave of (2.3.1–2), where:

$$\overline{E}(\overline{r},t) = \hat{x} E_o \cos(z - ct) \quad [V/m] \tag{2.7.34}$$

$$\overline{H}(\overline{r},t) = \hat{y} \sqrt{\frac{\varepsilon_o}{\mu_o}} E_o \cos(z - ct) \quad \left[A/m^2\right] \tag{2.7.35}$$

The flux density for this wave is given by the Poynting vector $\overline{S}(t)$ (2.7.12):

$$\overline{S}(t) = \overline{E} \times \overline{H} = \hat{z} \frac{E_o^2}{\eta_o} \cos^2(z - ct) \quad \left[W/m^2\right] \tag{2.7.36}$$

where the characteristic impedance of free space is $\eta_o = \sqrt{\mu_o/\varepsilon_o}$ ohms (2.2.19). The time average of $\overline{S}(t)$ in (2.7.36) is $\hat{z} E_o^2 / 2\eta_o$ $[W/m^2]$.

The electric and magnetic energy densities for this wave can be found from (2.7.7–8):

$$W_e = \frac{1}{2}\varepsilon |\overline{E}|^2 = \frac{1}{2}\varepsilon_o E_o^2 \cos^2(z - ct) \quad \left[J/m^3\right] \qquad \text{(electric energy density)} \tag{2.7.37}$$

$$W_m = \frac{1}{2}\mu |\overline{H}|^2 = \frac{1}{2}\varepsilon_o E_o^2 \cos^2(z - ct) \quad \left[J/m^3\right] \qquad \text{(magnetic energy density)} \tag{2.7.38}$$

Note that these two energy densities, W_e and W_m, are equal, non-negative, and sinusoidal in behavior at twice the spatial frequency $k = 2\pi/\lambda$ [radians/m] of the underlying wave, where $\cos^2(z - ct) = 0.5[1 + \cos 2(z - ct)]$, as illustrated in Figure 2.7.1; their frequency f[Hz] is also double that of the underlying wave. They have the same space/time form as the flux density, except with a different magnitude.

The complex electric and magnetic fields corresponding to (2.7.34–5) are:

$$\underline{\overline{E}}(\overline{r}) = \hat{x}\underline{E}_o \quad [V/m] \tag{2.7.39}$$

$$\underline{\overline{H}}(\overline{r}) = \hat{y}\sqrt{\varepsilon_o/\mu_o}\,\underline{E}_o \quad \left[A/m^2\right] \tag{2.7.40}$$

The real and imaginary parts of the complex power for this uniform plane wave indicate the time average and reactive powers, respectively:

$$\langle \overline{S}(t) \rangle = \frac{1}{2}R_e[\underline{\overline{S}}] = \frac{1}{2}R_e\left[\underline{\overline{E}} \times \underline{\overline{H}}^*\right] = \hat{z}\frac{1}{2\eta_o}|\underline{E}_o|^2 = \hat{z}\frac{1}{2}\eta_o|\underline{H}_o|^2 \quad \left[W/m^2\right] \tag{2.7.41}$$

$$I_m \left\{ \underline{S}(\bar{r}) \right\} = I_m \left\{ \overline{\underline{E}} \times \overline{\underline{H}}^* \right\} = 0 \qquad (2.7.42)$$

If two superimposed plane waves are propagating in opposite directions with the same polarization, then the imaginary part of the Poynting vector is usually non-zero. Positive reactive power flowing into a volume is generally associated with an excess of time-average magnetic energy storage over electric energy storage in that volume, and vice-versa, with negative reactive power input corresponding to excess electric energy storage.

$$\overline{E}(\bar{r}, t) = \hat{x} E_o \cos(z - ct)$$

$$\overline{S}(\bar{r}, t) = \hat{z} \left(E_o^2 / \eta_o \right) \cos^2(z - ct)$$
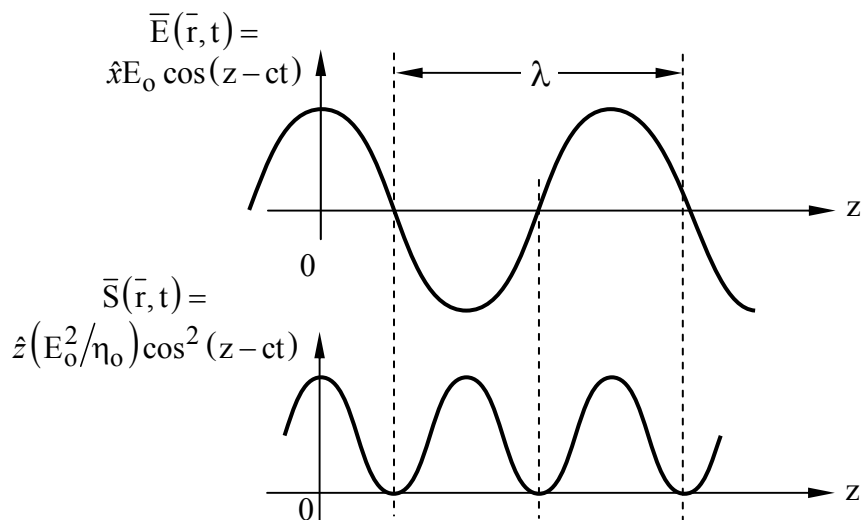
Figure 2.71   Electric field, electric and magnetic storage, and wave intensity for a uniform plane wave.

Example 2.7B

Two equal x-polarized plane waves are propagating along the z axis in opposite directions such that $\overline{E}(t) = 0$ at $z = 0$ for all time t. What is $\underline{S}(z)$?

Solution: $\overline{\underline{E}}(z) = \hat{x} E_o \left(e^{-jkz} - e^{+jkz} \right) = 0$ at z = 0; and $\overline{\underline{H}}(z) = \hat{y} \left(E_o / \eta_o \right) \left(e^{-jkz} + e^{+jkz} \right)$, so

$\underline{S} = \overline{\underline{E}} \times \overline{\underline{H}}^* = \hat{z} \left(E_o^2 / \eta_o \right) \left(e^{-j2kz} - e^{+j2kz} \right) = -\hat{z} 2j \left(E_o^2 / \eta_o \right) \sin(2kz)$. $\underline{S}$ is pure imaginary and varies sinusoidally along z between inductive and capacitive reactive power, with nulls at intervals of $\lambda/2$.

2.8   *Uniqueness theorem*

Throughout this text we often implicitly assume uniqueness when we first guess the solution to Maxwell's equations for a given set of boundary conditions and then test that solution against those equations. This process does not guarantee that the resulting solution is unique, and often there are an infinite number of possible solutions, of which we might guess only one. The

uniqueness theorem is quite useful for it sets forth constraints on the boundary conditions that guarantee there is only one solution to Maxwell's equations, which we can find as usual.

To prove the uniqueness theorem we begin by considering a volume V enclosed by surface S and governed by Maxwell's equations:

$$\nabla \bullet \overline{D}_i = \rho \tag{2.8.1}$$

$$\nabla \bullet \overline{B}_i = 0 \qquad \nabla \times \overline{E}_i = -\frac{\partial \overline{B}_i}{\partial t} \qquad \nabla \times \overline{H}_i = \overline{J} + \frac{\partial \overline{D}_i}{\partial t} \tag{2.8.2}$$

where i = 1, 2 correspond to two possible solutions consistent with the given source distributions ρ and $\overline{J}$. We can now show that the difference $\overline{A}_d = \overline{A}_1 - \overline{A}_2$ between these two solutions must be zero under certain conditions, and therefore there can then be no more than one solution: $\overline{A}$ represents $\overline{D}, \overline{B}, \overline{E}, \overline{H}$, or $\overline{J}$.

If we subtract (2.8.1) for i = 2 from (2.8.1) for i = 1 we obtain:

$$\nabla \bullet (\overline{D}_1 - \overline{D}_2) = \nabla \bullet \overline{D}_d = 0 \tag{2.8.3}$$

Similar subtraction of corresponding equations for (2.8.2) yield three more Maxwell's equations that the difference fields $\overline{B}_d$ and $\overline{D}_d$ must satisfy:

$$\nabla \bullet \overline{B}_d = 0 \qquad \nabla \times \overline{E}_d = -\frac{\partial \overline{B}_d}{\partial t} \qquad \nabla \times \overline{H}_d = \frac{\partial \overline{D}_d}{\partial t} \tag{2.8.4}$$

where we note that the source terms ρ and $\overline{J}$ have vanished from (2.8.3) and (2.8.4) because they are given and fixed.

The boundary constraints that ensure uniqueness are:

(1) At some time t = 0 the fields are known everywhere so that at that instant $\overline{E}_d = \overline{D}_d = \overline{H}_d = \overline{B}_d = 0$.

(2) At all times and at each point on the surface S either the tangential $\overline{E}$ or the tangential $\overline{H}$ is known.

Applying Poynting's theorem (2.7.10) to the difference fields at time t proves uniqueness subject to these constraints:

$$\iiint_V \left[\overline{H}_d \bullet \frac{\partial \overline{B}_d}{\partial t} + \overline{E}_d \bullet \frac{\partial \overline{D}_d}{\partial t} \right] dv + \oiint_S (\overline{E}_d \times \overline{H}_d) \bullet d\overline{a} = 0 \tag{2.8.5}$$

Boundary constraint (2) ensures that the tangential component of either $\overline{E}_d$ or $\overline{H}_d$ is always zero, thus forcing the cross product in the second integral of (2.8.5) to zero everywhere on the enclosing surface S. The first integral can be simplified if $\overline{D} = \varepsilon\overline{E}$ and $\overline{B} = \mu\overline{H}$, where both ε and μ can be functions of position. Because this volume integral then involves only the time derivative of the squares of the difference fields $\left(|\overline{H}_d|^2 \text{ and } |\overline{E}_d|^2\right)$, and because these fields are zero at t = 0 by virtue of constraint (1), the difference fields can never depart from zero while satisfying (2.8.5). Since (2.8.5) holds for all time, the difference fields must therefore always be zero everywhere, meaning there can be no more than one solution to Maxwell's equations subject to the two constraints listed above.

Chapter 3: Electromagnetic Fields in Simple Devices and Circuits

3.1   Resistors and capacitors

3.1.1   Introduction

One important application of electromagnetic field analysis is to simple electronic components such as resistors, capacitors, and inductors, all of which exhibit at higher frequencies characteristics of the others. Such structures can be analyzed in terms of their: 1) static behavior, for which we can set $\partial/\partial t = 0$ in Maxwell's equations, 2) quasistatic behavior, for which $\partial/\partial t$ is non-negligible, but we neglect terms of the order $\partial^2/\partial t^2$, and 3) dynamic behavior, for which terms on the order of $\partial^2/\partial t^2$ are not negligible either; in the dynamic case the wavelengths of interest are no longer large compared to the device dimensions. Because most such devices have either cylindrical or planar geometries, as discussed in Sections 1.3 and 1.4, their fields and behavior are generally easily understood. This understanding can be extrapolated to more complex structures.

One approach to analyzing simple structures is to review the basic constraints imposed by symmetry, Maxwell's equations, and boundary conditions, and then to hypothesize the electric and magnetic fields that would result. These hypotheses can then be tested for consistency with any remaining constraints not already invoked. To illustrate this approach resistors, capacitors, and inductors with simple shapes are analyzed in Sections 3.1–2 below.

All physical elements exhibit varying degrees of resistance, inductance, and capacitance, depending on frequency. This is because: 1) essentially all conducting materials exhibit some resistance, 2) all currents generate magnetic fields and therefore contribute inductance, and 3) all voltage differences generate electric fields and therefore contribute capacitance. R's, L's, and C's are designed to exhibit only one dominant property at low frequencies. Section 3.3 discusses simple examples of ambivalent device behavior as frequency changes.

Most passive electronic components have two or more terminals where voltages can be measured. The voltage difference between any two terminals of a passive device generally depends on the histories of the currents through all the terminals. Common passive linear two-terminal devices include resistors, inductors, and capacitors (R's, L's. and C's, respectively), while transformers are commonly three- or four-terminal devices. Devices with even more terminals are often simply characterized as N-port networks. Connected sets of such passive linear devices form passive linear circuits which can be analyzed using the methods discussed in Section 3.4. RLC resonators and RL and RC relaxation circuits are most relevant here because their physics and behavior resemble those of common electromagnetic systems. RLC resonators are treated in Section 3.5, and RL, RC, and LC circuits are limiting cases when one of the three elements becomes negligible.

3.1.2   Resistors

Resistors are two-terminal passive linear devices characterized by their *resistance* R [ohms]:

$$v = iR \tag{3.1.1}$$

where v(t) and i(t) are the associated voltage and current. That is, one volt across a one-ohm resistor induces a one-ampere current through it; this defines the *ohm*.

The resistor illustrated in Figure 3.1.1 is comprised of two parallel perfectly conducting end-plates between which is placed a medium of conductivity σ, permittivity ε, permeability μ, and thickness d; the two end plates and the medium all have a constant cross-sectional area A [m^2] in the x-y plane. Let's assume a static voltage v exists across the resistor R, and that a current i flows through it.
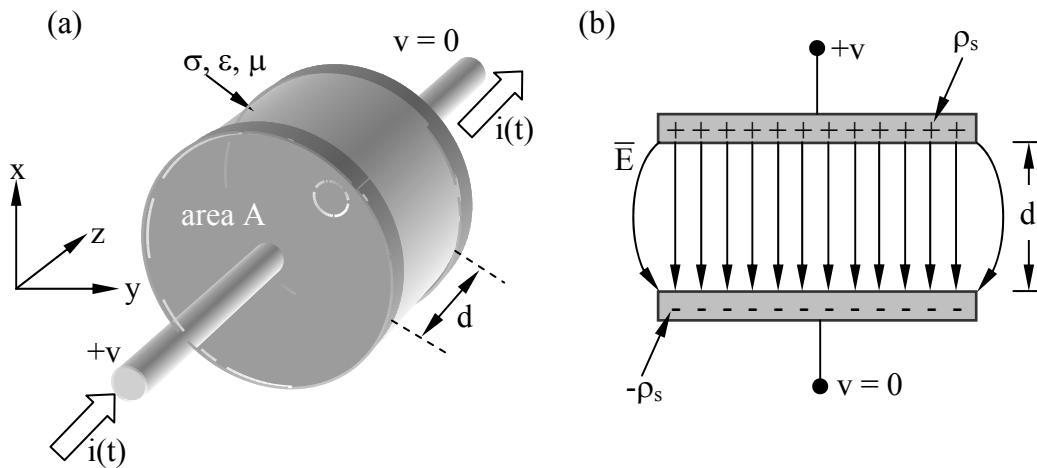
Figure 3.1.1   Simple resistor.

Boundary conditions require the electric field $\overline{E}$ at any perfectly conducting plate to be perpendicular to it [see (2.6.16); $\overline{E} \times \hat{n} = 0$], and Faraday's law requires that any line integral of $\overline{E}$ from one iso-potential end plate to the other must equal the voltage v regardless of the path of integration (1.3.13). Because the conductivity σ [Siemens/m] is uniform within walls parallel to $\hat{z}$, these constraints are satisfied by a static uniform electric field $\overline{E} = \hat{z} E_o$ everywhere within the conducting medium, which would be charge-free since our assumed $\overline{E}$ is non-divergent. Thus:

$$\int_0^d \overline{E} \bullet \hat{z} \, dz = E_o d = v \tag{3.1.2}$$

where $E_o = v/d \; \left[Vm^{-1} \right]$.

Such an electric field within the conducting medium induces a current density $\overline{J}$, where:

$$\overline{J} = \sigma \overline{E} \; \left[Am^{-2} \right] \tag{3.1.3}$$

The total current i flowing is the integral of $\overline{J} \bullet \hat{z}$ over the device cross-section A, so that:

$$i = \iint_A \overline{J} \bullet \hat{z} \ dxdy = \iint_A \sigma\overline{E} \bullet \hat{z} \ dxdy = \iint_A \sigma E_o \ dxdy = \sigma E_o A = v\sigma A/d \qquad (3.1.4)$$

But i = v/R from (3.1.1), and therefore the static resistance of a simple *planar resistor* is:

$$R = v/i = d/\sigma A \ [ohms] \qquad (3.1.5)$$

The instantaneous power p [W] dissipated in a resistor is $i^2 R = v^2/R$, and the time-average power dissipated in the sinusoidal steady state is $|\underline{I}|^2 R/2 = |\underline{V}|^2/2R$ watts. Alternatively the local instantaneous power density $P_d = \overline{E} \bullet \overline{J} \ [W \ m^{-3}]$ can be integrated over the volume of the resistor to yield the total instantaneous power dissipated:

$$p = \iiint_V \overline{E} \bullet \overline{J} \ dv = \iiint_V \overline{E} \bullet \sigma\overline{E} \ dv = \sigma|\overline{E}|^2 Ad = \sigma A v^2/d = v^2/R \ [W] \qquad (3.1.6)$$

which is the expected answer, and where we used (2.1.17): $\overline{J} = \sigma\overline{E}$.

Surface charges reside on the end plates where the electric field is perpendicular to the perfect conductor. The boundary condition $\hat{n} \bullet \overline{D} = \rho_s$ (2.6.15) suggests that the surface charge density ρ_s on the positive end-plate face adjacent to the conducting medium is:

$$\rho_s = \varepsilon E_o \ [Cm^{-2}] \qquad (3.1.7)$$

The total static charge Q on the positive resistor end plate is therefore $\rho_s A$ coulombs. By convention, the subscript s distinguishes surface charge density ρ_s [C m^{-2}] from volume charge density ρ [C m^{-3}]. An equal negative surface charge resides on the other end-plate. The total stored charged Q = $\rho_s A$ = CV, where C is the device capacitance, as discussed further in Section 3.1.3.

The static currents and voltages in this resistor will produce fields outside the resistor, but these produce no additional current or voltage at the device terminals and are not of immediate concern here. Similarly, μ and ε do not affect the static value of R. At higher frequencies, however, this resistance R varies and both inductance and capacitance appear, as shown in the following three sections. Although this static solution for charge, current, and electric field within the conducting portion of the resistor satisfies Maxwell's equations, a complete solution would also prove uniqueness and consistency with $\overline{H}$ and Maxwell's equations outside the device. Uniqueness is addressed by the uniqueness theorem in Section 2.8, and approaches to finding fields for arbitrary device geometries are discussed briefly in Sections 4.4–6.

Example 3.1A
Design a practical 100-ohm resistor. If thermal dissipation were a problem, how might that change the design?

Solution: Resistance $R = d/\sigma A$ (3.1.5), and if we arbitrarily choose a classic cylindrical shape with resistor length $d = 4r$, where r is the radius, then $A = \pi r^2 = \pi d^2/16$ and $R = 16/\pi d\sigma = 100$. Discrete resistors are smaller for modern low power compact circuits, so we might set $d = 1$ mm, yielding $\sigma = 16/\pi dR = 16/(\pi 10^{-3} \times 100) \cong 51$ S m^{-1}. Such conductivities roughly correspond, for example, to very salty water or carbon powder. The surface area of the resistor must be sufficient to dissipate the maximum power expected, however. Flat resistors thermally bonded to a heat sink can be smaller than air-cooled devices, and these are often made of thin metallic film. Some resistors are long wires wound in coils. Resistor failure often occurs where the local resistance is slightly higher, and the resulting heat typically increases the local resistance further, causing even more local heating.

3.1.3   Capacitors

Capacitors are two-terminal passive linear devices storing charge Q and characterized by their *capacitance* C [Farads], defined by:

$$Q = Cv \quad \text{[Coulombs]} \tag{3.1.8}$$

where v(t) is the voltage across the capacitor. That is, one static volt across a one-Farad capacitor stores one Coulomb on each terminal, as discussed further below; this defines the *Farad* [Coulombs per volt].

The resistive structure illustrated in Figure 3.1.1 becomes a pure capacitor at low frequencies if the media conductivity $\sigma \to 0$. Although some capacitors are air-filled with $\varepsilon \cong \varepsilon_o$, usually dielectric filler with permittivity $\varepsilon > \varepsilon_o$ is used. Typical values for the *dielectric constant* $\varepsilon/\varepsilon_o$ used in capacitors are ~1-100. In all cases boundary conditions again require that the electric field $\overline{E}$ be perpendicular to the perfectly conducting end plates, i.e., to be in the $\pm z$ direction, and Faraday's law requires that any line integral of $\overline{E}$ from one iso-potential end plate to the other must equal the voltage v across the capacitor. These constraints are again satisfied by a static uniform electric field $\overline{E} = \hat{z}E_o$ within the medium separating the plates, which is uniform and charge-free.

We shall neglect temporarily the effects of all fields produced outside the capacitor if its plate separation d is small compared to its diameter, a common configuration. Thus $E_o = v/d$ [V m^{-1}] (3.1.2). The surface charge density on the positive end-plate face adjacent to the conducting medium is $\sigma_s = \varepsilon E_o$ [C m^{-2}], and the total static charge Q on the positive end plate of area A is therefore:

$$Q = A\sigma_s = A\varepsilon E_o = A\varepsilon v/d = Cv \quad \text{[C]} \tag{3.1.9}$$

Therefore, for a *parallel-plate capacitor*:

$$C = \varepsilon A/d \ \ [\text{Farads}] \qquad\qquad \text{(parallel-plate capacitor)} \quad (3.1.10)$$

Using (3.1.2) and the fact that the charge $Q(t)$ on the positive plate is the time integral of the current $i(t)$ into it, we obtain the relation between voltage and current for a capacitor:

$$v(t) = Q(t)/C = (1/C)\int_{-\infty}^{t} i(t)\,dt \qquad\qquad (3.1.11)$$

$$i(t) = C \ dv(t)/dt \qquad\qquad (3.1.12)$$

When two capacitors are connected in parallel as shown in Figure 3.1.2, they are equivalent to a single capacitor of value C_{eq} storing charge Q_{eq}, where these values are easily found in terms of the charges (Q_1, Q_2) and capacitances (C_1, C_2) associated with the two separate devices.
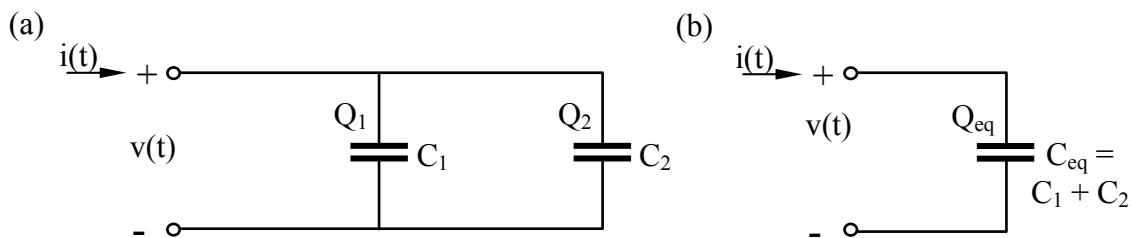
Figure 3.1.2   Capacitors in parallel.

Because the total charge Q_{eq} is the sum of the charges on the two separate capacitors, and capacitors in parallel have the same voltage v, it follows that:

$$Q_{eq} = Q_1 + Q_2 = (C_1 + C_2)v = C_{eq}v \qquad\qquad (3.1.13)$$

$$C_{eq} = C_1 + C_2 \qquad\qquad \text{(capacitors in parallel)} \quad (3.1.14)$$
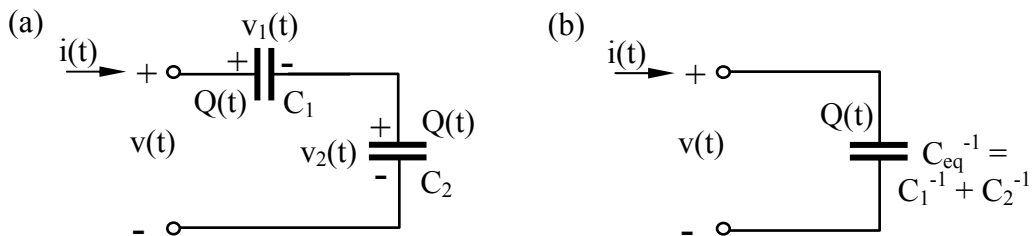
Figure 3.1.3   Capacitors in series.

When two capacitors are connected in series, as illustrated in Figure 3.1.3, then their two charges Q_1 and Q_2 remain equal if they were equal before current i(t) began to flow, and the total voltage is the sum of the voltages across each capacitor:

$$C_{eq}^{-1} = v/Q = (v_1 + v_2)/Q = C_1^{-1} + C_2^{-1} \qquad \text{(capacitors in series)} \qquad (3.1.15)$$

The instantaneous electric energy density W_e [J m^{-3}] between the capacitor plates is given by Poynting's theorem: $W_e = \varepsilon|\overline{E}|^2\big/2$ (2.7.7). The total electric energy w_e stored in the capacitor is the integral of W_e over the volume Ad of the dielectric:

$$w_e = \iiint_V \left(\varepsilon|\overline{E}|^2\big/2\right)dv = \varepsilon Ad|\overline{E}|^2\big/2 = \varepsilon Av^2\big/2d = Cv^2\big/2 \quad [J] \qquad (3.1.16)$$

The corresponding expression for the time-average energy stored in a capacitor in the sinusoidal steady state is:

$$w_e = C|\underline{V}|^2\big/4 \quad [\ ] \qquad (3.1.17)$$

The extra factor of two relative to (3.1.9) enters because the time average of a sinusoid squared is half its peak value.

To prove (3.1.16) for any capacitor C, not just parallel-plate devices, we can compute $w_e = \int_0^t iv\ dt$ where $i = dq/dt$ and $q = Cv$. Therefore $w_e = \int_0^t C(dv/dt)v\ dt = \int_0^v Cv\ dv = Cv^2\big/2$ in general.

We can also analyze other capacitor geometries, such as the cylindrical capacitor illustrated in Figure 3.1.4. The inner radius is "a", the outer radius is "b", and the length is D; its interior has permittivity ε.
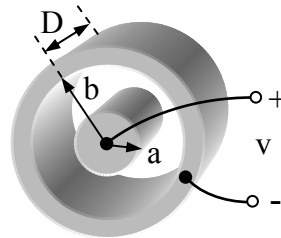
Figure 3.1.4   Cylindrical capacitor.

The electric field again must be divergence- and curl-free in the charge-free regions between the two cylinders, and must be perpendicular to the inner and outer cylinders at their perfectly conducting walls. The solution can be cylindrically symmetric and independent of ϕ. A purely radial electric field has these properties:

$$\overline{E}(r) = \hat{r}E_o/r \tag{3.1.18}$$

The electric potential $\Phi(r)$ is the integral of the electric field, so the potential difference v between the inner and outer conductors is:

$$v = \Phi_a - \Phi_b = \int_a^b \frac{E_o}{r} dr = E_o \ln r \Big|_a^b = E_o \ln\left(\frac{b}{a}\right) \text{ [V]} \tag{3.1.19}$$

This capacitor voltage produces a surface charge density ρ_s on the inner and outer conductors, where $\rho_s = \varepsilon E = \varepsilon E_o/r$. If $\Phi_a > \Phi_b$, then the inner cylinder is positively charged, the outer cylinder is negatively charged, and E_o is positive. The total charge Q on the inner cylinder is then:

$$Q = \rho_s 2\pi aD = \varepsilon E_o 2\pi D = \varepsilon v 2\pi D / \left[\ln(b/a)\right] = Cv \text{ [ ]} \tag{3.1.20}$$

Therefore this *cylindrical capacitor* has capacitance C:

$$C = \varepsilon 2\pi D / \left[\ln(b/a)\right] \text{ [ ]} \qquad \text{(cylindrical capacitor)} \tag{3.1.21}$$

In the limit where $b/a \to 1$ and $b - a = d$, then we have approximately a parallel-plate capacitor with $C \to \varepsilon A/d$ where the plate area $A = 2\pi aD$; see (3.1.10).

Example 3.1B
Design a practical 100-volt 10^{-8} farad (0.01 mfd) capacitor using dielectric having $\varepsilon = 20\varepsilon_o$ and a breakdown field strength E_B of 10^7 [V m^{-1}].

Solution: For parallel-plate capacitors $C = \varepsilon A/d$ (3.1.10), and the device breakdown voltage is $E_B d = 100$ [V]. Therefore the plate separation $d = 100/E_B = 10^{-5}$ [m]. With a safety factor of two, d doubles to 2×10^{-5}, so $A = dC/\varepsilon = 2\times10^{-5} \times 10^{-8}/(20 \times 6.85\times10^{-12}) \cong 1.5\times10^3$ [m^2]. If the capacitor is a cube of side D, then the capacitor volume is $D^3 =$ Ad and $D = (Ad)^{0.333} = (1.5\times10^{-3} \times 2\times10^{-5})^{0.333} \cong 3.1$ mm. To simplify manufacture, such capacitors are usually wound in cylinders or cut from flat stacked sheets.

3.2 Inductors and transformers

3.2.1 Solenoidal inductors

All currents in devices produce magnetic fields that store magnetic energy and therefore contribute inductance to a degree that depends on frequency. When two circuit branches share magnetic fields, each will typically induce a voltage in the other, thus *coupling* the branches so they form a transformer, as discussed in Section 3.2.4.

Inductors are two-terminal passive devices specifically designed to store magnetic energy, particularly at frequencies below some design-dependent upper limit. One simple geometry is shown in Figure 3.2.1 in which current i(t) flows in a loop through two perfectly conducting parallel plates of width W and length D, spaced d apart, and short-circuited at one end.
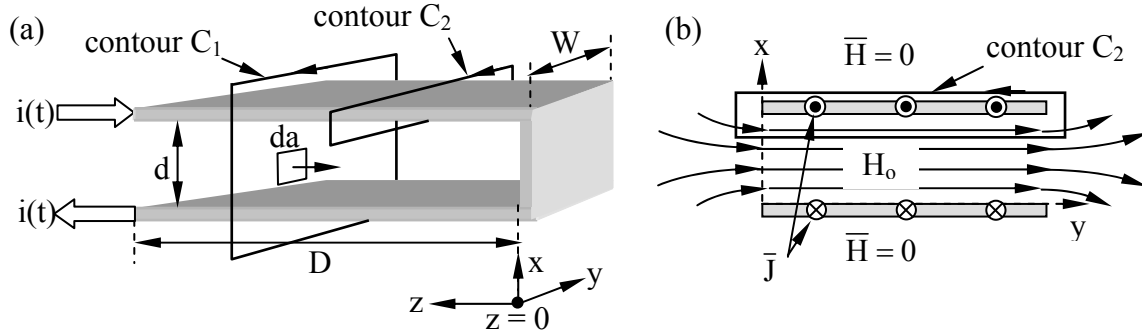
Figure 3.2.1 Parallel-plate inductor.

To find the magnetic field from the currents we can use the integral form of Ampere's law, which links the variables $\overline{H}$ and $\overline{J}$:

$$\oint_C \overline{H} \bullet d\overline{s} = \iint_A \left(\overline{J} + \partial\overline{D}/\partial t \right) \bullet d\overline{a} \tag{3.2.1}$$

The contour C_1 around both currents in Figure 3.2.1 encircles zero net current, and (3.2.1) says the contour integral of $\overline{H}$ around zero net current must be zero in the static case. Contour C_2 encircles only the current i(t), so the contour integral of $\overline{H}$ around any C_2 in the right-hand sense equals i(t) for the static case. The values of these two contour integrals are consistent with zero magnetic field outside the pair of plates and a constant field $\overline{H} = H_o \hat{y}$ between them, although a uniform magnetic field could be superimposed everywhere without altering those integrals. Since such a uniform field would not have the same symmetry as this device, such a field would have to be generated elsewhere. These integrals are also exactly consistent with fringing fields at the edges of the plate, as illustrated in Figure 3.2.1(b) in the x-y plane for z > 0. Fringing fields can usually be neglected if the plate separation d is much less than the plate width W.

It follows that:

$$\oint_{C2} \overline{H} \bullet d\overline{s} = i(t) = H_o W \tag{3.2.2}$$

$$\overline{H} = \hat{y}\, H_o = \hat{y}\, i(t)/W \left[A\ m^{-1} \right] \qquad (\overline{H} \text{ between the plates}) \tag{3.2.3}$$

and $\overline{H} \cong 0$ elsewhere.
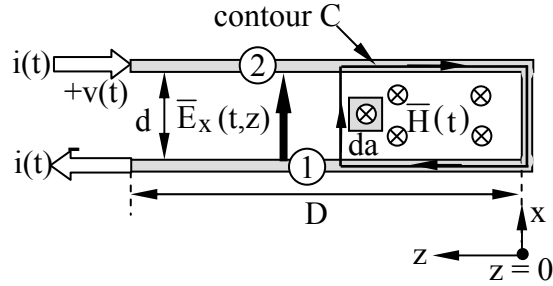
Figure 3.2.2   Voltages induced on a parallel-plate inductor.

The voltage v(t) across the terminals of the inductor illustrated in Figures 3.2.1 and 3.2.2 can be found using the integral form of Faraday's law and (3.2.3):

$$\oint_C \overline{E} \bullet d\overline{s} = -\frac{\partial}{\partial t} \iint_A \mu \overline{H} \bullet d\overline{a} = -\frac{\mu Dd}{W} \frac{di(t)}{dt} = \int_1^2 E_x(t,z)dx = -v(t,z) \qquad (3.2.4)$$

where z = D at the inductor terminals. Note that when we integrate $\overline{E}$ around contour C there is zero contribution along the path inside the perfect conductor; the non-zero portion is restricted to the illustrated path 1-2. Therefore:

$$v(t) = \frac{\mu Dd}{W} \frac{di(t)}{dt} = L \frac{di(t)}{dt} \qquad (3.2.5)$$

where (3.2.5) defines the *inductance* L [Henries] of any inductor. Therefore L_1 for a single-turn current loop having length W >> d and area A = Dd is:

$$L_1 = \frac{\mu Dd}{W} = \frac{\mu A}{W} \ [H] \qquad \text{(single-turn wide inductor)} \qquad (3.2.6)$$

To simplify these equations we define *magnetic flux* ψ_m as[8]:

$$\psi_m = \iint_A \mu \overline{H} \bullet d\overline{a} \ [\text{Webers} = \text{Vs}] \qquad (3.2.7)$$

Then Equations (3.2.4) and (3.2.7) become:

$$v(t) = d\psi_m(t)/dt \qquad (3.2.8)$$

$$\psi_m(t) = L\ i(t) \qquad \text{(single-turn inductor)} \qquad (3.2.9)$$

---

[8] The symbol ψ_m for magnetic flux [Webers] should not be confused with Ψ for magnetic potential [Amperes].

Since we assumed fringing fields could be neglected because W >> d, large single-turn inductors require very large structures. The standard approach to increasing inductance L in a limited volume is instead to use multi-turn coils as illustrated in Figure 3.2.3.
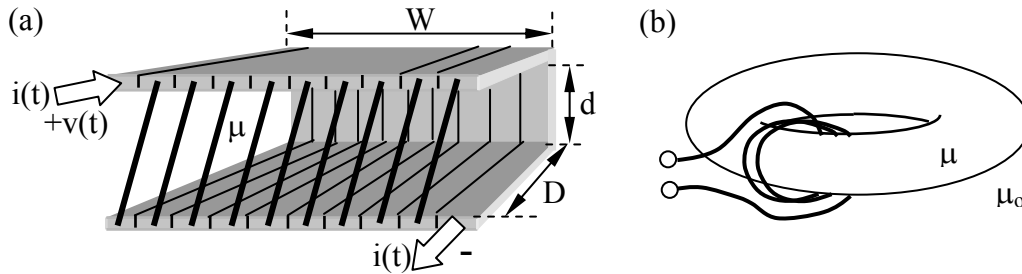
Figure 3.2.3 N-turn inductor: (a) solenoid, (b) toroid.

The N-turn coil of Figure 3.2.3 duplicates the current flow geometry illustrated in Figures 3.2.1 and 3.2.2, but with N times the intensity (A m^{-1}) for the same terminal current i(t), and therefore the magnetic field H_o and flux ψ_m are also N times stronger than before. At the same time the voltage induced in each turn is proportional to the flux ψ_m through it, which is now N times greater than for a single-turn coil ($\psi_m = Ni\mu A/W$), and the total voltage across the inductor is the sum of the voltages across the N turns. Therefore, provided that W >> d, the total voltage across an N-turn inductor is N^2 times its one-turn value, and the inductance L_N of an N-turn coil is also N^2 greater than L_1 for a one-turn coil:

$$v(t) = L_N \frac{di(t)}{dt} = N^2 L_1 \frac{di(t)}{dt} \qquad\qquad (3.2.10)$$

$$L_N = N^2 \frac{\mu A}{W} \;[\;\;] \qquad\qquad \text{(N-turn solenoidal inductor)} \quad (3.2.11)$$

where A is the coil area and W is its length; $W >> \sqrt{A} > d$.

Equation (3.2.11) also applies to cylindrical coils having W >> d, which is the most common form of inductor. To achieve large values of N the turns of wire can be wound on top of each other with little adverse effect; (3.2.11) still applies.

These expressions can also be simplified by defining *magnetic flux linkage* Λ as the magnetic flux ψ_m (3.2.7) linked by N turns of the current i, where:

$$\Lambda = N\psi_m = N(Ni\mu A/W) = (N^2\mu A/W)i = Li \qquad \text{(flux linkage)} \quad (3.2.12)$$

This equation $\Lambda = Li$ is dual to the expression $Q = Cv$ for capacitors. We can use (3.2.5) and (3.2.12) to express the voltage v across N turns of a coil as:

$$v = L\, di/dt = d\Lambda/dt \qquad\qquad \text{(any coil linking magnetic flux } \Lambda) \quad (3.2.13)$$

The net inductance L of two inductors L_1 and L_2 in series or parallel is related to L_1 and L_2 in the same way two connected resistors are related:

$$L = L_1 + L_2 \qquad \text{(series combination)} \qquad (3.2.14)$$

$$L^{-1} = L_1^{-1} + L_2^{-1} \qquad \text{(parallel combination)} \qquad (3.2.15)$$

For example, two inductors in series convey the same current i but the total voltage across the pair is the sum of the voltages across each – so the inductances add.

Example 3.2A
Design a 100-Henry air-wound inductor.

Solution: Equation (3.2.11) says $L = N^2\mu A/W$, so N and the form factor A/W must be chosen. Since $A = \pi r^2$ is the area of a cylindrical inductor of radius r, then $W = 4r$ implies $L = N^2\mu\pi r/4$. Although tiny inductors (small r) can be achieved with a large number of turns N, N is limited by the ratio of the cross-sectional areas of the coil rW and of the wire πr_w^2, and is $N \cong r^2/r_w^2$. N is further limited if we want the resistive impedance R $\ll j\omega L$. If ω_{min} is the lowest frequency of interest, then we want $R \cong \omega_{min}L/100 = d/(\sigma\pi r_w^2)$ [see (3.1.5)], where the wire length $d \cong 2\pi rN$. These constraints eventually yield the desired values for r and N that yield the smallest inductor. Example 3.2B carries these issues further.

3.2.2   Toroidal inductors

The prior discussion assumed μ filled all space. If μ is restricted to the interior of a solenoid, L is diminished significantly, but coils wound on a high-μ *toroid*, a donut-shaped structure as illustrated in Figure 3.2.3(b), yield the full benefit of high values for μ. Typical values of μ are ~5000 to 180,000 for iron, and up to ~10^6 for special materials.

Coils wound on high-permeability toroids exhibit significantly less flux leakage than solenoids. Consider the boundary between air and a high-permeability material ($\mu/\mu_o \gg 1$), as illustrated in Figure 3.2.4.
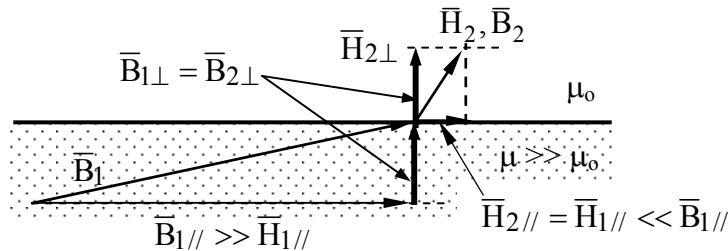
Figure 3.2.4   Magnetic fields at high-permeability boundaries.

The degree to which $\overline{B}$ is parallel or perpendicular to the illustrated boundary has been diminished substantially for the purpose of clarity. The boundary conditions are that both $\overline{B}_\perp$ and $\overline{H}_{//}$ are continuous across any interface (2.6.5, 2.6.11). Since $\overline{B} = \mu\overline{H}$ in the permeable core and $\overline{B} = \mu_o\overline{H}$ in air, and since $\overline{H}_{//}$ is continuous across the boundary, therefore $\overline{B}_{//}$ changes across the boundary by the large factor μ/μ_o. In contrast, $\overline{B}_\perp$ is the same on both sides. Therefore, as suggested in Figure 3.2.4, $\overline{B}_2$ in air is nearly perpendicular to the boundary because $\overline{H}_{//}$, and therefore $\overline{B}_{2//}$, is so very small; note that the figure has been scaled so that the arrows representing $\overline{H}_2$ and $\overline{B}_2$ have the same length when $\mu = \mu_0$.
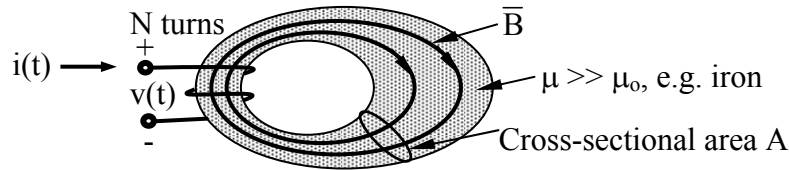
Figure 3.2.5   Toroidal inductor.

In contrast, $\overline{B}_1$ is nearly parallel to the boundary and is therefore largely trapped there, even if that boundary curves, as shown for a toroid in Figure 3.2.5. The reason magnetic flux is largely trapped within high-μ materials is also closely related to the reason current is trapped within high-σ wires, as described in Section 4.3.

The inductance of a *toroidal inductor* is simply related to the linked magnetic flux Λ by (3.2.12) and (3.2.7):

$$L = \frac{\Lambda}{i} = \frac{\mu N \iint_A \overline{H} \bullet d\overline{a}}{i} \qquad \text{(toroidal inductor)} \quad (3.2.16)$$

where A is any cross-sectional area of the toroid.

Computing $\overline{H}$ is easier if the toroid is circular and has a constant cross-section A which is small compared to the major radius R so that $R \gg \sqrt{A}$. From Ampere's law we learn that the integral of $\overline{H}$ around the $2\pi R$ circumference of this toroid is:

$$\oint_C \overline{H} \bullet d\overline{s} \cong 2\pi R H \cong Ni \qquad (3.2.17)$$

where the only linked current is i(t) flowing through the N turns of wire threading the toroid. Equation (3.2.17) yields $H \cong Ni/2\pi R$ and (3.2.16) relates H to L. Therefore the inductance L of such a toroid found from (3.2.16) and (3.2.17) is:

$$L \cong \frac{\mu NA}{i} \frac{Ni}{2\pi R} = \frac{\mu N^2 A}{2\pi R} \quad \text{[Henries]} \qquad \text{(toroidal inductor)} \quad (3.2.18)$$

The inductance is proportional to μ, N^2, and cross-sectional area A, but declines as the toroid major radius R increases. The most compact large-L toroids are therefore fat (large A) with almost no hole in the middle (small R); the hole size is determined by N (made as large as possible) and the wire diameter (made small). The maximum acceptable series resistance of the inductor limits N and the wire diameter; for a given wire mass [kg] this resistance is proportional to N^2.
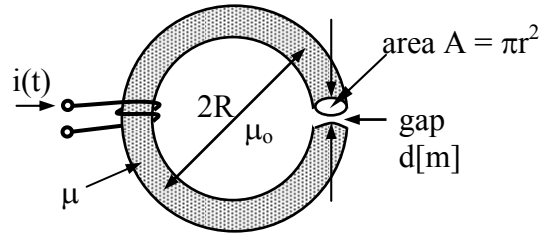
Figure 3.2.6   Toroidal inductor with a small gap.

The inductance of a high-permeability toroid is strongly reduced if even a small gap of width d exists in the magnetic path, as shown in Figure 3.2.6. The inductance L of a toroid with a gap of width d can be found using (3.2.16), but first we must find the magnitude of H_μ within the toroid. Again we can use the integral form of Ampere's law for a closed contour along the axis of the toroid, encircling the hole.

$$\oint_C \overline{H} \bullet d\overline{s} \cong (2\pi R - d)H_\mu + H_g d \cong Ni \qquad (3.2.19)$$

where H_g is the magnitude of H within the gap. Since $\overline{B}_\perp$ is continuous across the gap faces, $\mu_o H_g = \mu H_\mu$ and these two equations can be solved for the two unknowns, H_g and H_μ. The second term $H_g d$ can be neglected if the gap width $d \ll 2\pi R \mu_o/\mu$. In this limiting case we have the same inductance as before, (3.2.18). However, if $A^{0.5} > d \gg 2\pi R \mu_o/\mu$, then $H_g \cong Ni/d$ and:

$$L = \Lambda/i \cong N\psi_m/i \cong N\mu_o H_g A/i \cong N^2 \mu_o A \big/ d \quad [\text{H}] \qquad (\textit{toroid with a gap}) \qquad (3.2.20)$$

Relative to (3.2.18) the inductance has been reduced by a factor of μ_o/μ and increased by a much smaller factor of $2\pi R/d$, a significant net reduction even though the gap is small.

Equation (3.2.20) suggests how small air gaps in magnetic motors limit motor inductance and sometimes motor torque, as discussed further in Section 6.3. Gaps can be useful too. For example, if μ is non-linear [$\mu = f(H)$], then $L \neq f(H)$ if the gap and μ_o dominate L. Also, inductance dominated by gaps can store more energy when H exceeds saturation (i.e., $B^2/2\mu_o \gg B_{SAT}^2/2\mu$).

3.2.3   Energy storage in inductors

The energy stored in an inductor resides in its magnetic field, which has an instantaneous energy density of:

$$W_m(t) = \mu|\overline{H}|^2 \big/ 2 \;\left[\mathrm{J\;m^{-3}}\right] \tag{3.2.21}$$

Since the magnetic field is uniform within the volume Ad of the rectangular inductor of Figure 3.2.1, the total instantaneous magnetic energy stored there is:

$$w_m \cong \mu AW|\overline{H}|^2 \big/ 2 \cong \mu AW(i/W)^2 \big/ 2 \cong Li^2 \big/ 2 \;\;[\mathrm{J}] \tag{3.2.22}$$

That (3.2.22) is valid and exact for any inductance L can be shown using Poynting's theorem, which relates power P = vi at the device terminals to changes in energy storage:

$$w_m = \int_{-\infty}^{t} v(t)i(t)\,dt = \int_{-\infty}^{t} L(di/dt)i\;dt = \int_{0}^{i} Li\;di = Li^2 \big/ 2 \,[\mathrm{J}] \tag{3.2.23}$$

Earlier we neglected fringing fields, but they store magnetic energy too. We can compute them accurately using the Biot-Savart law (10.2.21), which is derived later and expresses $\overline{H}$ directly in terms of the currents flowing in the inductor:

$$\overline{H}(\overline{r}) = \iiint_{V'} dv'\left[\overline{J}(\overline{r'}) \times (\overline{r} - \overline{r'})\right]\big/\left[4\pi|\overline{r} - \overline{r'}|^3\right] \tag{3.2.24}$$

The magnetic field produced by current $\overline{J}(\overline{r'})$ diminishes with distance squared, and therefore the magnitude of the uniform field $\overline{H}$ within the inductor is dominated by currents within a distance of ~d of the inductor ends, where d is the nominal diameter or thickness of the inductor [see Figure 3.2.3(a) and assume $d \cong D << W$]. Therefore $|\overline{H}|$ at the center of the end-face of a semi-infinite cylindrical inductor has precisely half the strength it has near the middle of the same inductor because the Biot-Savart contributions to $\overline{H}$ at the end-face arise only from one side of the end-face, not from both sides.

The energy density within a solenoidal inductor therefore diminishes within a distance of ~d from each end, but this is partially compensated in (3.2.23) by the neglected magnetic energy outside the inductor, which also decays within a distance ~d. For these reasons fringing fields are usually neglected in inductance computations when d << W. Because magnetic flux is non-divergent, the reduced field intensity near the ends of solenoids implies that some magnetic field lines escape the coil there; they are fully trapped within the rest of the coil.

The energy stored in a thin toroidal inductor can be found using (3.2.21):

$$w_m \cong \left(\mu|\overline{H}|^2 \big/ 2\right) A 2\pi R \tag{3.2.25}$$

The energy stored in a toroidal inductor with a non-negligible gap of width d can be easily found knowing that the energy storage in the gap dominates that in the high-permeability toroid, so that:

$$w_m \cong \left(\mu_o H_g^2 / 2\right) Ad \cong \mu_o \left(Ni/d\right)^2 Ad/2 \cong Li^2/2 \qquad (3.2.26)$$

Example 3.2B

Design a practical 100-Henry inductor wound on a toroid having $\mu = 10^4 \mu_o$; it is to be used for $\omega \cong 400$ [r s^{-1}] (~60 Hz). How many Joules can it store if the current is one Ampere? If the residual flux density B_r of the toroid is 0.2 Tesla, how does this affect design?

Solution: We have at least three unknowns, i,e., size, number of turns N, and wire radius r_w, and therefore need at least three equations. Equation (3.2.18) says $L \cong \mu N^2 A / 2\pi R_m$ where $A = \pi r^2$. A fat toroid might have major radius $R_m \cong 3r$, corresponding to a central hole of radius 2r surrounded by an iron torus 2r thick, yielding an outer diameter of 4r. Our first equation follows: $L = 100 \cong \mu N^2 r/6$. Next, the number N of turns is limited by the ratio of the cross-sectional area of the hole in the torus ($\pi 4r^2$) and the cross-sectional area of the wire πr_w^2; our second equation is $N \cong 4r^2/r_w^2$. Although tiny inductors (small r) can be achieved with large N, N is limited if we want the resistive impedance $R \ll \omega L$. If ω_{min} is the lowest frequency of interest, then we obtain our third equation, $R \cong \omega_{min} L/100 = 400 = d/(\sigma \pi r_w^2)$ [see (3.1.5)], where the wire length $d \cong 4\pi r N$. Eliminating r_w^2 from the second and third equation yields $N^2 \cong 400\sigma r$, and eliminating N^2 from the first equation yields $r = (600/400\sigma\mu)^{0.5} \cong 1.5$mm, where for typical wires $\sigma \cong 5\times10^7$; the maximum diameter of this toroid is $8r \cong 1.2$ cm. Since $N^2 \cong 400\sigma r$, therefore $N \cong 5600$, and $r_w \cong 2r/\sqrt{N} \cong 40$ microns.

We might suppose the stored energy $w_m = Li^2/2 = 100\times1^2/2 = 50$ joules. However, if 1 ampere flows through 5600 turns, and if $H = 5600/2\pi 3r = 5600/0.031 = 1.8\times10^5$ [A m^{-1}], then $B = \mu H \cong 2300$ Tesla, well above the limit of $B_r = 0.2$ Tesla where saturation was said to occur. Since the incremental μ_o applies at high currents, this device is quite non-linear and the computed stored energy of 50J should be reduced by a factor of $\sim\mu_o/\mu$ to yield ~5 mJ. If linearity and low loss ($R \ll \omega L$) are desired, either this toroid must be made much larger so that the upper limit on μH inside the toroid is not exceeded, or the maximum current must be reduced to the ~100 μA level. Moreover, a sinusoidal current of 1 ampere through this small 400-ohm resistance would dissipate 200 W, enough to damage it. Note that if ω_{min} is increased by a factor of F, then r decreases by $F^{0.5}$.

3.2.4   Transformers

Transformers are passive devices used to raise or lower the voltages of alternating currents or transients. The voltage v across two terminals of any coil can be found using Faraday's law (2.4.14):

$$\oint_C \overline{E} \bullet d\overline{s} = -\frac{d}{dt} \iint_A \mu_o \overline{H} \bullet d\overline{a}$$

(3.2.27)

which leads to the voltage across any N turns of a coil, as given by (3.2.13):

$$v = d\Lambda/dt$$

(3.2.28)

where the flux linkage $\Lambda = N\psi_m$ and the magnetic flux ψ_m within the cross-sectional area A of the coil is defined by (3.2.7):

$$\Psi_m = \iint_A \mu \overline{H} \bullet d\overline{a} \quad [\text{Webers} = \text{Vs}]$$

(3.2.29)

Consider the ideal toroidal transformer of Figure 3.2.7.
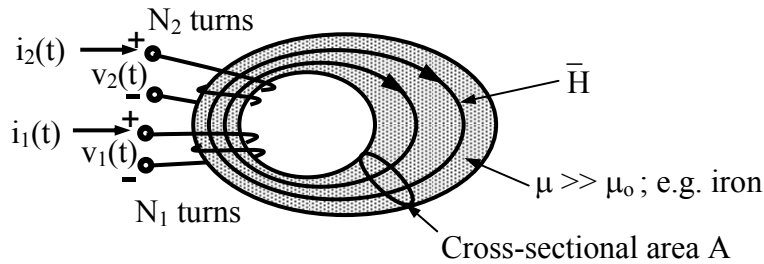
Figure 3.2.7   Toroidal transformer.

Its high permeability traps the magnetic flux within it so that ψ_m is constant around the toroid, even though A varies. From (3.2.28) we see that the voltage v_k across coil k is therefore:

$$v_k = d\Lambda_k/dt = N_k d\Psi_m/dt$$

(3.2.30)

The ratio between the voltages across two coils k = 1,2 is therefore:

$$v_2/v_1 = N_2/N_1$$

(3.2.31)

where N_2/N_1 is the transformer turns ratio.

   If current i_2 flows in the output coil, then there will be an added contribution to v_1 and v_2 due to the contributions of i_2 to the original ψ_m from the input coil alone. Note that current flowing into the "+" terminal of both coils in the figure contribute to $\overline{H}$ in the illustrated direction; this

distinguishes the positive terminal from the negative terminal of each coil. If the flux coupling between the two coils is imperfect, then the output voltage is correspondingly reduced. Any resistance in the wires can increment these voltages in proportion to the currents.

Figure 3.2.8 suggests traditional symbols used to represent ideal transformers and some common configurations used in practice. The polarity dot at the end of each coil indicates which terminals would register the same voltage for a given change in the linked magnetic flux. In the absence of dots, the polarity indicated in (a) is understood. Note that many transformers consist of a single coil with multiple taps. Sometimes one of the taps is a commutator that can slide across the coil windings to provide a continuously variable transformer turns ratio. As illustrated, the presence of an iron core is indicated by parallel lines and an auto-transformer consists of only one tapped coil.
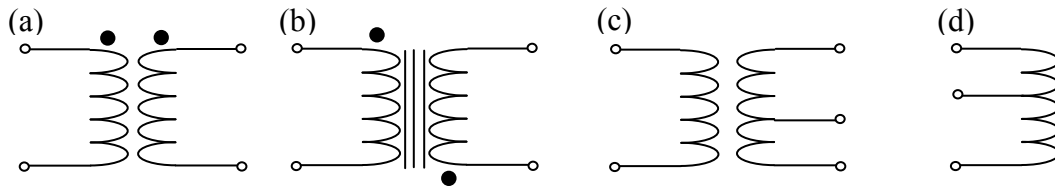
Figure 3.2.8   Transformer configurations:
(a) air-core, (b) iron-core, (c) tapped, and (d) auto-transformer.

The terminal voltages of linear transformers for which $\mu \neq f(H)$ are linearly related to the various currents flowing through the windings. Consider a simple toroid for which H, B, and the cross-sectional area A are the same everywhere around the average circumference πD. In this case the voltage $\underline{V}_1$ across the N_1 turns of coil (1) is:

$$\underline{V}_1 = j\omega N_1 \qquad\qquad (3.2.32)$$

$$\underline{\Psi} = \mu \underline{H} A \qquad\qquad (3.2.33)$$

$$\underline{H} = (N_1\underline{I}_1 + N_2\underline{I}_2)/\pi D \qquad\qquad (3.2.34)$$

Therefore:

$$\underline{V}_1 = j\omega[\mu A N_1(N_1\underline{I}_1 + N_2\underline{I}_2)/\pi D] = j\omega(L_{11}\underline{I}_1 + L_{12}\underline{I}_2) \qquad\qquad (3.2.35)$$

where the *self-inductance* L_{11} and *mutual inductance* L_{12} [Henries] are:

$$L_{11} = \mu A N_1{}^2/\pi D \qquad\qquad L_{12} = \mu A N_1 N_2/\pi D \qquad\qquad (3.2.36)$$

Equation (3.2.35) can be generalized for a two-coil transformer:

$$\begin{bmatrix} \underline{V}_1 \\ \underline{V}_2 \end{bmatrix} = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} \underline{I}_1 \\ \underline{I}_2 \end{bmatrix} \qquad\qquad (3.2.37)$$

Consider the simple toroidal step-up transformer illustrated in Figure 3.2.9 in which the voltage source drives the load resistor R through the transformer, which has N_1 and N_2 turns on its input and output, respectively. The toroid has major diameter D and cross-sectional area A.
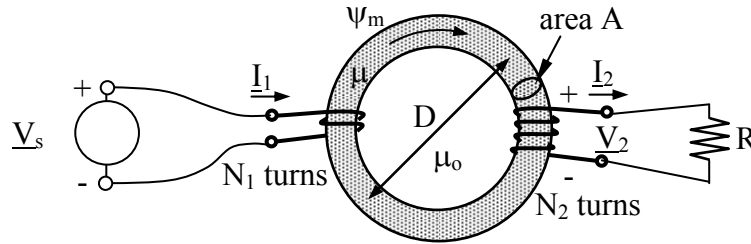
Figure 3.2.9   Toroidal step-up transformer loaded with resistor R.

Combining (3.2.33) and (3.2.34), and noting that the sign of I_2 has been reversed in the figure, we obtain the expression for total flux:

$$\underline{\Psi} = \mu A\, (N_1\underline{I}_1 + N_2\underline{I}_2)/\pi D \tag{3.2.38}$$

We can find the admittance seen by the voltage source by solving (3.2.38) for $\underline{I}_1$ and dividing by $\underline{V}_s$:

$$\underline{I}_1 = (\pi D\underline{\Psi}/\mu A N_1) + \underline{I}_2\, N_2/N_1 \tag{3.2.39}$$

$$\underline{V}_s = j\omega N_1\underline{\Psi} = \underline{V}_2 N_1/N_2 = \underline{I}_2 R N_1/N_2 \tag{3.2.40}$$

$$\underline{I}_1/\underline{V}_s = (\pi D\underline{\Psi}/\mu A N_1)/j\omega N_1\underline{\Psi} + \underline{I}_2 N_2{}^2/(N_1{}^2\underline{I}_2 R) \tag{3.2.41}$$

$$= -\, j\pi D/(\omega N_1{}^2 \mu A) + (N_2/N_1)^2/R = 1/j\omega L_{11} + (N_2/N_1)^2/R \tag{3.2.42}$$

Thus the admittance seen at the input to the transformer is that of the self-inductance $(1/j\omega L_{11})$ in parallel with the admittance of the transformed resistance $[(N_2/N_1)^2/R]$. The power delivered to the load is $|\underline{V}_2|^2/2R = |\underline{V}_1|^2(N_2/N_1)^2/2R$, which is the time-average power delivered to the transformer, since $|\underline{V}_2|^2 = |\underline{V}_1|^2(N_2/N_1)^2$; see (3.2.31).

The *transformer equivalent circuit* is thus L_{11} in parallel with the input of an ideal transformer with turns ratio N_2/N_1. Resistive losses in the input and output coils could be represented by resistors in series with the input and output lines. Usually $j\omega L_{11}$ for an iron-core transformer is so great that only the ideal transformer is important.

One significant problem with iron-core transformers is that the changing magnetic fields within them can generate considerable voltages and *eddy currents* by virtue of Ohm's Law ($\overline{J} = \sigma\, \overline{E}$) and Faraday's law:

$$\oint_C \overline{E} \bullet d\overline{s} = -j\omega\mu\int_A \overline{H} \bullet d\overline{a} \qquad (3.2.43)$$

where the contour C circles each conducting magnetic element. A simple standard method for reducing the eddy currents $\underline{J}$ and the associated dissipated power $\int_V (\sigma|\underline{J}|^2/2)dv$ is to reduce the area A by laminating the core; i.e., by fabricating it with thin stacked insulated slabs of iron or steel oriented so as to interrupt the eddy currents. The eddy currents flow perpendicular to $\overline{H}$, so the slab should be sliced along the direction of $\overline{H}$. If N stacked slabs replace a single slab, then A, $\underline{E}$, and $\underline{J}$ are each reduced roughly by a factor of N, so the power dissipated, which is proportional to the square of J, is reduced by a factor of $\sim N^2$. Eddy currents and laminated cores are discussed further at the end of Section 4.3.3.

3.3   *Quasistatic behavior of devices*

3.3.1   Electroquasistatic behavior of devices

The voltages and currents associated with all interesting devices sometimes vary. If the wavelength $\lambda = c/f$ associated with these variations is much larger than the device size D, no significant wave behavior can occur. The device behavior can then be characterized as electroquasistatic if the device stores primarily electric energy, and magnetoquasistatic if the device stores primarily magnetic energy. *Electroquasistatics* involves the behavior of electric fields plus the first-order magnetic consequences of their variations. The electroquasistatic approximation includes the magnetic field $\overline{H}$ generated by the varying dominant electric field (Ampere's law), where:

$$\nabla \times \overline{H} = \sigma\overline{E} + \frac{\partial \overline{D}}{\partial t} \qquad (3.3.1)$$

The quasistatic approximation neglects the second-order electric field contributions from the time derivative of the resulting $\overline{H}$ in Faraday's law: $\nabla \times \overline{E} = -\mu_o \partial \overline{H}/\partial t \cong 0$.

One simple geometry involving slowly varying electric fields is a *capacitor* charged to voltage V(t), as illustrated in Figure 3.3.1. It consists of two circular parallel conducting plates of diameter D and area A that are separated in vacuum by the distance d << D. Boundary conditions require $\overline{E}$ to be perpendicular to the plates, where E(t) = V(t)/d, and the surface charge density is given by (2.6.15):

$$\overline{E} \bullet \hat{n} = \rho_s / \varepsilon_o = V/d \qquad (3.3.2)$$

$$\rho_s = \varepsilon_o V/d \quad [C\,m^{-2}] \qquad (3.3.3)$$
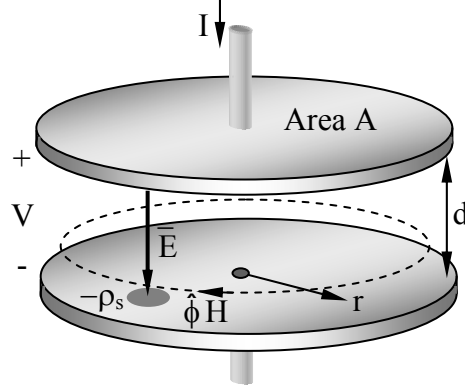
Figure 3.3.1   Quasistatic electric and magnetic fields in a circular capacitor.

Since the voltage across the plates is the same everywhere, so are $\overline{E}$ and ρ_s, and therefore the total charge is:

$$Q(t) \cong \rho_s A \cong (\varepsilon_o A/d)V = CV(t) \tag{3.3.4}$$

where $C \cong \varepsilon_o A/d$ is the capacitance, as shown earlier (3.1.10). The same surface charge density $\rho_s(t)$ can also be found by evaluating first the magnetic field $\overline{H}(r,t)$ produced by the slowly varying (quasistatic) electric field $\overline{E}(t)$, and then the surface current $\overline{J}_s(r,t)$ associated with $\overline{H}(r,t)$; charge conservation then links $\overline{J}_s(r,t)$ to $\rho_s(t)$.

Ampere's law requires a non-zero magnetic field between the plates where $\overline{J} = 0$:

$$\oint_C \overline{H} \bullet d\overline{s} = \varepsilon_o \iint_{A'} \left(\partial \overline{E}/\partial t \right) \bullet d\overline{a} \tag{3.3.5}$$

Symmetry of geometry and excitation requires that $\overline{H}$ between the plates be in the $\hat{\phi}$ direction and a function only of radius r, so (3.3.5) becomes:

$$2\pi r\, H(r) = \varepsilon_o \pi r^2\, dE/dt = (\varepsilon_o \pi r^2/d)\, dV/dt \tag{3.3.6}$$

$$H(r) = (\varepsilon_o r/2d)\, dV/dt \tag{3.3.7}$$

If V(t) and the magnetic field H are varying so slowly that the electric field given by Faraday's law for H(r) is much less than the original electric field, then that incremental electric field can be neglected, which is the essence of the electroquasistatic approximation. If it cannot be neglected, then the resulting solution becomes more wavelike, as discussed in later sections.

The boundary condition $\hat{n} \times \overline{H} = \overline{J}_s$ (2.6.17) then yields the associated surface current $\overline{J}_s(r)$ flowing on the interior surface of the top plate:

$$\overline{J}_s(r) = \hat{r}\, (\varepsilon_o r/2d)\, dV/dt = \hat{r}\, J_{sr} \tag{3.3.8}$$

This in turn is related to the surface charge density ρ_s by conservation of charge (2.1.19), where the del operator is in cylindrical coordinates:

$$\nabla \bullet \overline{J}_s = -\partial \rho_s / \partial t = - r^{-1} \partial (r\, J_{sr}) / \partial r \qquad\qquad (3.3.9)$$

Substituting J_{sr} from (3.3.8) into the right-hand side of (3.3.9) yields:

$$\partial \rho_s / \partial t = (\varepsilon_o / d)\, dV/dt \qquad\qquad (3.3.10)$$

Multiplying both sides of (3.3.10) by the plate area A and integrating over time then yields Q(t) = CV(t), which is the same as (3.3.4). Thus we could conclude that variations in V(t) will produce magnetic fields between capacitor plates by virtue of Ampere's law and the values of either $\partial \overline{D} / \partial t$ between the capacitor plates or $\overline{J}_s$ within the plates. These two approaches to finding $\overline{H}$ (using $\partial \overline{D} / \partial t$ or $\overline{J}_s$) yield the same result because of the self-consistency of Maxwell's equations.

Because the curl of $\overline{H}$ in Ampere's law equals the sum of current density $\overline{J}$ and $\partial \overline{D} / \partial t$, the derivative $\partial \overline{D} / \partial t$ is often called the *displacement current* density because the units are the same, A/m^2. For the capacitor of Figure 3.3.1 the curl of $\overline{H}$ near the feed wires is associated only with $\overline{J}$ (or I), whereas between the capacitor plates the curl of $\overline{H}$ is associated only with displacement current.

Section 3.3.4 treats the electroquasistatic behavior of electric fields within conductors and relaxation phenomena.

3.3.2 Magnetoquasistatic behavior of devices

All currents produce magnetic fields that in turn generate electric fields if those magnetic fields vary. *Magnetoquasistatics* characterizes the behavior of such slowly varying fields while neglecting the second-order magnetic fields generated by $\partial \overline{D} / \partial t$ in Ampere's law, (2.1.6):

$$\nabla \times \overline{H} = \overline{J} + \partial \overline{D} / \partial t \cong \overline{J} \qquad \text{(quasistatic Ampere's law)} \quad (3.3.11)$$

The associated electric field $\overline{E}$ can then be found from Faraday's law:

$$\nabla \times \overline{E} = -\partial \overline{B} / \partial t \qquad \text{(Faraday's law)} \quad (3.3.12)$$

Section 3.2.1 treated an example for which the dominant effect of the quasistatic magnetic field in a current loop is voltage induced via Faraday's law, while the example of a short wire follows; both are inductors. Section 3.3.4 treats the magnetoquasistatic example of magnetic diffusion, which is dominated by currents induced by the first-order induced voltages, and resulting modification of the original magnetic field by those induced currents. In every quasistatic problem wave effects can be neglected because the associated wavelength $\lambda \gg D$, where D is the maximum device dimension.

We can roughly estimate the inductance of a short wire segment by modeling it as a perfectly conducting cylinder of radius r_o and length D carrying a current i(t), as illustrated in Figure 3.3.2. An exact computation would normally be done using computer tools designed for such tasks because analytic solutions are practical only for extremely simple geometries. In this analysis we neglect any contributions to $\overline{H}$ from currents in nearby conductors, which requires those nearby conductors to have much larger diameters or be far away.   We also make the quasistatic assumption $\lambda >> D$.
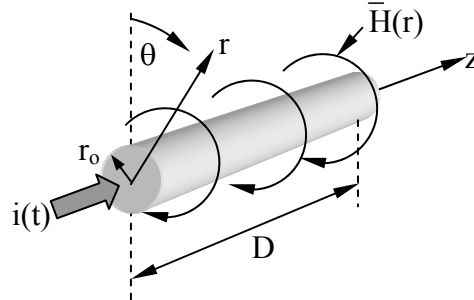
Figure 3.3.2   Inductance of an isolated wire segment.

We know from (3.2.23) that the inductance of any device can be expressed in terms of the magnetic energy stored as a function of its current i:

$$L = 2w_m / i^2 \quad [H]$$

(3.3.13)

Therefore to estimate L we first estimate $\overline{H}$ and w_m. If the cylinder were infinitely long then $\overline{H} \cong \hat{\theta} H(r)$ must obey Ampere's law and exhibit the same cylindrical symmetry, as suggested in the figure. Therefore:

$$\oint_C \overline{H} \bullet d\overline{s} = 2\pi r \, H(r) = i(t)$$

(3.3.14)

and $H(r) \cong i/2\pi r$. Therefore the instantaneous magnetic energy density is:

$$\langle W_m \rangle = \frac{1}{2}\mu_o H^2(r) = \frac{1}{2}\mu_o (i/2\pi r)^2 \quad [J/m^3]$$

(3.3.15)

To find the total average stored magnetic energy we must integrate over volume. Laterally we can neglect fringing fields and simply integrate over the length D.   Integration with respect to radius will produce a logarithmic answer that becomes infinite if the maximum radius is infinite. A plausible outer limit for r is ~D because the Biot-Savart law (1.4.6) says fields decrease as r^2 from their source if that source is local; the transition from slow cylindrical field decay as r^{-1} to decay as r^{-2} occurs at distances r comparable to the largest dimension of the source: $r \cong D$. With these approximations we find:

$$w_m \cong \int_0^D dz \int_{r_o}^D \langle W_m \rangle 2\pi r \, dr \cong D \int_{r_o}^D \frac{1}{2} \mu_o \left(\frac{i}{2\pi r} \right)^2 2\pi r \, dr$$

$$= \left(\mu_o D i^2 / 4\pi \right) \ln r \Big|_{r_o}^D = \left(\mu_o D i^2 / 4\pi \right) \ln \left(D/r_o \right) \quad [J] \tag{3.3.16}$$

Using (3.3.13) we find the inductance L for this wire segment is:

$$L \cong \left(\mu_o D / 2\pi \right) \ln \left(D/r_o \right) \quad [Hy] \tag{3.3.17}$$

where the units "Henries" are abbreviated here as "Hy". Note that superposition does not apply here because we are integrating energy densities, which are squares of field strengths, and the outer limit of the integral (3.3.16) is wire length D, so longer wires have slightly more inductance than the sum of shorter elements into which they might be subdivided.

3.3.3 Equivalent circuits for simple devices

Section 3.1 showed how the parallel plate resistor of Figure 3.1.1 would exhibit resistance R = d/σA ohms and capacitance C = εA/d farads, connected in parallel. The currents in the same device also generate magnetic fields and add inductance.

Referring to Figure 3.1.1 of the original parallel plate resistor, most of the inductance will arise from the wires, since they have a very small radius r_o compared to that of the plates. This inductance L will be in series with the RC portions of the device because their two voltage drops add. The R and C components are in parallel because the total current through the device is the sum of the conduction current and the displacement current, and the voltages driving these two currents are the same, i.e., the voltage between the parallel plates. The corresponding first-order equivalent circuit is illustrated in Figure 3.3.3.
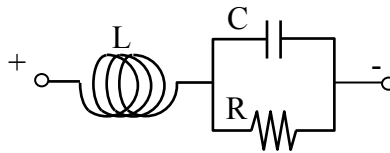
Figure 3.3.3   Equivalent RLC circuit of a parallel-plate capacitor.

Examination of Figure 3.3.3 suggests that at very low frequencies the resistance R dominates because, relative to the resistor, the inductor and capacitor become approximate short and open circuits, respectively. At the highest frequencies the inductor dominates. As f increases from zero beyond where R dominates, either the RL or the RC circuit first dominates, depending on whether C shorts the resistance R at lower frequencies than when L open-circuits R; that is, RC dominates first when $R > \sqrt{L/C}$. At still higher frequencies the LC circuit dominates, followed by L alone. For certain combinations of R, L, and C, some transitions can merge.

Even this model for a resistor is too simple; for example, the wires also exhibit resistance and there is magnetic energy stored between the end plates because $\partial D/dt \neq 0$ there. Since such parasitic effects typically become important only at frequencies above the frequency range specified for the device, they are normally neglected. Even more complex behavior can result if the frequencies are so high that the device dimensions exceed $\sim\lambda/8$, as discussed later in Section 7.1. Similar considerations apply to every resistor, capacitor, inductor, or transformer manufactured. Components and circuits designed for very high frequencies minimize unwanted *parasitic capacitance* and *parasitic inductance* by their very small size and proper choice of materials and geometry. It is common for circuit designers using components or wires near their design limits to model them with simple lumped-element equivalent circuits like that of Figure 3.3.3, which include the dominant parasitic effects. The form of these circuits obviously depends on the detailed structure of the modeled device; for example, R and C might be in series.

Example 3.3A

What are the approximate values L and C for the 100-Ω resistor designed in Example 3.1A if $\varepsilon = 4\varepsilon_o$, and what are the three critical frequencies $(RC)^{-1}$, R/L, and $(LC)^{-0.5}$?

Solution: The solution to 3.1A said the conducting caps of the resistor have area $A = \pi r^2 = \pi(2.5\times10^{-4})^2$, and the length of the dielectric d is 1 mm. The permittivity $\varepsilon = 4\varepsilon_o$, so the capacitance (3.1.10) is $C = \varepsilon A/d = 4\times8.85\times10^{-12}\times\pi(2.5\times10^{-4})^2/10^{-3} \cong 7\times10^{-15}$ farads. The inductance L of this device would probably be dominated by that of the connecting wires because their diameters would be smaller and their length longer. Assume the wire length is $D = 4d = 4\times10^{-3}$, and its radius r is 10^{-4}. Then (3.3.17) yields $L \cong (\mu_o D/16\pi)\ln(D/r) = (1.26\times10^{-6} \times 4\times10^{-3}/16\pi)\ln(40) = 3.7\times10^{-10}$ [Hy]. The critical frequencies R/L, $(RC)^{-1}$, and $(LC)^{-0.5}$ are 2.7×10^{11}, 6.2×10^{11}, and 1.4×10^{12} [r s^{-1}], respectively, so the maximum frequency for which reasonably pure resistance is obtained is $\sim$10 GHz ($\sim R/2\pi L4$).

3.4   *General circuits and solution methods*

3.4.1   Kirchoff's laws

Circuits are generally composed of *lumped elements* or "*branches*" connected at *nodes* to form two- or three-dimensional structures, as suggested in Figure 3.4.1. They can be characterized by the voltages v_i at each node or across each branch, or by the currents i_j flowing in each branch or in a set of current loops. To determine the behavior of such circuits we develop simultaneous linear equations that must be satisfied by the unknown voltages and currents. Kirchoff's laws generally provide these equations.

Although circuit analysis is often based in part on Kirchoff's laws, these laws are imperfect due to electromagnetic effects. For example, *Kirchoff's voltage law* (KVL) says that the voltage drops v_i associated with each lumped element around any loop must sum to zero, i.e.:

$$\sum_i v_i = 0 \qquad\qquad \text{(Kirchoff's voltage law [KVL])} \qquad (3.4.1)$$
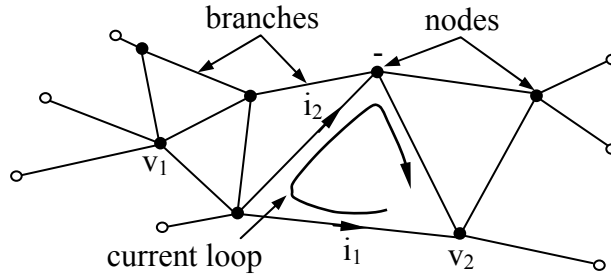
Figure 3.4.1   Circuit with branches and current loops.

which can be derived from the integral form of Faraday's law:

$$\oint_C \overline{E} \bullet d\overline{s} = -(\partial/\partial t) \oiint_A \overline{B} \bullet d\overline{a} \tag{3.4.2}$$

This integral of $\overline{E} \bullet d\overline{s}$ across any branch yields the voltage across that branch. Therefore the sum of branch voltages around any closed contour is zero if the net magnetic flux through that contour is constant; this is the basic assumption of KVL.

KVL is clearly valid for any static circuit. However, any branch carrying time varying current will contribute time varying magnetic flux and therefore voltage to all adjacent loops plus others nearby. These voltage contributions are typically negligible because the currents and loop areas are small relative to the wavelengths of interest ($\lambda = c/f$) and the KVL approximation then applies. A standard approach to analyzing circuits that violate KVL is to determine the magnetic energy or inductance associated with any extraneous magnetic fields, and to model their effects in the circuit with a lumped *parasitic inductance* in each affected current loop.

The companion relation to KVL is *Kirchoff's current law* (KCL), which says that the sum of the currents i_j flowing into any node is zero:

$$\sum_j i_j = 0 \qquad\qquad \text{(Kirchoff's current law)} \qquad (3.4.3)$$

This follows from conservation of charge (2.4.19) when no charge storage on the nodes is allowed:

$$(\partial/\partial t) \iiint_V \rho \ dv = -\oiint_A \overline{J} \bullet d\overline{a} \qquad\qquad \text{(conservation of charge)} \qquad (3.4.4)$$

If no charge can be stored on the volume V of a node, then $(\partial/\partial t) \iiint_V \rho \ dv = 0$, and there can be no net current into that node.

For static problems, KCL is exact. However, the physical nodes and the wires connecting those nodes to lumped elements typically exhibit varying voltages and $\overline{D}$, and therefore have

- 89 -

capacitance and the ability to store charge, violating KCL. If the frequency is sufficiently high that such *parasitic capacitance* at any node becomes important, that parasitic capacitance can be modeled as an additional lumped element attached to that node.

3.4.2 Solving circuit problems

To determine the behavior of any given linear lumped element circuit a set of simultaneous equations must be solved, where the number of equations must equal or exceed the number of unknowns. The unknowns are generally the voltages and currents on each branch; if there are b branches there are 2b unknowns.

Figure 3.4.2(a) illustrates a simple circuit with $b = 12$ branches, $p = 6$ loops, and $n = 7$ nodes. A set of *loop currents* uniquely characterizes all currents if each loop circles only one "hole" in the topology and if no additional loops are added once every branch in the circuit is incorporated in at least one loop. Although other definitions for the loop currents can adequately characterize all *branch currents*, they are not explored here. Figure 3.4.2(b) illustrates a bridge circuit with $b = 6$, $p = 3$, and $n = 4$.
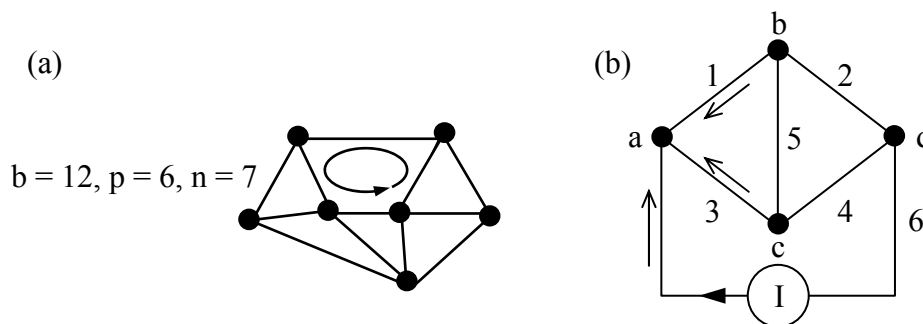
Figure 3.4.2   12-branch circuit and bridge circuit.

The simplest possible circuit has one node and one branch, as illustrated in Figure 3.4.3(a).
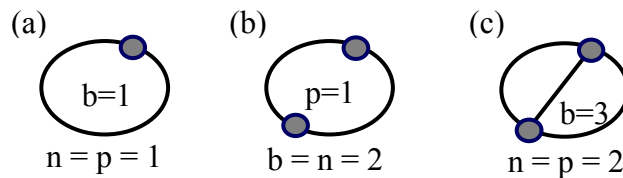
Figure 3.4.3   Simple circuit topologies;
n, p, and b are the numbers of nodes, loops, and branches, respectively.

It is easy to see from the figure that the number b of branches in a circuit is:

$$b = n + p - 1 \tag{3.4.5}$$

As we add either nodes or branches to the illustrated circuit in any sequence and with any placement, Equation (3.4.5) is always obeyed. If we add voltage or current sources to the circuit, they too become branches.

The voltage and current for each branch are initially unknown and therefore any circuit has 2b unknowns. The number of equations is also $b + (n - 1) + p = 2b$, where the first b in this expression corresponds to the equations relating voltage to current in each branch, n-1 is the number of independent KCL equations, and p is the number of loops and KVL equations; (3.4.5) says $(n - 1) + p = b$. Therefore, since the numbers of unknowns and linear equations match, we may solve them. The equations are linear because Maxwell's equations are linear for RLC circuits.

Often circuits are so complex that it is convenient for purposes of analysis to replace large sections of them with either a two-terminal *Thevenin equivalent circuit* or *Norton equivalent circuit*. This can be done only when that circuit is incrementally linear with respect to voltages imposed at its terminals. Thevenin equivalent circuits consist of a voltage source $V_{Th}(t)$ in series with a passive linear circuit characterized by its frequency-dependent impedance $\underline{Z}(\omega) = R + jX$, while Norton equivalent circuits consist of a current source $I_{No}(t)$ in parallel with an impedance $\underline{Z}(\omega)$.

An important example of the utility of equivalent circuits is the problem of designing a *matched load* $\underline{Z}_L(\omega) = R_L(\omega) + jX_L(\omega)$ that accepts the maximum amount of power available from a linear source circuit, and reflects none. The solution is simply to design the load so its impedance $\underline{Z}_L(\omega)$ is the complex conjugate of the source impedance: $\underline{Z}_L(\omega) = \underline{Z}^*(\omega)$. For both Thevenin and Norton equivalent sources the reactance of the matched load cancels that of the source $[X_L(\omega) = - X(\omega)]$ and the two resistive parts are set equal, $R = R_L$.

One proof that a matched load maximizes power transfer consists of computing the time-average power P_d dissipated in the load as a function of its impedance, equating to zero its derivative $dP_d/d\omega$, and solving the resulting complex equation for R_L and X_L. We exclude the possibility of negative resistances here unless those of the load and source have the same sign; otherwise the transferred power can be infinite if $R_L = -R$.

Example 3.4A
The *bridge circuit* of Figure 3.4.2(b) has five branches connecting four nodes in every possible way except one. Assume both parallel branches have 0.1-ohm and 0.2-ohm resistors in series, but in reverse order so that $R_1 = R_4 = 0.1$, and $R_2 = R_3 = 0.2$. What is the resistance R of the bridge circuit between nodes a and d if $R_5 = 0$? What is R if $R_5 = \infty$? What is R if R_5 is 0.5 ohms?

Solution: When $R_5 = 0$ then the node voltages $v_b = v_c$, so R_1 and R_3 are connected in parallel and have the equivalent resistance $R_{13//}$. Kirchoff's current law "KCL" (3.4.3) says the current flowing into node "a" is $I = (v_a - v_b)(R_1^{-1} + R_3^{-1})$. If $V_{ab} \equiv (v_a - v_b)$, then $V_{ab} = IR_{13//}$ and $R_{13//} = (R_1^{-1} + R_3^{-1})^{-1} = (10+5)^{-1} = 0.067\Omega = R_{24//}$. These two circuits are in series so their resistances add: $R = R_{13//} + R_{24//} \cong 0.133$ ohms. When $R_5 = \infty$, R_1

and R_2 are in series with a total resistance R_{12s} of $0.1 + 0.2 = 0.3\Omega = R_{34s}$. These two resistances, R_{12s} and R_{34s} are in parallel, so $R = (R_{12s}^{-1} + R_{34s}^{-1})^{-1} = 0.15\Omega$. When R_5 is finite, then simultaneous equations must be solved. For example, the currents flowing into each of nodes a, b, and c sum to zero, yielding three simultaneous equations that can be solved for the vector $\overline{V} = [v_a, v_b, v_c]$; we define $v_d = 0$. Thus $(v_a - v_b)/R_1 + (v_a - v_c)/R_3 = I = v_a(R_1^{-1} + R_3^{-1}) - v_b R_1^{-1} - v_c R_3^{-1} = 15v_a - 10v_b - 5v_c$. KCL for nodes b and c similarly yield: $-10v_a + 17v_b - 2v_c = 0$, and $-5v_a - 2v_b + 17v_c = 0$. If we define the current vector $\overline{I} = [I, 0, 0]$, then these three equations can be written as a matrix equation:

$$\overline{\overline{G}}\,\overline{v} = \overline{I}, \text{ where } \overline{\overline{G}} = \begin{bmatrix} 15 & -10 & -5 \\ -10 & 17 & -2 \\ -5 & -2 & 17 \end{bmatrix}.$$

Since the desired circuit resistance between nodes a and d is $R = v_a/I$, we need only solve for v_a in terms of I, which follows from $\overline{v} = \overline{\overline{G}}^{-1}\overline{I}$, provided the conductance matrix $\overline{\overline{G}}$ is not singular (here it is not). Thus $R = 0.146\Omega$, which is intermediate between the first two solutions, as it should be.

3.5   *Two-element circuits and RLC resonators*

3.5.1   Two-element circuits and uncoupled RLC resonators

RLC resonators typically consist of a resistor R, inductor L, and capacitor C connected in series or parallel, as illustrated in Figure 3.5.1. RLC resonators are of interest because they behave much like other electromagnetic systems that store both electric and magnetic energy, which slowly dissipates due to resistive losses. First we shall find and solve the differential equations that characterize RLC resonators and their simpler sub-systems: RC, RL, and LC circuits. This will lead to definitions of resonant frequency ω_o and Q, which will then be related in Section 3.5.2 to the frequency response of RLC resonators that are coupled to circuits.
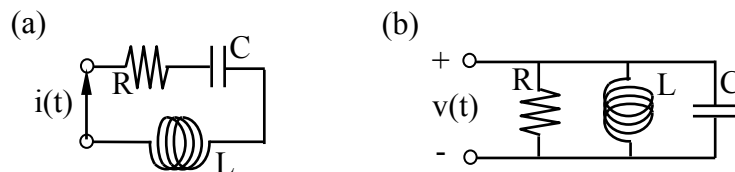
Figure 3.5.1   Series and parallel RLC resonators.

The differential equations that govern the voltages across R's, L's, and C's are, respectively:

$$v_R = iR \tag{3.5.1}$$

$$v_L = L\,di/dt \tag{3.5.2}$$

$$v_C = (1/C)\int i\ dt \tag{3.5.3}$$

Kirchoff's voltage law applied to the series RLC circuit of Figure 3.5.1(a) says that the sum of the voltages (3.5.1), (3.5.2), and (3.5.3) is zero:

$$d^2i/dt^2 + (R/L)\,di/dt + (1/LC)i = 0 \tag{3.5.4}$$

where we have divided by L and differentiated to simplify the equation. Before solving it, it is useful to solve simpler versions for RC, RL, and LC circuits, where we ignore one of the three elements.

In the RC limit where L = 0 we add (3.5.1) and (3.5.3) to yield the differential equation:

$$di/dt + (1/RC)i = 0 \tag{3.5.5}$$

This says that i(t) can be any function with the property that the first derivative is the same as the original signal, times a constant. This property is restricted to exponentials and their sums, such as sines and cosines. Let's represent i(t) by $I_o e^{st}$, where:

$$i(t) = R_e\left\{I_o e^{st}\right\} \tag{3.5.6}$$

where the *complex frequency* s is:

$$s \equiv \alpha + j\omega \tag{3.5.7}$$

We can substitute (3.5.6) into (3.5.5) to yield:

$$Re\left\{\left[s + (1/RC)\right]I_o e^{st}\right\} = 0 \tag{3.5.8}$$

Since e^{st} is not always zero, to satisfy (3.5.8) it follows that s = - 1/RC and:

$$i(t) = I_o e^{-(1/RC)t} = I_o e^{-t/\tau} \qquad \text{(RC current response)} \tag{3.5.9}$$

where τ equals RC seconds and is the *RC time constant*. I_o is chosen to satisfy initial conditions, which were not given here.

A simple example illustrates how initial conditions can be incorporated in the solution. We simply need as many equations for t = 0 as there are unknown variables. In the present case we need one equation to determine I_o. Suppose the RC circuit [of Figure 3.5.1(a) with L = 0] was at

rest at t = 0, but the capacitor was charged to V_o volts. Then we know that the initial current I_o at t = 0 must be V_o/R.

In the RL limit where C = ∞ we add (3.5.1) and (3.5.2) to yield di/dt + (R/L)i = 0, which has the same form of solution (3.5.6), so that s = -R/L and:

$$i(t) = I_o e^{-(R/L)t} = I_o e^{-t/\tau}$$
(RL current response)   (3.5.10)

where the *RL time constant* τ is L/R seconds.

In the LC limit where R = 0 we add (3.5.2) and (3.5.3) to yield:

$$d^2i/dt^2 + (1/LC)i = 0$$
(3.5.11)

Its solution also has the form (3.5.6). Because i(t) is real and $e^{j\omega t}$ is complex, it is easier to assume sinusoidal solutions, where the phase ϕ and magnitude I_o would be determined by initial conditions. This form of the solution would be:

$$i(t) = I_o \cos(\omega_o t + \phi)$$
(LC current response)   (3.5.12)

where $\omega_o = 2\pi f_o$ is found by substituting (3.5.12) into (3.5.11) to yield $[\omega_o^2 - (LC)^{-1}]i(t) = 0$, so:

$$\omega_o = \frac{1}{\sqrt{LC}} \left[\text{radians s}^{-1} \right]$$
(*LC resonant frequency*)   (3.5.13)

We could alternatively express this solution (3.5.12) as the sum of two exponentials using the identity $\cos \omega t \equiv (e^{j\omega t} + e^{-j\omega t})/2$.

RLC circuits exhibit both oscillatory resonance and exponential decay. If we substitute the generic solution $\underline{I}_o e^{st}$ (3.5.6) into the RLC differential equation (3.5.4) for the *series RLC resonator* of Figure 3.5.1(a) we obtain:

$$(s^2 + sR/L + 1/LC)\underline{I}_o e^{st} = (s - s_1)(s - s_2)\underline{I}_o e^{st} = 0$$
(3.5.14)

The RLC resonant frequencies s_1 and s_2 are solutions to (3.5.14) and can be found by solving this *quadratic equation*[9] to yield:

$$s_i = -R/2L \pm j\left[(1/LC) - (R/2L)^2\right]^{0.5}$$
(series *RLC resonant frequencies*)   (3.5.15)

When R = 0 this reduces to the LC resonant frequency solution (3.5.13).

---

[9] A quadratic equation in x has the form $ax^2 + bx + c = 0$ and the solution $x = (-b \pm [b^2 - 4ac]^{0.5})/2a$.

The generic solution $i(t) = \underline{I}_o' e^{st}$ is complex, where $\underline{I}_o' \equiv I_o e^{j\phi}$:

$$i(t) = R_e \left\{ \underline{I}_o' e^{s_1 t} \right\} = R_e \left\{ I_o e^{j\phi} e^{-(R/2L)t} e^{j\omega t} \right\} = I_o e^{-(R/2L)t} \cos(\omega t + \phi) \qquad (3.5.16)$$

where $\omega = [(LC)^{-1} + (R/2L)^2]^{0.5} \cong (LC)^{-0.5}$. I_o and ϕ can be found from the initial conditions, which are the initial current through L and the initial voltage across C, corresponding to the initial energy storage terms. If we choose the time origin so that the phase $\phi = 0$, the instantaneous magnetic energy stored in the inductor (3.2.23) is:

$$w_m(t) = Li^2/2 = \left(LI_o^2/2 \right) e^{-Rt/L} \cos^2 \omega t = \left(LI_o^2/4 \right) e^{-Rt/L} (1 + \cos 2\omega t) \qquad (3.5.17)$$

Because $w_m = 0$ twice per cycle and energy is conserved, the peak electric energy $w_e(t)$ stored in the capacitor must be intermediate between the peak magnetic energies stored in the inductor $(e^{Rt/L} LI_o^2/2)$ during the preceding and following cycles. Also, since $dv_C/dt = i/C$, the cosine variations of i(t) produce a sinusoidal variation in the voltage $v_C(t)$ across the capacitor. Together these two facts yield: $w_e(t) \cong (LI_o^2/2)e^{-Rt/L} \sin^2 \omega t$. If we define V_o as the maximum initial voltage corresponding to the maximum initial current I_o, and recall the expression (3.1.16) for $w_e(t)$, we find:

$$w_e(t) = Cv^2/2 \cong \left(CV_o^2/2 \right) e^{-Rt/L} \sin^2 \omega t = \left(CV_o^2/4 \right) e^{-Rt/L} (1 - \cos 2\omega t) \qquad (3.5.18)$$

Comparison of (3.5.17) and (3.5.18) in combination with conservation of energy yields:

$$V_o \cong \left(L/C \right)^{0.5} I_o \qquad (3.5.19)$$

Figure 3.5.2 illustrates how the current and energy storage decays exponentially with time while undergoing conversion between electric and magnetic energy storage at 2ω radians s^{-1}; the time constant for current and voltage is $\tau = 2L/R$ seconds, and that for energy is L/R.

One useful way to characterize a resonance is by the dimensionless quantity Q, which is the number of radians required before the total energy w_T decays to 1/e of its original value, as illustrated in Figure 3.5.2(b). That is:

$$w_T = w_{To} e^{-2\alpha t} = w_{To} e^{-\omega t/Q} \ [J] \qquad (3.5.20)$$

The decay rate α for current and voltage is therefore simply related to Q:

$$\alpha = \omega/2Q \qquad (3.5.21)$$
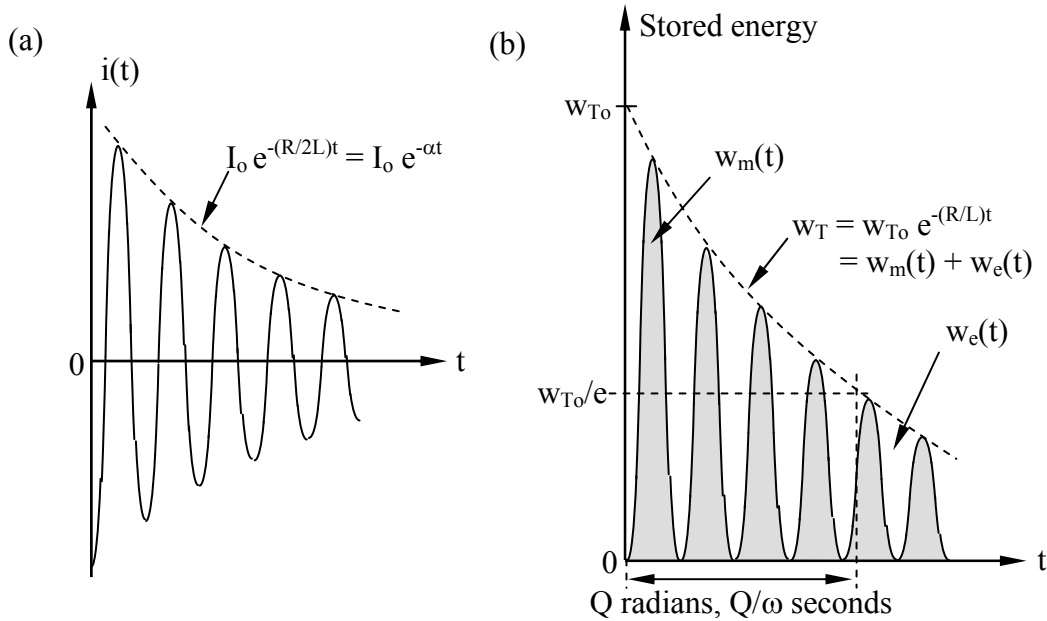
Figure 3.5.2   Time variation of current and energy storage in RLC circuits.

If we find the power dissipated P_d [W] by differentiating total energy w_T with respect to time using (3.5.20), we can then derive a common alternative definition for Q:

$$P_d = -dw_T/dt = (\omega/Q)w_T \tag{3.5.22}$$

$$Q = \omega w_T/P_d \qquad \text{(one definition of Q)} \tag{3.5.23}$$

For the series RLC resonator $\alpha = R/2L$ and $\omega \cong (LC)^{-0.5}$, so (3.5.21) yields:

$$Q = \omega/2\alpha = \omega L/R \cong (L/C)^{0.5}\big/R \qquad \text{(Q of series RLC resonator)} \tag{3.5.24}$$

Figure 3.5.1(b) illustrates a *parallel RLC resonator*. KCL says that the sum of the currents into any node is zero, so:

$$C \ dv/dt + v/R + (1/L)\int v \ dt = 0 \tag{3.5.25}$$

$$d^2v/dt^2 + (1/RC)dv/dt + (1/LC)v = 0 \tag{3.5.26}$$

If $v = V_o e^{st}$, then:

$$\left[s^2 + (1/RC)s + (L/C) \right] = 0 \tag{3.5.27}$$

$$s = -(1/2RC) \pm j\left[(1/LC) - (1/2RC)^2\right]^{0.5} \qquad \text{(parallel RLC resonance)} \qquad (3.5.28)$$

Analogous to (3.5.16) we find:

$$v(t) = R_e\left\{\underline{V}_o{}'e^{s_1 t}\right\} = V_o e^{-(1/2RC)t} \cos(\omega t + \phi) \qquad (3.5.29)$$

where $\underline{V}_o{}' = V_o e^{j\phi}$. It follows that for a parallel RLC resonator:

$$\omega = \left[(LC)^{-1} - (2RC)^{-2}\right]^{0.5} \cong (LC)^{-0.5} \qquad (3.5.30)$$

$$Q = \omega/2\alpha = \omega RC = R\left(C/L\right)^{0.5} \qquad \text{(Q of parallel RLC resonator)} \qquad (3.5.31)$$

Example 3.5A
What values of L and C would give a parallel resonator at 1 MHz a Q of 100 if $R = 10^6/2\pi$?

Solution: $LC = 1/\omega_o{}^2 = 1/(2\pi 10^6)^2$, and $Q = 100 = \omega RC = 2\pi 10^6(10^6/2\pi)C$ so $C = 10^{-10}$ [F] and
$L = 1/\omega_o{}^2 c \cong 2.5 \times 10^{-4}$ [Hy].

3.5.2   Coupled RLC resonators

RLC resonators are usually coupled to an environment that can be represented by either its Thevenin or Norton equivalent circuit, as illustrated in Figure 3.5.3(a) and (b), respectively, for purely resistive circuits.

(a) Thevenin equivalent source   (b) Norton equivalent source
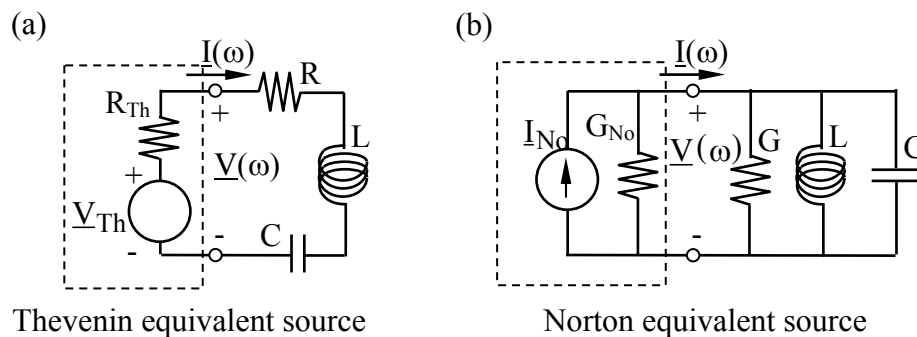
Figure 3.5.3   Series and parallel RLC resonators
driven by Thevenin and Norton equivalent circuits.

A Thevenin equivalent consists of a voltage source V_{Th} in series with an impedance $\underline{Z}_{Th} = R_{Th} + jX_{Th}$, while a Norton equivalent circuit consists of a current source I_{No} in parallel with an admittance $\underline{Y}_{No} = G_{No} + jU_{No}$.   The Thevenin equivalent of a resistive Norton

equivalent circuit has open-circuit voltage $V_{Th} = I_{No}/G_{No}$, and $R_{Th} = 1/G_{No}$; that is, their open-circuit voltages, short-circuit currents, and impedances are the same. No single-frequency electrical experiment performed at the terminals can distinguish ideal linear circuits from their Thevenin or Norton equivalents.

An important characteristic of a resonator is the frequency dependence of its power dissipation. If $R_{Th} = 0$, the series RLC resonator of Figure 3.5.3(a) dissipates:

$$P_d = R|\underline{I}|^2/2 \quad [W] \tag{3.5.32}$$

$$P_d = \left[R|\underline{V}_{Th}|^2/2\right]\Big/\left|R + Ls + C^{-1}s^{-1}\right|^2 = \left[R|\underline{V}_{Th}|^2/2\right]|s/L|^2\Big/\left|(s-s_1)(s-s_2)\right|^2 \tag{3.5.33}$$

where s_1 and s_2 are given by (3.5.15):

$$s_i = -R/2L \pm j\left[(1/LC) - (R/2L)^2\right]^{0.5} = -\alpha \pm j\omega'_o \qquad \text{(series RLC resonances)} \tag{3.5.34}$$

The maximum value of P_d is achieved when $\omega \cong \omega'_o$:

$$P_{d\,max} = |\underline{V}_{Th}|^2/2R \tag{3.5.35}$$

This simple expression is expected since the reactive impedances of L and C cancel at ω_o, leaving only R.

If $(1/LC) \gg (R/2L)$ so that $\omega_o \cong \omega'_o$, then as $\omega - \omega_o$ increases from zero to α, $|s - s_1| = |j\omega_o - (j\omega_o + \alpha)|$ increases from α to $\sqrt{2}\,\alpha$. This departure from resonance approximately doubles the denominator of (3.5.33) and halves P_d. As ω departs still further from ω_o and resonance, P_d eventually approaches zero because the impedances of L and C approach infinity at infinite and zero frequency, respectively. The total frequency response $P_d(f)$ of this series RLC resonator is suggested in Figure 3.5.4. The *resonator bandwidth* or *half-power bandwidth* $\Delta\omega$ is said to be the difference between the two half-power frequencies, or $\Delta\omega \cong 2\alpha = R/L$ for this series circuit. $\Delta\omega$ is simply related to ω_o and Q for both series and parallel resonances, as follows from (3.5.21):

$$Q = \omega_o/2\alpha = \omega_o/\Delta\omega \qquad\qquad \text{(Q versus bandwidth)} \tag{3.5.36}$$

Parallel RLC resonators behave similarly except that:

$$s_i = -G/2L \pm j\left[(1/LC) - (G/2L)^2\right]^{0.5} = -\alpha \pm j\omega'_o \quad \text{(parallel RLC resonances)} \tag{3.5.37}$$

where R, L, and C in (3.5.34) have been replaced by their duals G, C, and L, respectively.
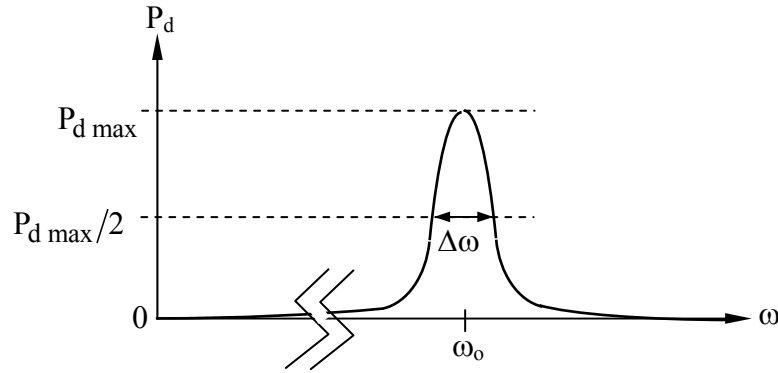
Figure 3.5.4   RLC power dissipation near resonance.

Resonators reduce to their resistors at resonance because the impedance of the LC portion approaches zero or infinity for series or parallel resonators, respectively. At resonance P_d is maximized when the source R_s and load R resistances match, as is easily shown by setting the derivative $dP_d/dR = 0$ and solving for R. In this case we say the resonator is *critically matched* to its source, for all available power is then transferred to the load at resonance.

This critically matched condition can also be related to the Q's of a coupled resonator with zero Thevenin voltage applied from outside, where we define *internal Q* (or Q_I) as corresponding to power dissipated internally in the resonator, *external Q* (or Q_E) as corresponding to power dissipated externally in the source resistance, and *loaded Q* (or Q_L) as corresponding to the total power dissipated both internally (P_{DI}) and externally (P_{DE}). That is, following (3.5.23):

$$Q_I \equiv \omega w_T / P_{DI} \qquad \text{(internal Q)} \qquad (3.5.38)$$

$$Q_E \equiv \omega w_T / P_{DE} \qquad \text{(external Q)} \qquad (3.5.39)$$

$$Q_L \equiv \omega w_T / (P_{DI} + P_{DE}) \qquad \text{(loaded Q)} \qquad (3.5.40)$$

Therefore these Q's are simply related:

$$Q_L^{-1} = Q_I^{-1} + Q_E^{-1} \qquad (3.5.41)$$

It is Q_L that corresponds to $\Delta\omega$ for coupled resonators $(Q_L = \omega_o / \Delta\omega)$.

For example, by applying Equations (3.5.38–40) to a series RLC resonator, we readily obtain:

$$Q_I = \omega_o L / R \qquad (3.5.42)$$

$$Q_E = \omega_o L / R_{Th} \qquad\qquad (3.5.43)$$

$$Q_L = \omega_o L / (R_{Th} + R) \qquad\qquad (3.5.44)$$

For a parallel RLC resonator the Q's become:

$$Q_I = \omega_o RC \qquad\qquad (3.5.45)$$

$$Q_E = \omega_o R_{Th} C \qquad\qquad (3.5.46)$$

$$Q_L = \omega_o C R_{Th} R / (R_{Th} + R) \qquad\qquad (3.5.47)$$

Since the source and load resistances are matched for maximum power dissipation at resonance, it follows from Figure 3.5.3 that a *critically coupled resonator* or *matched resonator* results when $Q_I = Q_E$. These expressions for Q are in terms of energies stored and power dissipated, and can readily be applied to electromagnetic resonances of cavities or other structures, yielding their bandwidths and conditions for maximum power transfer to loads, as discussed in Section 9.4.

Chapter 4: Static and Quasistatic Fields

4.1 Introduction

Static electric and magnetic fields are governed by the static forms of Maxwell's equations in differential and integral form for which $\partial/\partial t \to 0$:

$$\nabla \times \overline{E} = 0 \qquad \oint_C \overline{E} \bullet d\overline{s} = 0 \qquad\qquad\qquad \textit{Faraday's Law} \qquad (4.1.1)$$

$$\nabla \times \overline{H} = \overline{J} \qquad \oint_C \overline{H} \bullet d\overline{s} = \iint_A \overline{J} \bullet \hat{n} \, da \qquad\qquad \textit{Ampere's Law} \qquad (4.1.2)$$

$$\nabla \bullet \overline{D} = \rho \qquad \oiint_A (\overline{D} \bullet \hat{n}) da = \iiint_V \rho \, dv = Q \qquad \textit{Gauss's Law} \qquad (4.1.3)$$

$$\nabla \bullet \overline{B} = 0 \qquad \oiint_A (\overline{B} \bullet \hat{n}) da = 0 \qquad\qquad\qquad \textit{Gauss's Law} \qquad (4.1.4)$$

As shown in (1.3.5), Gauss's law (4.1.3) leads to the result that a single point charge Q at the origin in vacuum yields produces an electric field at radius r of:

$$\overline{E}(r) = \hat{r} \, Q/4\pi\varepsilon_o r^2 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.1.5)$$

Superposition of such contributions to $\overline{E}(\overline{r})$ from a charge distribution $\rho(\overline{r}')$ located within the volume V' yields:

$$\overline{E}(\overline{r}) = \hat{r} \iiint_{V'} \frac{\rho(\overline{r}')}{4\pi\varepsilon_O |\overline{r} - \overline{r}'|^2} dv' \qquad \textit{Coulomb's superposition integral} \qquad (4.1.6)$$

where $\hat{r}$ is outside the integral because $r \gg \sqrt[3]{V'}$. A more complex derivation given in Section 10.1 yields the corresponding equation for static magnetic fields:

$$\overline{H}(\overline{r}, t) = \iiint_{V'} \frac{\overline{J}' \times (\overline{r} - \overline{r}')}{4\pi \, |\overline{r} - \overline{r}'|^3} dv' \qquad\qquad \textit{Biot-Savart law} \qquad (4.1.7)$$

Any static electric field can be related to an electric potential distribution Φ[volts] because $\nabla \times \overline{E} = 0$ implies $\overline{E} = -\nabla\Phi$, where the voltage difference between two points (1.3.12) is:

$$\Phi_1 - \Phi_2 = \int_1^2 \overline{E} \bullet d\overline{s} \qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.1.8)$$

Similarly, in current-free regions of space $\nabla \times \overline{H} = 0$ implies $\overline{H} = -\nabla\Psi$ [Amperes], where Ψ is magnetic potential. Therefore the magnetic potential difference between two points is:

$$\Psi_1 - \Psi_2 = \int_1^2 \overline{H} \cdot d\overline{s} \qquad (4.1.9)$$

This definition of magnetic potential is useful in understanding the magnetic circuits discussed in Section 4.4.3.

Often not all source charges and currents are given because some reside on given equipotential surfaces and assume an unknown distribution consistent with that constraint. To address this case, Maxwell's equations can be simply manipulated to form Laplace's equation, which can sometimes be solved by separation of variables, as discussed in Section 4.5, or usually by numerical methods. Section 4.6 then discusses the utility of flux tubes and field mapping for understanding static field distributions.

Quasistatics assumes that the field strengths change so slowly that the electric and magnetic fields induced by those changes (the contributions to $\overline{E}$ and $\overline{H}$ from the $\partial/\partial t$ terms in Faraday's and Ampere's laws) are sufficiently small that their own induced fields ($\propto (\partial/\partial t)^2$) can be neglected; only the original and first-order induced fields are therefore of interest. Quasistatic examples were discussed in Chapter 3 in the context of resistors, capacitors, and inductors. The mirror image technique described in Section 4.2 is used for static, quasistatic, and dynamic problems and incidentally in the discussion in Section 4.3 concerning exponential relaxation of field strengths in conducting media and skin depth.

4.2   Mirror image charges and currents

One very useful problem solving technique is to change the problem definition to one that is easier to solve but is known to have the same answer. An excellent example of this approach is the use of mirror-image charges and currents, which also works for wave problems.[10]
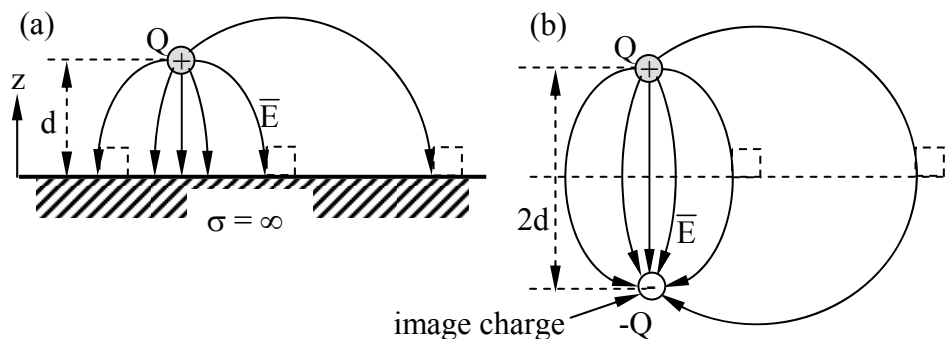
Figure 4.2.1   Image charge for an infinite planar perfect conductor.

---

[10] Another example of this approach is use of duality between $\overline{E}$ and $\overline{H}$, as discussed in Section 9.2.6.

Consider the problem of finding the fields produced by a charge located a distance d above an infinite perfectly conducting plane, as illustrated in Figure 4.2.1(a). Boundary conditions at the conductor require only that the electric field lines be perpendicular to its surface. Any other set of boundary conditions that imposes the same constraint must yield the same unique solution by virtue of the uniqueness theorem of Section 2.8.

One such set of equivalent boundary conditions invokes a duplicate *mirror image charge* a distance 2d away from the original charge and of opposite sign; the conductor is removed. The symmetry for equal and opposite charges requires the electric field lines $\overline{E}$ to be perpendicular to the original surface of the conductor at z = 0; this results in $\overline{E}$ being exactly as it was for z > 0 when the conductor was present, as illustrated in Figure 4.2.1(b). Therefore uniqueness says that above the half-plane the fields produced by the original charge plus its mirror image are identical to those of the original problem. The fields below the original half plane are clearly different, but they are not relevant to the original problem.

This equivalence applies for multiple charges or for a charge distribution, as illustrated in Figure 4.2.2. In fact the mirror image method remains valid so long as the charges change value or position slowly with respect to the relaxation time ε/σ of the conductor, as discussed in Section 4.4.1. The relaxation time is the 1/e time constant required for the charges within the conductor to approach new equilibrium positions after the source charge distribution outside changes.
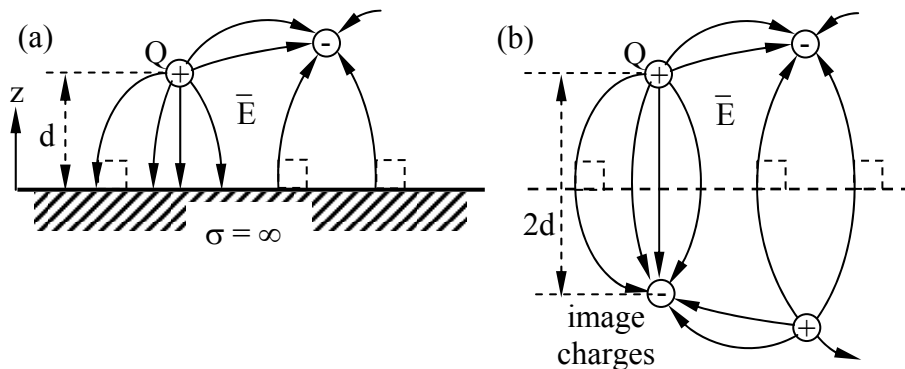
Figure 4.2.2 Multiple image charges.

Because the mirror image method works for varying or moving charges, it works for the currents that must be associated with them by conservation of charge (2.1.21), as suggested in Figure 4.2.3 (a) and (b). Figure 4.2.3(d) also suggests how the magnetic fields produced by these currents satisfy the boundary conditions for the conducting plane: at the surface of a perfect conductor $\overline{H}$ is only parallel.

The mirror image method continues to work if the upper half plane contains a conductor, as illustrated in Figure 4.2.4; the conductor must be imaged too. These conductors can even be at angles, as suggested in Figure 4.2.4(b). The region over which the deduced fields are valid is naturally restricted to the original opening between the conductors. Still more complex image

configurations can be used for other conductor placements, and may even involve an infinite series of progressively smaller image charges and currents.
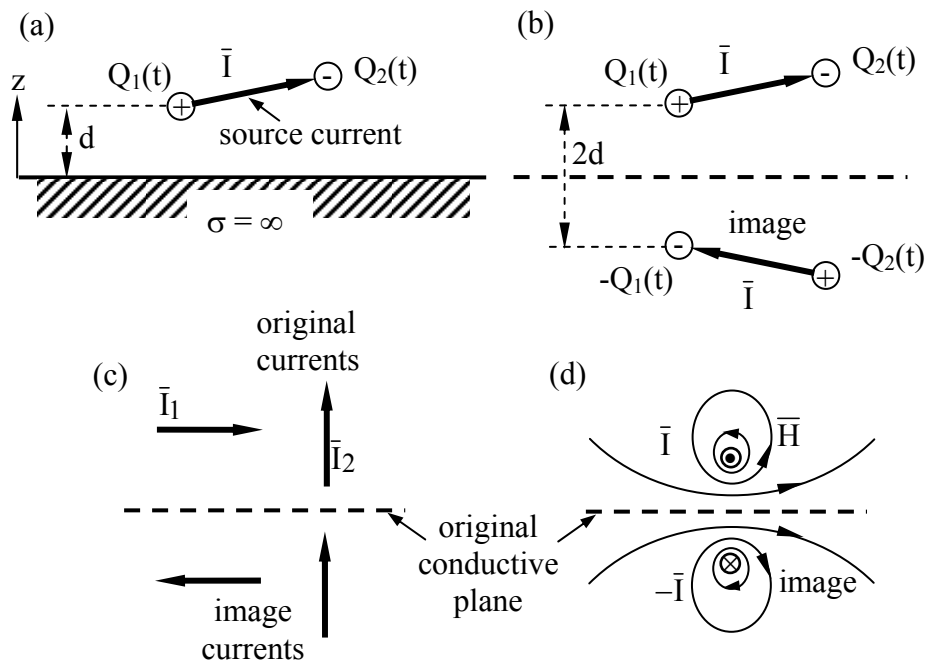
Figure 4.2.3   Image currents.
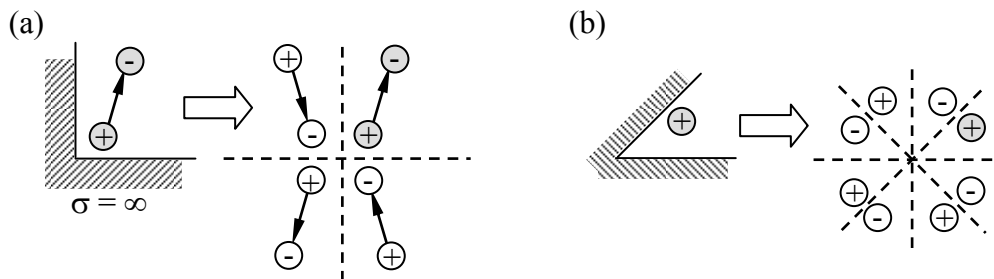
Figure 4.2.4   Image charges and currents for intersecting conductors.

4.3   Relaxation of fields, skin depth

4.3.1   Relaxation of electric fields and charge in conducting media

Electric and magnetic fields established in conducting time-invariant homogeneous media tend to decay exponentially unless maintained. Under the quasistatic assumption all time variations are sufficiently slow that contributions to $\overline{E}$ by $\partial\overline{B}/\partial t$ are negligible, which avoids wave-like behavior and simplifies the problem. This relaxation process is governed by the conservation-of-charge relation (2.1.21), Gauss's law $\left(\nabla \bullet \overline{D} = \rho\right)$, and Ohm's law $\left(\overline{J} = \sigma\overline{E}\right)$:

$$\nabla \bullet \overline{J} + \partial\rho/\partial t = 0 = \nabla \bullet (\sigma\overline{E}) + (\partial/\partial t)(\nabla \bullet \varepsilon\overline{E}) = \nabla \bullet \left[(\sigma + \varepsilon\partial/\partial t)\overline{E} \right] = 0 \qquad (4.3.1)$$

Since an arbitrary $\overline{E}$ can be established by initial conditions, the general solution to (4.3.1) requires $(\sigma + \varepsilon\partial/\partial t)\nabla\bullet\overline{E} = 0$, leading to the differential equation:

$$(\partial/\partial t + \sigma/\varepsilon)\rho = 0 \qquad (4.3.2)$$

where $\nabla\bullet\overline{E} = \rho/\varepsilon$. This has the solution that $\rho(\overline{r})$ relaxes exponentially with a charge *relaxation time* constant $\tau = \varepsilon/\sigma$ seconds:

$$\rho(\overline{r}) = \rho_o(\overline{r})e^{-\sigma t/\varepsilon} = \rho_o(\overline{r})e^{-t/\tau}\left[\text{m}^{-3} \right] \qquad \text{(\textit{charge relaxation})} \qquad (4.3.3)$$

It follows that an arbitrary initial electric field $\overline{E}(\overline{r})$ in a medium having uniform ε and σ will also decay exponentially with the same time constant ε/σ because Gauss's law relates $\overline{E}$ and ρ linearly:

$$\nabla\bullet\overline{E} = \rho(t)/\varepsilon \qquad (4.3.4)$$

where $\nabla\bullet\overline{E}_o \equiv \rho_o/\varepsilon$. Therefore *electric field relaxation* is characterized by:

$$\overline{E}(\overline{r},t) = \overline{E}_o(\overline{r})e^{-\sigma t/\varepsilon} \left[\text{v m}^{-1} \right] \qquad \text{(electric field relaxation)} \qquad (4.3.5)$$

We should expect such exponential decay because any electric fields in a conductor will generate currents and therefore dissipate power proportional to J^2 and E^2. But the stored electrical energy is also proportional to E^2, and power dissipation is the negative derivative of stored energy. That is, the energy decays at a rate proportional to its present value, which results in exponential decay. In copper $\tau = \varepsilon_o/\sigma \cong 9\times10^{-12}/(5\times10^7) \cong 2\times10^{-19}$ seconds, short compared to any delay of common interest. The special case of parallel-plate resistors and capacitors is discussed in Section 3.1.

Example 4.3A

What are the electric field relaxation time constants τ for sea water ($\varepsilon \cong 80\varepsilon_o$, $\sigma \cong 4$) and dry soil ($\varepsilon \cong 2\varepsilon_o$, $\sigma \cong 10^{-5}$)? For what radio frequencies can they be considered good conductors?

Solution: Equation (4.3.5) yields $\tau = \varepsilon/\sigma \cong (80\times8.8\times10^{-12})/4 \cong 1.8\times10^{-10}$ seconds for seawater, and $(2\times8.8\times10^{-12})/10^{-5} \cong 1.8\times10^{-6}$ seconds for dry soil. So long as $\overline{E}$ changes slowly with respect to τ, the medium has time to cancel $\overline{E}$; frequencies below ~5 GHz and ~500 kHz have this property for seawater and typical dry soil, respectively, which behave like good conductors at these lower frequencies. Moist soil behaves like a conductor up to ~5 MHz and higher.

4.3.2 Relaxation of magnetic fields in conducting media

Magnetic fields and their induced currents similarly decay exponentially in conducting media unless they are externally maintained; this decay process is often called *magnetic diffusion* or *magnetic relaxation*. We assume that the time variations are sufficiently slow that contributions to $\overline{H}$ by $\partial \overline{D}/\partial t$ are negligible. In this limit Ampere's law becomes:

$$\nabla \times \overline{H} = \overline{J} = \sigma \overline{E} \tag{4.3.6}$$

$$\nabla \times (\nabla \times \overline{H}) = \sigma \nabla \times \overline{E} = -\sigma \mu \partial \overline{H}/\partial t = -\nabla^2 \overline{H} + \nabla(\nabla \bullet \overline{H}) = -\nabla^2 \overline{H} \tag{4.3.7}$$

where Faraday's law, the vector identity (2.2.6), and Gauss's law $(\nabla \bullet \overline{B} = 0)$ were used.

The resulting differential equation:

$$\sigma \mu \partial \overline{H}/\partial t = \nabla^2 \overline{H} \tag{4.3.8}$$

has at least one simple solution:

$$\overline{H}(z,t) = \hat{x} H_0 e^{-t/\tau_m} \cos kz \tag{4.3.9}$$

where we assumed an x-polarized z-varying sinusoid. Substituting (4.3.9) into (4.3.8) yields the desired time constant:

$$\tau_m = \mu\sigma/k^2 = \mu\sigma\lambda^2/4\pi^2 \quad \text{[s]} \qquad \text{(magnetic relaxation time)} \tag{4.3.10}$$

Thus the lifetime of magnetic field distributions in conducting media increases with permeability (energy storage density), conductivity (reducing dissipation for a given current), and the wavelength squared ($\lambda = 2\pi/k$).

4.3.3 Induced currents

Quasistatic magnetic fields induce electric fields by virtue of Faraday's law: $\nabla \times \overline{E} = -\mu \partial \overline{H}/\partial t$. In conductors these induced electric fields drive currents that obey *Lenz's law*: "The direction of induced currents tends to oppose changes in magnetic flux." Induced currents find wide application, for example, in: 1) heating, as in induction furnaces that melt metals, 2) mechanical actuation, as in induction motors and impulse generators, and 3) electromagnetic shielding. In some cases these induced currents are undesirable and are inhibited by subdividing the conductors into elements separated by thin insulating barriers. All these examples are discussed below.

First consider a simple conducting hollow cylinder of length W driven circumferentially by current $I_o u(t)$, as illustrated in Figure 4.3.1, where $u(t)$ is the unit step function (the current is zero until $t = 0$, when it becomes I_o). Centered in the outer cylinder is an isolated second cylinder of conductivity σ and having a thin wall of thickness δ; its length and diameter are W and D << W, respectively.
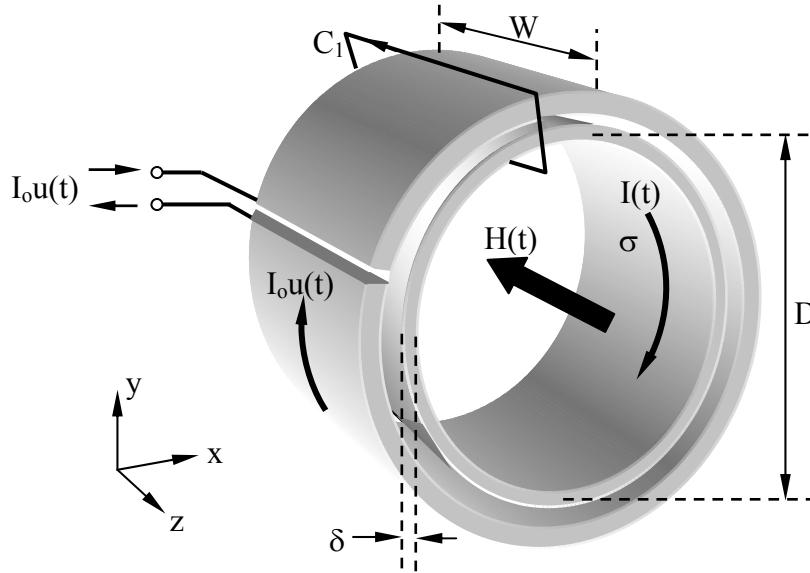
Figure 4.3.1   Relaxation penetration of a magnetic field into a conducting cylinder.

If the inner cylinder were a perfect conductor, then the current $I_o u(t)$ would produce an equal and opposite image current $\sim$-$I_o u(t)$ on the outer surface of the inner cylinder, thus producing a net zero magnetic field inside the cylinder formed by that image current. Consider the integral of $\overline{H} \bullet d\overline{s}$ around a closed contour C_1 that threads both cylinders and circles zero net current at t = 0+; this integral yields zero. If the inner conductor were slightly resistive, then the same equal and opposite current would flow on the inner cylinder, but it would slowly dissipate heat until the image current decayed to zero and the magnetic field inside reached the maximum value I_o/W [A m^{-1}] associated with the outer current I_o. These conclusions are quantified below.

The magnetic field H inside the inner cylinder depends on the currents flowing in the outer and inner cylinders, I_o and $I(t)$, respectively:

$$H(t) = u(t)[I_o + I(t)]/W \tag{4.3.11}$$

The current $I(t)$ flowing in the inner cylinder is driven by the voltage induced by $H(t)$ via Faraday's law (2.4.14):

$$\oint_{C_2} \overline{E} \bullet d\overline{s} = IR = \mu_o \int_A \left(d\overline{H}/dt \right) \bullet d\overline{a} = \mu_o A \, dH/dt \tag{4.3.12}$$

where the contour C_2 is in the x-y plane and circles the inner cylinder with diameter D. The area circled by the contour $A = \pi D^2/4$. The circumferential resistance of the inner cylinder is R =

$\pi D/\sigma\delta W$ ohms. For simplicity we assume that the permeability here is μ_o everywhere. Substituting (4.3.11) into (4.3.12) yields a differential equation for I(t):

$$I(t) = - (\mu_o A/WR) \, dI/dt \qquad\qquad\qquad (4.3.13)$$

Substituting the general solution $I(t) = Ke^{-t/\tau}$ into (4.3.13) yields:

$$Ke^{-t/\tau} = (\mu_o A/WR\tau)Ke^{-t/\tau} \qquad\qquad\qquad (4.3.14)$$

$$\tau = \mu_o A/WR = \mu_o A\sigma\delta/\pi D \ \ [\text{s}] \qquad\qquad (\text{magnetic relaxation time}) \quad (4.3.15)$$

Thus the greater the conductivity of the inner cylinder, and the larger its product μA, the longer it takes for transient magnetic fields to penetrate it. For the special case where $\delta = D/4\pi$ and $A = D^2$, we find $\tau = \mu_o\sigma(D/2\pi)^2$, which is the same magnetic relaxation time constant derived in (4.3.10) if we identify D with the wavelength λ of the magnetic field variations. Equation (4.3.15) is also approximately correct if $\mu_o \rightarrow \mu$ for the inner cylinder.

Since H(t) = 0 at t = 0+, (4.3.11) yields $I(t = 0+) = - I_o$, and the solution $I(t) = Ke^{-t/\tau}$ becomes:

$$I(t) = - I_o \, e^{-t/\tau} \ \ [\text{A}] \qquad\qquad\qquad (4.3.16)$$

The magnetic field inside the inner cylinder follows from (4.3.16) and (4.3.11):

$$H(t) = u(t)I_o \, (1 - e^{-t/\tau})/W \ \ [\text{A m}^{-1}] \qquad\qquad\qquad (4.3.17)$$

The geometry of Figure 4.3.1 can be used to heat resistive materials such as metals electrically by placing the metals in a ceramic container that sinusoidal magnetic fields penetrate easily. The induced currents can then melt the material quicker by heating the material throughout rather than just at the surface, as would a flame. The frequency f generally must be sufficiently low that the magnetic fields penetrate a significant fraction of the container diameter; $f << 1/\tau$.

The inner cylinder of Figure 4.3.1 can also be used to shield its interior from alternating magnetic fields by designing it so that its time constant τ is much greater than the period of the undesired AC signal; large values of $\mu\sigma\delta$ facilitate this since $\tau = \mu_o A\sigma\delta/\pi D$ (4.3.15). Since we can model a solid inner cylinder as a continuum of concentric thin conducting shells, it follows that the inner shells will begin to see significant magnetic fields only after the surrounding shells do, and therefore the time delay experienced increases with depth. This is consistent with $\tau \propto \delta$. The penetration of alternating fields into conducting surfaces is discussed further in Section 9.3 in terms of the exponential penetration skin depth $\delta = \sqrt{2/\omega\mu\sigma} \ \ [\text{m}]$.

Two actuator configurations are suggested by Figure 4.3.1. First, the inner cylinder could be inserted only part way into the outer cylinder. Then the net force on the inner cylinder would expel it when the outer cylinder was energized because the polarity of these two electromagnets

are reversed, the outer one powered by I_o and the inner one by $-I_o(1 - e^{-t/\tau})$. Electromagnetic forces are discussed more fully in Chapter 5; here it suffices to note that induced currents can be used to simplify electromechanical actuators. A similar "kick" can be applied to a flat plate placed across the end of the outer cylinder, for again the induced cylindrically shaped mirror image current would experience a transient repulsive force. Mirror-image currents were discussed in Section 4.2.

The inner cores in transformers and some inductors are typically iron and are circled by wires carrying alternating currents, as discussed in Section 3.2. The alternating currents induce circular currents in the core called *eddy currents* that dissipate power. To minimize such induced currents and losses, high-μ conducting cores are commonly composed of many thin sheets separated from each other by thin coats of varnish or other insulator that largely blocks those induced currents; these are called *laminated cores*. A rough estimate of the effectiveness of using N plates instead of one can be obtained by noting that the power P_d dissipated in each lamination is proportional to V^2/R, where $V = \oint_C \overline{E} \bullet d\overline{s}$ is the loop voltage induced by H(t) and R is the effective resistance of that loop. By design H(t) usually penetrates the full transformer core. Thus V is roughly proportional to the area of each lamination in the plane perpendicular to $\overline{H}$, which decreases as 1/N. The resistance R experienced by the induced current circulating in each lamination increases roughly by N since the width of the channel through which it can flow is reduced as N increases while the length of the channel changes only moderately. The total power dissipated for N laminations is thus roughly proportional to $NV^2/R \propto NN^{-2}/N = N^{-2}$. Therefore we need only increase N to the point where the power loss is tolerable and the penetration of the transformer core by H(t) is nearly complete each period.

Example 4.3B

How long does it take a magnetic field to penetrate a 1-mm thick metal cylinder of diameter D with conductivity 5×10^7 [S/m] if $\mu = \mu_o$? Design a shield for a ~10-cm computer that blocks 1-MHz magnetic fields emanating from an AM radio.

Solution: If we assume the geometry of Figure 4.3.1 and use (4.3.15), $\tau = \mu_o A\sigma\delta/\pi D$, we find τ = $1.3\times10^{-6}\times D\times5\times10^7\times10^{-3}/4 = 0.016D$ seconds, where $A = \pi D^2/4$ and $\delta = 10^{-3}$. If D = 0.1, then $\tau = 1.6\times10^{-2}$ seconds, which is ~10^5 longer than the rise time ~$10^{-6}/2\pi$ of a 1-MHz signal. If a smaller ratio of 10^2 is sufficient, then a one-micron thick layer of metal evaporated on thin plastic might suffice. If the metal had $\mu = 10^4\mu_o$, then a one-micron thick layer would provide a safety factor of 10^6.

4.4   *Static fields in inhomogeneous materials*

4.4.1   Static electric fields in inhomogeneous materials

Many practical problems involve *inhomogeneous media* where the boundaries may be abrupt, as in most capacitors or motors, or graded, as in many semiconductor or optoelectronic devices. The basic issues are well illustrated by the static cases discussed below. Sections 4.4.1 and 4.4.2 discuss static electric and magnetic fields, respectively, in inhomogeneous media. To simplify

the discussion, only media characterized by real scalar values for ε, μ, and σ will be considered, where all three properties can be a function of position.

Static electric fields in all media are governed by the static forms of Faraday's and Gauss's laws:

$$\nabla \times \overline{E} = 0 \tag{4.4.1}$$

$$\nabla \bullet \overline{D} = \rho_f \tag{4.4.2}$$

and by the constitutive relations:

$$\overline{D} = \varepsilon \overline{E} = \varepsilon_o \overline{E} + \overline{P} \tag{4.4.3}$$

$$\overline{J} = \sigma \overline{E} \tag{4.4.4}$$

A few simple cases illustrate how these laws can be used to characterize inhomogeneous conductors and dielectrics. Perhaps the simplest case is that of a wire or other conducting structure (1) imbedded in a perfectly insulating medium (2) having conductivity $\sigma = 0$. Since charge is conserved, the perpendicular components of current must be the same on both sides of the boundary so that $J_{1\perp} = J_{2\perp} = 0 = E_{2\perp}$. Therefore all currents in the conducting medium are trapped within it and at the surface must flow parallel to that surface.

Let's consider next the simple case of an inhomogeneous slab between two parallel perfectly conducting plates spaced L apart in the x direction at a potential difference of V_o volts, where the terminal at x = 0 has the greater voltage. Suppose that the medium has permittivity ε, current density J_o, and inhomogeneous conductivity $\sigma(x)$, where:

$$\sigma = \sigma_o \left/ \left[1 + \frac{x}{L} \right] \right. \left[\text{Siemens m}^{-1} \right] \tag{4.4.5}$$

The associated electric field follows from (4.4.4):

$$\overline{E} = \overline{J} / \sigma = \hat{x} \frac{J_o}{\sigma_o} \left(1 + \frac{x}{L} \right) \left[\text{Vm}^{-1} \right] \tag{4.4.6}$$

The free charge density in the medium then follows from (4.4.2) and is:

$$\rho_f = \nabla \bullet \overline{D} = \left(\varepsilon J_o / \sigma_o \right) \left(\partial / \partial x \right) \left(1 + x/L \right) = \varepsilon J_o / \sigma_o L \left[\text{Cm}^{-3} \right] \tag{4.4.7}$$

Note from the derivative in (4.4.7) that abrupt discontinuities in conductivity generally produce free surface charge ρ_s at the discontinuity. Although inhomogeneous conductors have a net free charge density throughout the volume, they may or may not also have a net polarization charge

density $\rho_p = -\nabla \bullet \overline{P}$, which is defined in (2.5.12) and can be deduced from the polarization vector $\overline{P} = \overline{D} - \varepsilon_o \overline{E} = (\varepsilon - \varepsilon_o) \overline{E}$ using (4.4.7):

$$\rho_p = -\nabla \bullet \overline{P} = -\nabla \bullet \left[(\varepsilon - \varepsilon_o) \overline{E} \right] = (\varepsilon - \varepsilon_o) J_o / \sigma_o L \ \left[\mathrm{Cm}^{-3} \right] \qquad (4.4.8)$$

Now let's consider the effects of inhomogeneous permittivity $\varepsilon(x)$ in an insulating medium ($\sigma = 0$) where:

$$\varepsilon = \varepsilon_o \left(1 + \frac{x}{L} \right) \qquad (4.4.9)$$

Since the insulating slab should contain no free charge and the boundaries force $\overline{D}$ to be in the x direction, therefore $\overline{D}$ cannot be a function of x because $\nabla \bullet \overline{D} = \rho_f = 0$. But $\overline{D} = \varepsilon(x) \overline{E}(x)$; therefore the x dependence of $\overline{E}$ must cancel that of ε, so:

$$\overline{E} = \hat{x} E_o \Big/ \left(1 + \frac{x}{L} \right) \qquad (4.4.10)$$

E_o is an unknown constant and can be found relative to the applied voltage V_o:

$$V_o = \int_0^L E_x \ dx = \int_0^L \left[E_o \Big/ \left(1 + \frac{x}{L} \right) \right] dx = L \ E_o \ln 2 \qquad (4.4.11)$$

Combining (4.4.9–11) leads to a displacement vector $\overline{D}$ that is independent of x (boundary conditions mandate continuity of $\overline{D}$), and a non-zero polarization charge density ρ_p distributed throughout the medium:

$$\overline{D} = \varepsilon \overline{E} = \hat{x} \varepsilon_o V_o / (L \ln 2) \qquad (4.4.12)$$

$$\begin{aligned} \rho_p &= -\nabla \bullet \overline{P} = -\nabla \bullet (\overline{D} - \varepsilon_o \overline{E}) = \varepsilon_o \nabla \bullet \overline{E} \\ &= \frac{\varepsilon_o V_o}{L \ln 2} \frac{\partial}{\partial x} (1 + x/L)^{-1} = \frac{-\varepsilon_o V_o}{(L+x)^2 \ln 2} \ \left[\mathrm{Cm}^{-3} \right] \end{aligned} \qquad (4.4.13)$$

A similar series of computations readily handles the case where both ε and σ are inhomogeneous.

Example 4.4A
A certain capacitor consists of two parallel conducting plates, one at z = 0 and +V volts and one at z = d and zero volts. They are separated by a dielectric slab of permittivity ε, for which the conductivity is small and different in the two halves of the dielectric, each of which is d/2 thick; $\sigma_1 = 3\sigma_2$. Assume the interface between σ_1 and σ_2 is parallel to the capacitor plates and is

located at z = 0. What is the free charge density $\rho_f(z)$ in the dielectric, and what is $\overline{E}(z)$ where z is the coordinate perpendicular to the plates?

Solution: Since charge is conserved, $\overline{J}_1 = \overline{J}_2 = \sigma_1 \overline{E}_1 = \sigma_2 \overline{E}_2$, so $\overline{E}_2 = \sigma_1 \overline{E}_1 / \sigma_2 = 3\overline{E}_1$. But $(E_1 + E_2)d/2 = V$, so $4E_1 d/2 = V$, and $E_1 = V/2d$. The surface charge on the lower plate is $\rho_{s(z=0)} = \overline{D}_{z=0} = \varepsilon E_1 = \varepsilon V/2d$ [C/m^2], and ρ_s on the upper plate is $-\overline{D}_{z=d} = -\varepsilon E_2 = -\varepsilon 3V/2d$. The free charge at the dielectric interface is $\rho_{s\ (z\ =\ d/2)} = D_2 - D_1 = \varepsilon(E_2 - E_1) = \varepsilon V/d$. Charge can accumulate at all three surfaces because the dielectric conducts. The net charge is zero. The electric field between capacitor plates was discussed in Section 3.1.2.

4.4.2   Static magnetic fields in inhomogeneous materials

Static magnetic fields in most media are governed by the static forms of Ampere's and Gauss's laws:

$$\nabla \times \overline{H} = 0 \qquad (4.4.14)$$

$$\nabla \bullet \overline{B} = 0 \qquad (4.4.15)$$

and by the constitutive relations:

$$\overline{B} = \mu\overline{H} = \mu_o\left(\overline{H} + \overline{M}\right) \qquad (4.4.16)$$

One simple case illustrates how these laws characterize inhomogeneous magnetic materials. Consider a magnetic material that is characterized by $\mu(x)$ and has an imposed magnetic field $\overline{B}$ in the x direction. Since $\nabla \bullet \overline{B} = 0$ it follows that $\overline{B}$ is constant $\left(\overline{B}_o\right)$ throughout, and that $\overline{H}$ is a function of x:

$$\overline{H} = \frac{\overline{B}_o}{\mu(x)} \qquad (4.4.17)$$

As a result, higher-permeability regions of magnetic materials generally host weaker magnetic fields $\overline{H}$, as shown in Section 3.2.2 for the toroidal inductors with gaps. In many magnetic devices μ might vary four to six orders of magnitude, as would $\overline{H}$.

4.4.3   Electric and magnetic flux trapping in inhomogeneous systems

Currents generally flow in conductors that control the spatial distribution of $\overline{J}$ and electric potential $\Phi(\overline{r})$. Similarly, high-permeability materials with $\mu \gg \mu_o$ can be used to form

magnetic circuits that guide $\overline{B}$ and control the spatial form of the static curl-free *magnetic potential* $\Psi(\overline{r})$.

Faraday's law says that static electric fields $\overline{E}$ are curl-free:

$$\nabla \times \overline{E} = -\frac{\partial \overline{B}}{\partial t} = 0 \qquad \text{(Faraday's law)} \qquad (4.4.18)$$

Since $\nabla \times \overline{E} = 0$ in static cases, it follows that:

$$\overline{E} = -\nabla\Phi \qquad (4.4.19)$$

where Φ is the *electric potential* [volts] as a function of position in space. But Gauss's law says $\nabla \cdot \overline{E} = \rho/\varepsilon$ in regions where ρ is constant. Therefore $\nabla \bullet \overline{E} = -\nabla^2\Phi = \rho/\varepsilon$ and:

$$\nabla^2\Phi = -\frac{\rho}{\varepsilon} \qquad \text{(Laplace's equation)} \qquad (4.4.20)$$

In static current-free regions of space with constant permeability μ, Ampere's law (2.1.6) says:

$$\nabla \times \overline{H} = 0 \qquad (4.4.21)$$

and therefore $\overline{H}$, like $\overline{E}$, can be related to a scalar *magnetic potential* [Amperes] Ψ:

$$\overline{H} = -\nabla\Psi \qquad (4.4.22)$$

Since $\nabla \bullet \overline{H} = 0$ when μ is independent of position, it follows that $\nabla \bullet (-\nabla\Psi) = \nabla^2\Psi$ and:

$$\nabla^2\Psi = 0 \qquad \text{(Laplace's equation for magnetic potential)} \qquad (4.4.23)$$

The perfect parallel between Laplace's equations (4.4.20) and (4.4.23) for electric and magnetic fields in charge-free regions offers a parallel between current density $\overline{J} = \sigma\overline{E}$ [A/m^2] and magnetic flux density $\overline{B} = \mu\overline{H}$, and also between conductivity σ and permeability μ as they relate to gradients of electric and magnetic potential, respectively:

$$\nabla^2\Phi = 0 \qquad\qquad \nabla^2\Psi = 0 \qquad (4.4.24)$$

$$\overline{E} = -\nabla\Phi \qquad\qquad \overline{H} = -\nabla\Psi \qquad (4.4.25)$$

$$\overline{J} = \sigma\overline{E} = -\sigma\nabla\Phi \qquad\qquad \overline{B} = \mu\overline{H} = -\mu\nabla\Psi \qquad (4.4.26)$$

Just as current is confined to flow within wires imbedded in insulating media having $\sigma \cong 0$, so is magnetic flux $\overline{B}$ trapped within high-permeability materials imbedded in very low permeability

media, as suggested by the discussion in Section 3.2.2 of how magnetic fields are confined within high-permeability toroids.

The boundary condition (2.6.5) that $\overline{B}_\perp$ is continuous requires that $\overline{B}_\perp \cong 0$ at boundaries with media having $\mu \cong 0$; thus essentially all magnetic flux $\overline{B}$ is confined within permeable magnetic media having $\mu \gg 0$.
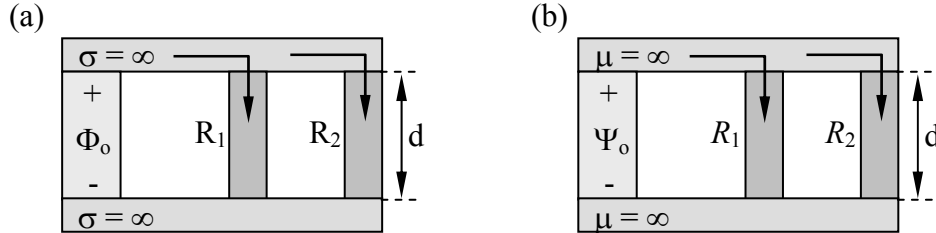
Figure 4.4.1   Current and magnetic flux-divider circuits.

Two parallel examples that help clarify the issues are illustrated in Figure 4.4.1. In Figure 4.4.1(a) a battery connected to perfect conductors apply the same voltage Φ_o across two conductors in parallel; A_i, σ_i, d_i, and I_i are respectively their cross-sectional area, conductivity, length, and current flow for i = 1,2. The current through each conductor is given by (4.4.26) and:

$$I_i = J_i A_i = \sigma_\iota \nabla \Phi_i A = \sigma_\iota \Phi_o A_i / d_i = \Phi_o / R_i \qquad (4.4.27)$$

where:

$$R_i = d_i / \sigma_i A_i \ \text{[ohms]} \qquad (4.4.28)$$

is the *resistance* of conductor i, and I = V/R is *Ohm's law*.

For the *magnetic circuit* of Figure 4.4.1(b) a parallel set of relations is obtained, where the total magnetic flux $\Lambda = BA$ [*Webers*] through a cross-section of area A is analogous to current I = JA. The magnetic flux Λ through each magnetic branch is given by (4.4.26) so that:

$$\Lambda_i = B_i A_i = \mu_\iota \nabla \Psi_i A_i = \mu_\iota \Psi_o A_i / d_i = \Psi_o / R_i \qquad (4.4.29)$$

where:

$$R_i = d_i / \mu_i A \qquad (4.4.30)$$

is the magnetic *reluctance* of branch i, analogous to the resistance of a conductive branch.

Because of the parallel between current I and magnetic flux Λ, they divide similarly between alternative parallel paths. That is, the total current is:

$$I_o = I_1 + I_2 = \Phi_0(R_1 + R_2)/R_1 R_2 \qquad (4.4.31)$$

The value of Φ_o found from (4.4.31) leads directly to the *current-divider equation*:

$$I_1 = \Phi_o/R_1 = I_oR_2/(R_1 + R_2) \tag{4.4.32}$$

So, if $R_2 = \infty$, all I_o flows through R_1; $R_2 = 0$ implies no current flows through R_1; and $R_2 = R_1$ implies half flows through each branch. The corresponding equations for total magnetic flux and flux division in magnetic circuits are:

$$\Lambda_o = \Lambda_1 + \Lambda_2 = \Psi_0(R_1 + R_2)/R_1R_2 \tag{4.4.33}$$

$$\Lambda_1 = \Psi_o/R_1 = \Lambda_oR_2/(R_1 + R_2) \tag{4.4.34}$$

Although the conductivity of insulators surrounding wires is generally over ten orders of magnitude smaller than that of the wires, the same is not true for the permeability surrounding high-μ materials, so there generally is some small amount of flux leakage from such media; the trapping is not perfect. In this case $\overline{H}$ outside the high-μ material is nearly perpendicular to its surface, as shown in (2.6.13).

Example 4.4B

The magnetic circuit of Figure 4.4.1(b) is driven by a wire that carries 3 amperes and is wrapped 50 times around the leftmost vertical member in a clockwise direction as seen from the top. That member has infinite permeability ($\mu = \infty$), as do the top and bottom members. If the rightmost member is missing, what is the magnetic field $\overline{H}$ in the vertical member R_1, for which the length is d and $\mu \gg \mu_o$? If both R_1 and R_2 are in place and identical, what then are $\overline{H}_1$ and $\overline{H}_2$? If R_2 is removed and R_1 consists of two long thin bars in series having lengths d_a and d_b, cross-sectional areas A_a and A_b, and permeabilities μ_a and μ_b, respectively, what then are $\overline{H}_a$ and $\overline{H}_b$?

Solution: For this static problem Ampere's law (4.1.2) becomes $\oint_C \overline{H} \cdot d\overline{s} = \oint\!\!\!\oint_A \overline{J} \cdot \hat{n}\, da = N\,I$
= 50×3 =150 [A] = Hd. Therefore $\overline{H} = \hat{z}\,150/d$ [A m^{-1}], where $\hat{z}$ and $\overline{H}$ are upward due to the right-hand rule associated with Ampere's law. If R_2 is added, both the integrals of $\overline{H}$ through the two branches must still equal NI, so $\overline{H}$ remains $\hat{z}\,150/d$ [A m^{-1}] in both branches. For the series case the integral of $\overline{H}$ yields $H_ad_a + H_bd_b = NI$. Because the magnetic flux is trapped within this branch, it is constant: $\mu_aH_aA_a = B_aA_a = B_bA_b = \mu_bH_bA_b$. Therefore $H_b = H_a(\mu_aA_a/\mu_bA_b)$ and $H_a[d_a + d_b(\mu_aA_a/\mu_bA_b)] = NI$, so $\overline{H}_a = \hat{z}\,NI/[d_a + d_b(\mu_aA_a/\mu_bA_b)]$ [A m^{-1}].

4.5   *Laplace's equation and separation of variables*

4.5.1   Laplace's equation

Electric and magnetic fields obey Faraday's and Ampere's laws, respectively, and when the fields are static and the charge and current are zero we have:

$$\nabla \times \overline{E} = 0 \qquad (4.5.1)$$

$$\nabla \times \overline{H} = 0 \qquad (4.5.2)$$

These equations are satisfied by any $\overline{E}$ or $\overline{H}$ that can be expressed as the gradient of a potential:

$$\overline{E} = -\nabla \Phi \qquad (4.5.3)$$

$$\overline{H} = -\nabla \Psi \qquad (4.5.4)$$

Therefore Maxwell's equations for static charge-free regions of space are satisfied for any arbitrary differentiable potential function $\Phi(\overline{r})$ or $\Psi(\overline{r})$, which can be determined as discussed below.

Any potential function must be consistent with the given boundary conditions, and with Gauss's laws in static charge- and current-free spaces:

$$\nabla \bullet \overline{D} = 0 \qquad (4.5.5)$$

$$\nabla \bullet \overline{B} = 0 \qquad (4.5.6)$$

where $\overline{D} = \varepsilon \overline{E}$ and $\overline{B} = \mu \overline{H}$. Substituting (4.5.3) into (4.5.5), and (4.5.4) into (4.5.6) yields *Laplace's equation*:

$$\nabla^2 \Phi = \nabla^2 \Psi = 0 \qquad \text{(Laplace's equation)} \qquad (4.5.7)$$

To find static electric or magnetic fields produced by any given set of boundary conditions we need only to solve Laplace's equation (4.5.7) for Φ or Ψ, and then use (4.5.3) or (4.5.4) to compute the gradient of the potential. One approach to solving Laplace's equation is developed in the following section.

Example 4.5A
Does the potential $\Phi = 1/r$ satisfy Laplace's equation $\nabla^2 \Phi = 0$, where $r = (x^2 + y^2 + z^2)^{0.5}$?

Solution: $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$. First: $(\partial/\partial x)(x^2 + y^2 + z^2)^{-0.5} = -0.5(x^2 + y^2 + z^2)^{-1.5}(2x)$, so $(\partial^2/\partial x^2)(x^2 + y^2 + z^2)^{-0.5} = 0.75(x^2 + y^2 + z^2)^{-2.5}(2x)^2 - (x^2 + y^2 + z^2)^{-1.5}$. Therefore $(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)(x^2 + y^2 + z^2)^{-0.5} = 3(x^2 + y^2 + z^2)^{-2.5}(x^2 + y^2 + z^2) - 3(x^2 + y^2 + z^2)^{-1.5} = 0$. So this potential satisfies Laplace's equation. The algebra could have been simplified if instead we wrote ∇^2 in spherical coordinates (see Appendix C), because only the radial term is potentially non-zero for $\Phi = 1/r$: $\nabla^2 = r^{-2}(\partial/\partial r)(r^2 \partial/\partial r)$. In this case the right-most factor is $r^2 \partial r^{-1}/\partial r = r^2(-r^{-2}) = -1$, and $\partial(-1)/\partial r = 0$, so again $\nabla^2 \Phi = 0$.

4.5.2   Separation of variables

We can find simple analytic solutions to Laplace's equation only in a few special cases for which the solutions can be factored into products, each of which is dependent only upon a single dimension in some coordinate system compatible with the geometry of the given boundaries. This process of separating Laplace's equation and solutions into uni-dimensional factors is called *separation of variables*. It is most easily illustrated in terms of two dimensions. Let's assume the solution can be factored:

$$\Phi(x,y) = X(x)Y(y) \tag{4.5.8}$$

Then Laplace's equation becomes:

$$\nabla^2\Phi = \partial^2\Phi/\partial x^2 + \partial^2\Phi/\partial y^2 = Y(y)\,d^2X/dx^2 + X(x)\,d^2Y/dy^2 = 0 \tag{4.5.9}$$

Dividing by X(x)Y(y) yields:

$$\left[d^2X(x)/dx^2\right]/X(x) = -\left[d^2Y(y)/dy^2\right]/Y(y) \tag{4.5.10}$$

Since (4.5.10) must be true for all values of x, y, it follows that each term must equal a constant k^2, called the *separation constant*, so that:

$$d^2X/dx^2 = -k^2X \qquad\qquad d^2Y/dy^2 = k^2Y \tag{4.5.11}$$

Generic solutions to (4.5.11) are, for $k \neq 0$:

$$X(x) = A \cos kx + B \sin kx \tag{4.5.12}$$

$$Y(y) = C \cosh ky + D \sinh ky \tag{4.5.13}$$

An equivalent alternative is $Y(y) = C'\,e^{ky} + D'\,e^{-ky}$. Generic solutions when k = 0 are:

$$X(x) = Ax + B \tag{4.5.14}$$

$$Y(y) = Cy + D \tag{4.5.15}$$

Note that by letting $k \to jk$, the sinusoidal x-dependence becomes hyperbolic, and the hyperbolic y dependence becomes sinusoidal--the roles of x and y are reversed. Whether k is zero, real, imaginary, or complex depends upon boundary conditions. Linear combinations of solutions to differential equations are also solutions to those same equations, and such combinations are often required to match boundary conditions.

These univariable solutions can be combined to yield the three solution forms for x-y coordinates:

$$\Phi(x,y) = (A + Bx)(C + Dy) \qquad\qquad \text{for } k = 0 \quad (4.5.16)$$

$$\Phi(x,y) = (A \cos kx + B \sin kx)(C \cosh ky + D \sinh ky) \qquad\qquad \text{for } k^2 > 0 \quad (4.5.17)$$

$$\Phi(x,y) = (A \cosh qx + B \sinh qx)(C \cos qy + D \sin qy) \qquad \text{for } k^2 < 0 \ (k = jq) \quad (4.5.18)$$

This approach can be extended to three cartesian dimensions by letting $\Phi(x,y,z) = X(x)Y(y)Z(z)$; this leads to the solution[11]:

$$\Phi(x,y,z) = (A \cos k_x x + B \sin k_x x)(C \cos k_y y + D \sin k_y y)(E \cosh k_z z + F \sinh k_z z) \quad (4.5.19)$$

where $k_x^2 + k_y^2 + k_z^2 = 0$. Since k_x^2, k_y^2, and k_z^2 must sum to zero, k_i^2 must be negative for one or two coordinates so that the solution is sinusoidal along either one or two axes and hyperbolic along the others.

Once the form of the solution is established, the correct form, (4.5.16) to (4.5.19), is selected and the unknown constants are determined so that the solution matches the given boundary conditions, as illustrated in the following example.
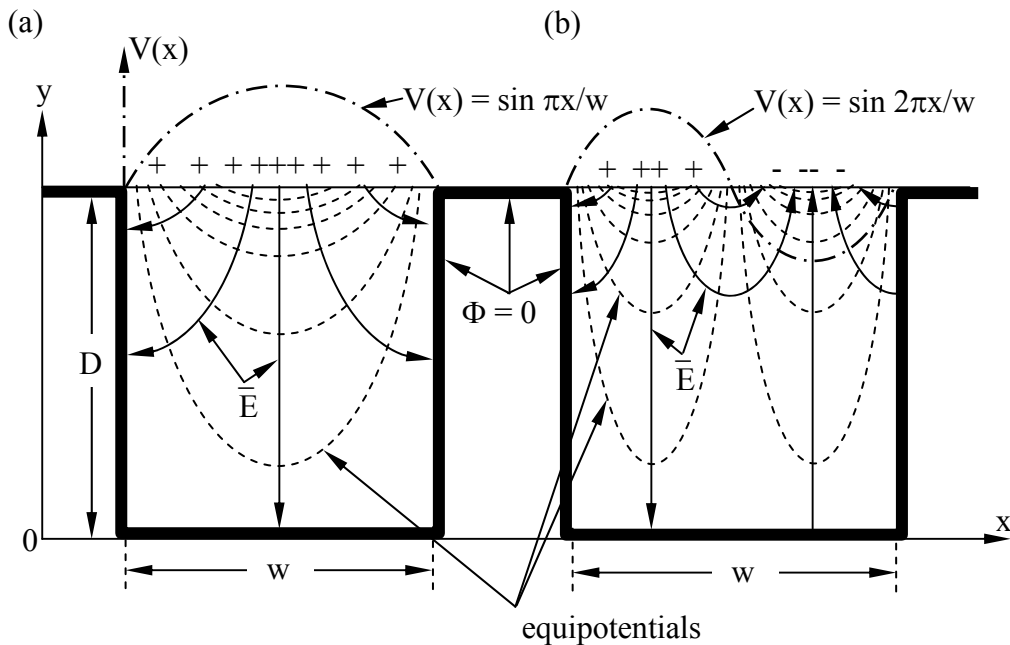
Figure 4.5.1   Static potentials and fields in a sinusoidally-driven conducting rectangular slot.

---

[11] If $\Phi(x,y,z) = X(x)Y(y)Z(z)$, then $\nabla^2\Phi = YZ d^2X/dx^2 + XZ d^2Y/dy^2 + XY d^2Z/dz^2$. Dividing by XYZ yields $X^{-1}d^2X/dx^2 + Y^{-1}d^2Y/dy^2 + Z^{-1}d^2Z/dz^2 = 0$, which implies all three terms must be constants if the equation holds for all x,y,z; let these constants be k_x^2, k_y^2, and k_z^2, respectively. Then $d^2X(y)/dx^2 = k_x^2 X(x)$, and the solution (4.5.19) follows when only $k_z^2 > 0$.

Consider an infinitely long slot of width w and depth d cut into a perfectly conducting slab, and suppose the cover to the slot has the voltage distribution $V(x) = 5 \sin(\pi x/w)$ volts, as illustrated in Figure 4.5.1(a). This is a two-dimensional cartesian-coordinate problem, so the solution (4.5.17) is appropriate, where we must ensure this expression yields potentials that have the given voltage across the top of the slot and zero potential over the side and bottom boundaries of the slot. Thus:

$$\Phi(x, y) = A \sin(\pi x/w) \sinh(\pi y/w) \quad [\text{volts}] \tag{4.5.20}$$

where the sine and sinh options[12] from (4.5.17) were chosen to match the given potentials on all four boundaries, and where $A = 5/\sinh(\pi D/w)$ in order to match the given potential across the top of the slot.

Figure 4.5.1(b) illustrates the solution for the case where the potential across the open top of the slot is given as $V(x) = \sin 2\pi x/w$. If an arbitrary voltage $V(x)$ is applied across the opening at the top of the slot, then a sum of sine waves can be used to match the boundary conditions.

Although all of these examples were in terms of static electric fields $\overline{E}$ and potentials Φ, they equally well could have been posed in terms of static $\overline{H}$ and magnetic potential Ψ; the forms of solutions for Ψ are identical.

Example 4.5B

A certain square region obeys $\nabla^2\Phi = 0$ and has $\Phi = 0$ along its two walls at $x = 0$ and at $y = 0$. $\Phi = V$ volts at the isolated corner $x = y = L$. Φ increases linearly from 0 to V along the other two walls. What are $\Phi(x,y)$ and $\overline{E}(x, y)$ within the square?

Solution: Separation of variables permits linear gradients in potentials in rectangular coordinates via (4.5.14) and (4.5.15), so the potential can have the form $\Phi = (Ax + B)(Cy + D)$ where $B = D = 0$ for this example. Boundary conditions are matched for $\Phi(x,y) = (V/L^2)xy$ [V]. It follows that: $\overline{E} = -\nabla\Phi = (V/L^2)(\hat{x}y + \hat{y}x)$.

4.5.3   Separation of variables in cylindrical and spherical coordinates

Laplace's equation can be separated only in four known coordinate systems: cartesian, cylindrical, spherical, and elliptical. Section 4.5.2 explored separation in cartesian coordinates, together with an example of how boundary conditions could then be applied to determine a total solution for the potential and therefore for the fields. The same procedure can be used in a few other coordinate systems, as illustrated below for cylindrical and spherical coordinates.

---

[12] $\sinh x = (e^x - e^{-x})/2$ and $\cosh x = (e^x + e^{-x})/2$.

When there is no dependence on the z coordinate, Laplace's equation in cylindrical coordinates reduces to circular coordinates and is:

$$\nabla^2 \Phi = r^{-1}(\partial/\partial r)(r\partial\Phi/\partial r) + r^{-2}(\partial^2\Phi/\partial\phi^2) = 0 \qquad (4.5.21)$$

Appendix C reviews the del operator in several coordinate systems. We again assume the solution can be separated:

$$\Phi = R(r)\Phi(\phi) \qquad (4.5.22)$$

Substitution of (4.5.22) into (4.5.21) and dividing by $R(r)\Phi(\phi)$ yields:

$$R^{-1}(d/dr)(r\ dR/dr) = -\Phi^{-1}(d^2\Phi/d\phi^2) = m^2 \qquad (4.5.23)$$

where m^2 is the separation constant.

The solution to (4.5.23) depends on whether m^2 is zero, positive, or negative:

$$\Phi(r,\phi) = [A + B\phi][C + D\ (\ln r)] \qquad \text{(for } m^2 = 0) \quad (4.5.24)$$

$$\Phi(r,\phi) = (A \sin m\phi + B \cos m\phi)(Cr^m + Dr^{-m}) \qquad \text{(for } m^2 > 0) \quad (4.5.25)$$

$$\Phi(r,\phi) = [A \sinh p\phi + B \cosh p\phi][C \cos(p \ln r) + D \sin(p \ln r)] \qquad \text{(for } m^2 < 0) \quad (4.5.26)$$

where A, B, C, and D are constants to be determined and $m \equiv jp$ for $m^2 < 0$.

A few examples of boundary conditions and the resulting solutions follow. The simplest case is a uniform field in the $+\hat{x}$ direction; the solution that matches these boundary conditions is (4.5.25) for $m = 1$:

$$\Phi(r,\phi) = Br \cos \phi \qquad (4.5.27)$$

Another simple example is that of a conducting cylinder of radius R and potential V. Then the potential inside the cylinder is V and that outside decays as ln r, as given by (1.3.12), when $m = C = 0$:

$$\Phi(r,\phi) = (V/\ln R)\ln r \qquad (4.5.28)$$

The electric field associated with this electric potential is:

$$\overline{E} = -\nabla\Phi = -\hat{r}\partial\Phi/\partial r = \hat{r}(V/\ln R)r^{-1} \qquad (4.5.29)$$

Thus $\overline{E}$ is radially directed away from the conducting cylinder if V is positive, and decays as r^{-1}.

A final interesting example is that of a dielectric cylinder perpendicular to an applied electric field $\overline{E} = \hat{x}E_0$. Outside the cylinder the potential follows from (4.5.25) for m = 1 and is:

$$\Phi(r,\phi) = -E_0 r \cos\phi + (AR/r)\cos\phi \qquad (4.5.30)$$

The potential inside can have no singularity at the origin and is:

$$\Phi(r,\phi) = -E_0(Br/R)\cos\phi \qquad (4.5.31)$$

which corresponds to a uniform electric field. The unknown constants A and B can be found by matching the boundary conditions at the surface of the dielectric cylinder, where both Φ and $\overline{D}$ must be continuous across the boundary between regions 1 and 2. The two linear equations for continuity ($\Phi_1 = \Phi_2$, and $\overline{D}_1 = \overline{D}_2$) can be solved for the two unknowns A and B. The electric fields for this case are sketched in Figure 4.5.2.
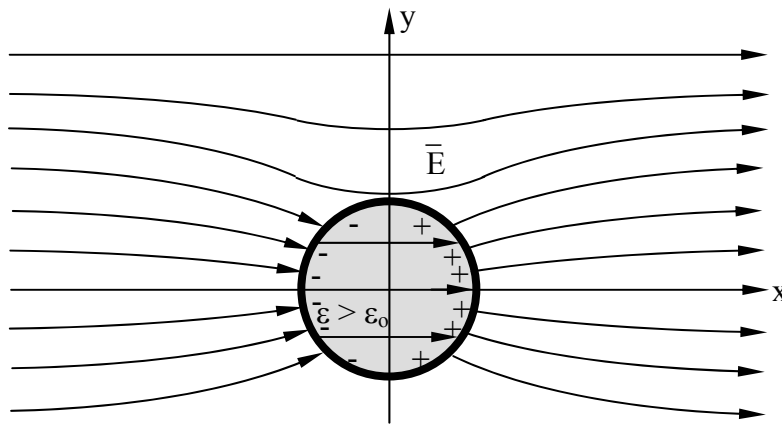
Figure 4.5.2   Electric fields perpendicular to a dielectric cylinder.

If these cylindrical boundary conditions also vary with z, the solution to Laplace's equation becomes:

$$\Phi(r,\phi,z) = \Phi_0[C_1 e^{kz} + C_2 e^{-kz}][C_3 \cos n\phi + C_4 \sin n\phi][C_5 J_n(kr) + C_6 N_n(kr)] \qquad (4.5.32)$$

where J_n and N_n are Bessel functions of order n of the first and second kind, respectively, and C_i are dimensionless constants that match the boundary conditions. The rapidly growing complexity of these solutions as the dimensionality of the problem increases generally mandates numerical solutions of such boundary value problems in practical cases.

Our final example involves spherical coordinates, for which the solutions are:

$$\Phi(r,\theta,\phi) = \Phi_0[C_1 r^n + C_2 r^{-n-1}][C_3 \cos m\phi + C_4 \sin m\phi][C_5 P_n^m(\cos\theta) + C_6 Q_n^m(\cos\theta)] \qquad (4.5.33)$$

where P_n^m and Q_n^m are associated Legendre functions of the first and second kind, respectively, and C_i are again dimensionless constants chosen to match boundary conditions. Certain spherical problems do not invoke Legendre functions, however, as illustrated below.

A dielectric sphere inserted in a uniform electric field $\hat{x} E_o$ exhibits the same general form of solution as does the dielectric rod perpendicular to a uniform applied electric field; the solution is the sum of the applied field and the dipole field produced by the induced polarization charges on the surface of the rod or sphere. Inside the sphere the field is uniform, as suggested in Figure 4.5.2. Polarization charges are discussed more fully in Section 2.5.3. The potential follows from (4.5.33) with n = 1 and m = 0, and is simply:

$$\Phi(r,\theta,\phi) = -E_o \cos\theta \left(C_1 r - C_2 R^3 r^{-2} \right) \tag{4.5.34}$$

where $C_2 = 0$ inside, and for the region outside the cylinder C_2 is proportional to the induced electric dipole. C_1 outside is unity and inside diminishes below unity as ε increases.

If the sphere in the uniform electric field is conducting, then in (4.5.34) $C_1 = C_2 = 0$ inside the sphere, and the field there is zero; the surface charge is:

$$\rho_s = -\varepsilon_o \hat{n} \bullet \nabla\Phi \big|_{r=R} = \varepsilon_o E_r = 3\varepsilon_o E_o \cos\theta \left[Cm^{-2} \right] \tag{4.5.35}$$

Outside the conducting sphere $C_1 = 1$, and to ensure $\Phi(r = R) = 0$, C_2 must also be unity.

The same considerations also apply to magnetic potentials. For example, a sphere of permeability μ and radius R placed in a uniform magnetic field would also have an induced magnetic dipole that produces a uniform magnetic field inside, and produces outside the superposition of the original uniform field with a magnetic dipole field produced by the sphere. A closely related example involves a sphere of radius R having surface current:

$$\overline{J}_s = \hat{\phi} \sin\theta \left[Am^{-1} \right] \tag{4.5.36}$$

This can be produced approximately by a coil wound on the surface of the sphere with a constant number of turns per unit length along the z axis.

For a permeable sphere in a uniform magnetic field $\overline{H} = -\hat{z} H_o$, the solution to Laplace's equation for magnetic potential $\nabla^2 \Psi = 0$ has a form similar to (4.5.34):

$$\Psi(r,\theta) = Cr\cos\theta \qquad\qquad \text{(inside the sphere; } r < R) \quad (4.5.37)$$

$$\Psi(r,\theta) = Cr^{-2}\cos\theta + H_o r\cos\theta \qquad\qquad \text{(outside the sphere; } r > R) \quad (4.5.38)$$

Using $\overline{H} = -\nabla\Psi$, we obtain:

$$\overline{H}(r,\theta) = -\hat{z}C \qquad\qquad\qquad\qquad \text{(inside the sphere; } r < R) \quad (4.5.39)$$

$$\overline{H}(r,\theta) = -C(R/r)^2 (\hat{r}\cos\theta + 0.5\hat{\theta}\sin\theta) - \hat{z}H_o \qquad \text{(outside the sphere; } r > R) \quad (4.5.40)$$

Matching boundary conditions at the surface of the sphere yields C; e.g. equate $\overline{B} = \mu\overline{H}$ inside to $\overline{B} = \mu_o H$ outside by equating (4.5.39) to (4.5.40) for $\theta = 0$.

4.6   Flux tubes and field mapping

4.6.1   <u>Static field flux tubes</u>

Flux tubes are arbitrarily designated bundles of static electric or magnetic field lines in charge-free regions, as illustrated in Figure 4.6.1.
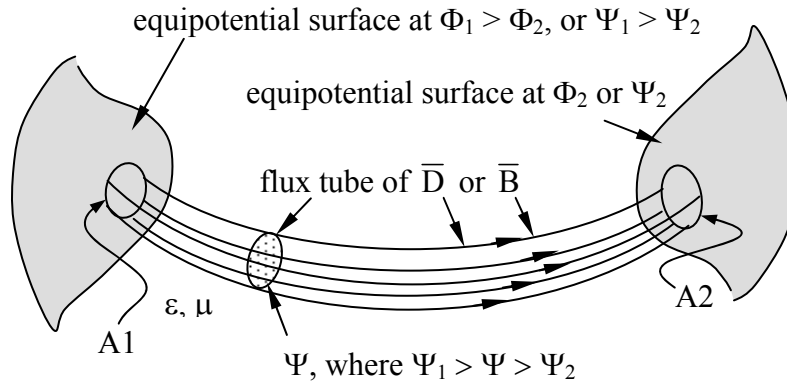
Figure 4.6.1   Electric or magnetic flux tube between two equipotential surfaces.

The divergence of such static fields is zero by virtue of Gauss's laws, and their curl is zero by virtue of Faraday's and Ampere's laws. The integral forms of Gauss's laws, (2.4.17) and (2.4.18), say that the total electric displacement $\overline{D}$ or magnetic flux $\overline{B}$ crossing the surface A of a volume V must be zero in a charge-free region:

$$\oiint_A (\overline{D}\bullet\hat{n})\,da = 0 \qquad\qquad\qquad\qquad (4.6.1)$$

$$\oiint_A (\overline{B}\bullet\hat{n})\,da = 0 \qquad\qquad\qquad\qquad (4.6.2)$$

Therefore if the walls of flux tubes are parallel to the fields then the walls contribute nothing to the integrals (4.6.1) and (4.6.2) and the total flux entering the area A1 of the flux tube at one end (A1) must equal that exiting through the area A2 at the other end, as illustrated:

$$\oiint_{A1} (\overline{D}\bullet\hat{n})\,da = -\oiint_{A2} (\overline{D}\bullet\hat{n})\,da \qquad\qquad\qquad (4.6.3)$$

$$\oiint_{A1} (\overline{B} \bullet \hat{n}) \, da = -\oiint_{A2} (\overline{B} \bullet \hat{n}) \, da \qquad (4.6.4)$$

Consider two surfaces with potential differences between them, as illustrated in Figure 4.6.1. A representative flux tube is shown and all other fields are omitted from the figure. The field lines could correspond to either $\overline{D}$ or $\overline{B}$. Constant ε and μ are not required for $\overline{D}$ and $\overline{B}$ flux tubes because $\overline{D}$ and $\overline{B}$ already incorporate the effects of inhomogeneous media. If the permittivity ε and permeability μ were constant then the figure could also apply to $\overline{E}$ or $\overline{H}$, respectively.

4.6.2 Field mapping

$\overline{E}$ and $\overline{H}$ are gradients of the potentials Φ and Ψ, respectively [see (4.6.2) and (4.6.5)], and therefore the equipotential surfaces are perpendicular to their corresponding fields, as suggested in Figure 4.6.1. This orthogonality leads to a useful technique called *field mapping* for sketching approximately correct field distributions given arbitrarily shaped surfaces at known potentials. The method is particularly simple for "two-dimensional" geometries that depend only on the x,y coordinates and are independent of z, such as the pair of circular surfaces illustrated in Figure 4.6.2(a) and the pair of ovals in Figure 4.6.2(b). Assume that the potential of the inner surface is Φ_1 or Ψ_1, and that at the outer surface is Φ_2 or Ψ_2.
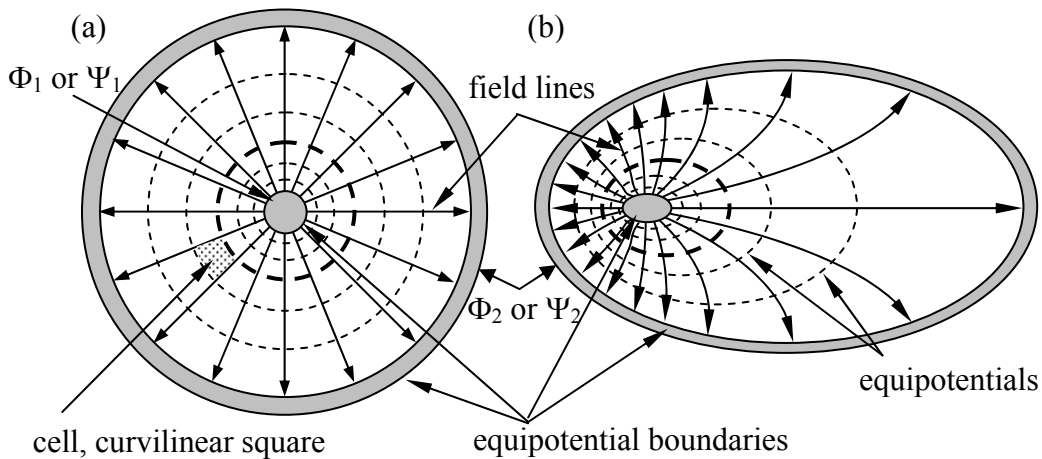
Figure 4.6.2   Field mapping of static electric and magnetic fields.

Because: 1) the lateral spacing between adjacent equipotential surfaces and (in two-dimensional geometries) between adjacent field lines are both inversely proportional to the local field strength, and 2) the equipotentials and field lines are mutually orthogonal, it follows that the rectangular shape of the cells formed by these adjacent lines is preserved over the field even as the field strengths and cell sizes vary. That is, the curvilinear square illustrated in Figure 4.6.2(a) has approximately the same shape (but not size) as all other cells in the figure, and approaches a perfect square as the cells are subdivided indefinitely. If sketched perfectly, any two-

dimensional static potential distribution can be subdivided indefinitely into such curvilinear square cells.

One algorithm for performing such a subdivision is to begin by sketching a first-guess equipotential surface that: 1) separates the two (or more) equipotential boundaries and 2) is orthogonal to the first-guess field lines, which also can be sketched. These field lines must be othogonal to the equipotential boundaries. For example, this first sketched surface might have potential $(\Phi_1 + \Phi_2)/2$, where Φ_1 and Φ_2 are the applied potentials. The spacing between the initially sketched field lines and between the initial equipotential surfaces should form approximate curvilinear squares. Each such square can then be subdivided into four smaller curvilinear squares using the same algorithm. If the initial guesses were correct, then the curvilinear squares approach true squares when infinitely subdivided. If they do not, the first guess is revised appropriately and the process can be repeated until the desired insight or perfection is achieved. In general there will be some fractional squares arranged along one of the field lines, but these become negligible in the limit.

Figure 4.6.2(a) illustrates how the flux tubes in a co-axial geometry are radial with field strength inversely proportional to radius. Therefore, when designing systems limited by the maximum allowable field strength, one avoids incorporating surfaces with small radii of curvature or sharp points. Figure 4.6.2(b) illustrates how the method can be adapted to arbitrarily shaped boundaries, albeit with more difficulty. Computer-based algorithms using relaxation techniques can implement such strategies rapidly for both two-dimensional and three-dimensional geometries. In three dimensions, however, the spacing between field lines varies inversely with the square root of their strength, and so the height-to-width ratio of the curvilinear 3-dimensional rectangles formed by the field lines and potentials is not preserved across the structure.

Chapter 5: Electromagnetic Forces

5.1 Forces on free charges and currents

5.1.1 Lorentz force equation and introduction to force

The Lorentz force equation (1.2.1) fully characterizes electromagnetic forces on stationary and moving charges. Despite the simplicity of this equation, it is highly accurate and essential to the understanding of all electrical phenomena because these phenomena are observable only as a result of forces on charges. Sometimes these forces drive motors or other actuators, and sometimes they drive electrons through materials that are heated, illuminated, or undergoing other physical or chemical changes. These forces also drive the currents essential to all electronic circuits and devices.

When the electromagnetic fields and the location and motion of free charges are known, the calculation of the forces acting on those charges is straightforward and is explained in Sections 5.1.2 and 5.1.3. When these charges and currents are confined within conductors instead of being isolated in vacuum, the approaches introduced in Section 5.2 can usually be used. Finally, when the charges and charge motion of interest are bound within stationary atoms or spinning charged particles, the Kelvin force density expressions developed in Section 5.3 must be added. The problem usually lies beyond the scope of this text when the force-producing electromagnetic fields are not given but are determined by those same charges on which the forces are acting (e.g., plasma physics), and when the velocities are relativistic.

The simplest case involves the forces arising from known electromagnetic fields acting on free charges in vacuum. This case can be treated using the *Lorentz force equation* (5.1.1) for the *force vector* $\bar{f}$ acting on a charge q [Coulombs]:

$$\bar{f} = q\left(\bar{E} + \bar{v} \times \mu_o \bar{H}\right) \quad \text{[Newtons]} \qquad \text{(Lorentz force equation)} \qquad (5.1.1)$$

where $\bar{E}$ and $\bar{H}$ are the local electric and magnetic fields and $\bar{v}$ is the charge velocity vector [m s^{-1}].

5.1.2 Electric Lorentz forces on free electrons

The *cathode-ray tube* (CRT) used for displays in older computers and television sets, as illustrated in Figure 5.1.1, provides a simple example of the Lorentz force law (5.1.1). Electrons thermally excited by a heated *cathode* at -V volts escape at low energy and are accelerated in vacuum at acceleration $\bar{a}$ [m s^{-2}] toward the grounded *anode* by the electric field $\bar{E} \cong -\hat{z}V/s$ between anode and cathode[13]; V and s are the voltage across the tube and the cathode-anode

---

[13] The anode is grounded for safety reasons; it lies at the tube face where users may place their fingers on the other side of the glass faceplate. Also, the cathode and anode are sometimes shaped so that the electric field $\bar{E}$, the force $\bar{f}$, and the acceleration $\bar{a}$ are functions of z instead of being constant; i.e., $\bar{E} \neq -\hat{z}V/D$.

separation, respectively. In electronics the anode always has a more positive potential Φ than the cathode, by definition.
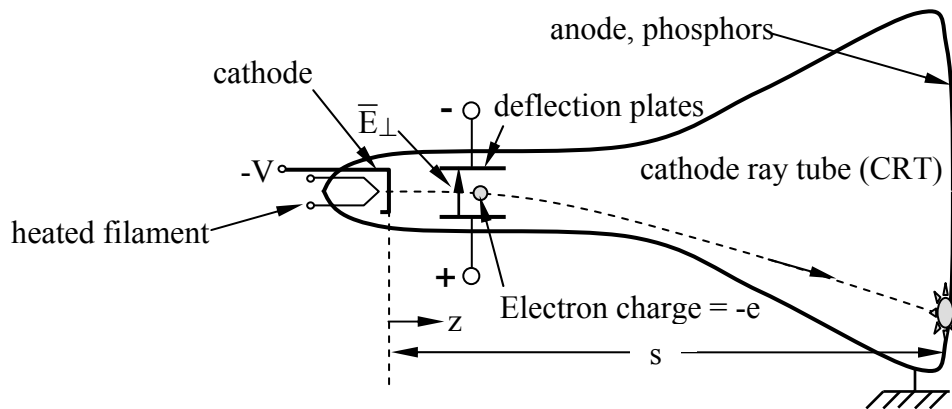
Figure 5.1.1   Cathode ray tube.

The acceleration $\overline{a}$ is governed by *Newton's law*:

$$\overline{f} = m\overline{a} \qquad \text{(Newton's law)} \qquad (5.1.2)$$

where m is the mass of the unconstrained accelerating particle. Therefore the acceleration a of the electron charge q = -e in an electric field E = V/s is:

$$a = f/m = qE/m \cong eV/ms \quad [ms^{-2}] \qquad (5.1.3)$$

The subsequent velocity $\overline{v}$ and position z of the particle can be found by integration of the acceleration $\hat{z}a$:

$$\overline{v} = \int_0^t \overline{a}(t)dt = \overline{v}_0 + \hat{z}at \quad [ms^{-1}] \qquad (5.1.4)$$

$$z = z_0 + \hat{z} \bullet \int_0^t \overline{v}(t)dt = z_0 + \hat{z} \bullet \overline{v}_0 t + at^2/2 \quad [m] \qquad (5.1.5)$$

where we have defined the initial electron position and velocity at t = 0 as z_0 and $\overline{v}_0$, respectively.

The increase w_k in the kinetic energy of the electron equals the accumulated work done on it by the electric field $\overline{E}$. That is, the increase in the kinetic energy of the electron is the product of the constant force f acting on it and the distance s the electron moved in the direction of $\overline{f}$ while experiencing that force. If s is the separation between anode and cathode, then:

$$w_k = fs = (eV/s)s = eV \quad [J] \qquad (5.1.6)$$

- 128 -

Thus the kinetic energy acquired by the electron in moving through the potential difference V is eV Joules. If V = 1 volt, then w_k is one "*electron volt*", or "e" Joules, where e $\cong 1.6 \times 10^{-19}$ Coulombs. The kinetic energy increase equals eV even when $\overline{E}$ is a function of z because:

$$w_k = \int_0^D eE_z dz = eV \tag{5.1.7}$$

Typical values for V in television CRT's are generally less than 50 kV so as to minimize dangerous x-rays produced when the electrons impact the phosphors on the CRT faceplate, which is often made of x-ray-absorbing leaded glass.

Figure 5.1.1 also illustrates how time-varying lateral electric fields $\overline{E}_\perp(t)$ can be applied by deflection plates so as to scan the electron beam across the CRT faceplate and "paint" the image to be displayed. At higher tube voltages V the electrons move so quickly that the lateral electric forces have no time to act, and magnetic deflection is used instead because lateral magnetic forces increase with electron velocity v.

Example 5.1A

Long interplanetary or interstellar voyages might eject charged particles at high speeds to obtain thrust. What particles are most efficient at imparting total momentum if the rocket has only E joules and M kg available to expend for this purpose?

Solution: Particles of charge e accelerating through an electric potential of V volts acquire energy eV [J] = $mv^2/2$; such energies can exceed those available in chemical reactions. The total increase in rocket momentum = nmv [N], where n is the total number of particles ejected, m is the mass of each particle, and v is their velocity. The total mass and energy available on the rocket is M = nm and E = neV. Since v = $(2eV/m)^{0.5}$, the total momentum ejected is Mv = nmv = $(n^2 2eVm)^{0.5}$ = $(2EM)^{0.5}$. Thus any kind of charged particles can be ejected, only the total energy E and mass ejected M matter.

5.1.3 Magnetic Lorentz forces on free charges

An alternate method for laterally scanning the electron beam in a CRT utilizes magnetic deflection applied by coils that produce a magnetic field perpendicular to the electron beam, as illustrated in Figure 5.1.2. The magnetic Lorentz force on the charge q = -e (1.6021×10^{-19} Coulombs) is easily found from (5.1.1) to be:

$$\overline{f} = -e\overline{v}\times\mu_o\overline{H} \quad [N] \tag{5.1.8}$$

Thus the illustrated CRT electron beam would be deflected upwards, where the magnetic field $\overline{H}$ produced by the coil is directed out of the paper; the magnitude of the force on each electron is $ev\mu_o H$ [N].
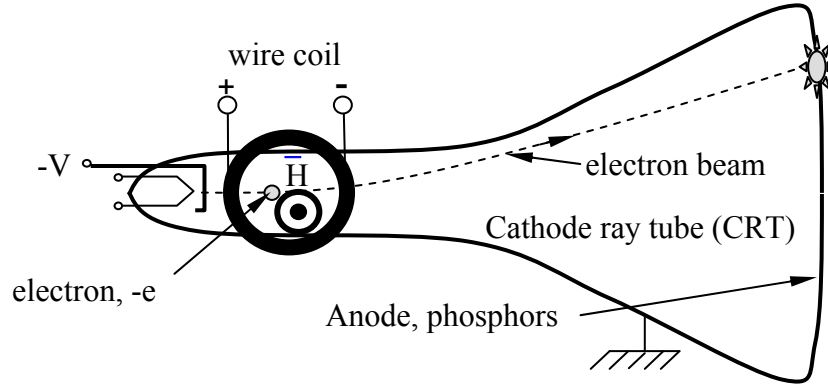
Figure 5.1.2   Magnetic deflection of electrons in a cathode ray tube.

The lateral force on the electrons $ev\mu_o H$ can be related to the CRT voltage V. Electrons accelerated from rest through a potential difference of V volts have kinetic energy eV [J], where:

$$eV = mv^2/2 \qquad (5.1.9)$$

Therefore the electron velocity $v = (2eV/m)^{0.5}$, where m is the electron mass (9.107×10^{-31} kg), and the lateral deflection increases with tube voltage V, whereas it decreases if electrostatic deflection is used instead.

Another case of magnetic deflection is illustrated in Figure 5.1.3 where a free electron moving perpendicular to a magnetic field $\overline{B}$ experiences a force $\overline{f}$ orthogonal to its velocity vector $\overline{v}$, since $\overline{f} = q\overline{v} \times \mu_o \overline{H}$.
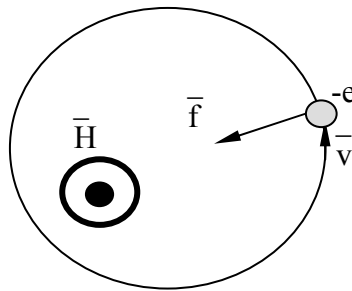
Figure 5.1.3   Cyclotron motion of an electron.

This force $|\overline{f}|$ is always orthogonal to $\overline{v}$ and therefore the trajectory of the electron will be circular with radius R at angular frequency ω_e [radians s^{-1}]:

$$|f| = ev\mu_o H = m_e a = m_e \omega_e^2 R = m_e v \omega_e \qquad (5.1.10)$$

where $v = \omega_e R$. We can solve (5.1.9) for this "electron *cyclotron frequency*" ω_e:

$$\omega_e = e\mu_o H/m_e \qquad\qquad \text{(electron cyclotron frequency)} \qquad (5.1.11)$$

which is independent of v and the electron energy, provided the electron is not relativistic. Thus the magnitudes of magnetic fields can be measured by observing the radiation frequency ω_e of free electrons in the region of interest.

Example 5.1B

What is the radius r_e of cyclotron motion for a 100 e.v. free electron in the terrestrial magnetosphere[14] where B $\cong 10^{-6}$ Tesla? What is the radius r_p for a free proton with the same energy? The masses of electrons and protons are ~9.1×10^{-31} and 1.7×10^{-27} kg, respectively.

Solution: The magnetic force on a charged particle is $qv\mu_o H = ma = mv^2/r$, where the velocity v follows from (5.1.9): $eV = mv^2/2 \Rightarrow v = (2eV/m)^{0.5}$. Solving for r_e yields $r_e = m_e v/e\mu_o H = (2Vm/e)^{0.5}/\mu_o H \cong (2\times100\times9.1\times10^{-31}/1.6\times10^{-19})^{0.5}/10^{-6} \cong 34$ m for electrons and ~2.5 km for protons.

5.2 Forces on charges and currents within conductors

5.2.1 Electric Lorentz forces on charges within conductors

Static electric forces on charges within conductors can also be calculated using the Lorentz force equation (5.1.1), which becomes $\overline{f} = q\overline{E}$. For example, consider the capacitor plates illustrated in Figure 5.2.1(a), which have total surface charges of ±Q coulombs on the two conductor surfaces facing each other. The fields and charges for capacitor plates were discussed in Section 3.1.3.
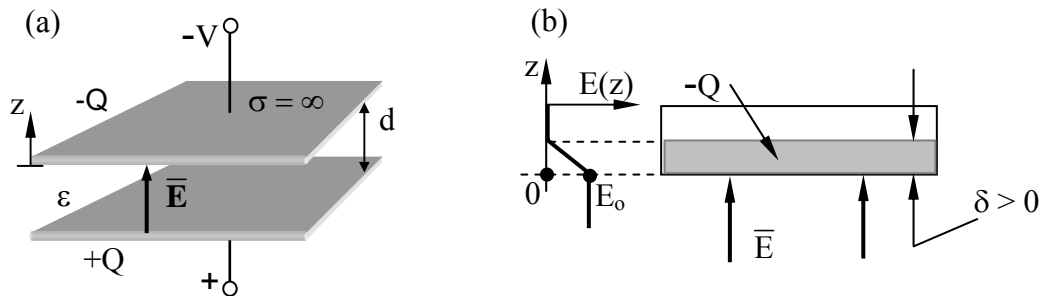
Figure 5.2.1   Charge distribution within conducting capacitor plates.

To compute the total attractive *electric pressure* P_e [N m^{-2}] on the top plate, for example, we can integrate the Lorentz force density $\overline{F}$ [N m^{-3}] acting on the charge distribution $\rho(z)$ over depth z and unit area:

---

[14] The magnetosphere extends from the ionosphere to several planetary radii; particle collisions are rare compared to the cyclotron frequency.

$$\overline{F} = \rho \overline{E} \quad [N \ m^{-3}] \tag{5.2.1}$$

$$\overline{P}_e = \int_0^\infty \overline{F}(z)dz = \hat{z}\int_0^\infty \rho(z)E_z(z)dz \quad [N \ m^{-2}] \tag{5.2.2}$$

where we have defined $\overline{E} = \hat{z}E_z$, as illustrated.

Care is warranted, however, because surface the charge $\rho(z)$ is distributed over some infinitesimal depth δ, as illustrated in Figure 5.2.1(b), and those charges at greater depths are shielded by the others and therefore see a smaller electric field $\overline{E}$. If we assume $\varepsilon = \varepsilon_o$ inside the conductors and a planar geometry with $\partial/\partial x = \partial/\partial y = 0$, then Gauss's law, $\nabla \bullet \varepsilon \overline{E} = \rho$, becomes:

$$\varepsilon_o dE_z/dz = \rho(z) \tag{5.2.3}$$

This expression for $\rho(z)$ can be substituted into (5.2.2) to yield the pressure exerted by the electric fields on the capacitor plate and perpendicular to it:

$$P_e = \int_{E_o}^0 \varepsilon_o dE_z E_z = -\varepsilon_o E_o^2 / 2 \qquad \text{(electric pressure on conductors)} \tag{5.2.4}$$

The charge density ρ and electric field E_z are zero at levels below δ, and the field strength at the surface is E_o. If the conductor were a dielectric with $\varepsilon \neq \varepsilon_o$, then the Kelvin polarization forces discussed in Section 5.3.2 would also have to be considered.

Thus the electric pressure P_e [N m^{-2}] pulling on a charged conductor is the same as the immediately adjacent electric energy density [Jm^{-3}], and is independent of the sign of ρ and $\overline{E}$. These dimensions are identical because [J] = [Nm]. The maximum achievable electric field strength thus limits the maximum achievable electric pressure P_e, which is negative because it pulls rather than pushes conductors.

An alternate form for the electric pressure expression is:

$$P_e = -\varepsilon_o E_o^2 / 2 = -\rho_s^2 / 2\varepsilon_o \quad [Nm^{-2}] \text{ (electric pressure on conductors)} \tag{5.2.5}$$

where ρ_s is the surface charge density [cm^{-2}] on the conductor and ε_o is its permittivity; boundary conditions at the conductor require $D = \varepsilon_o E = \sigma_s$. Therefore if the conductor were adjacent to a dielectric slab with $\varepsilon \neq \varepsilon_o$, the electrical pressure on the conductor would still be determined by the surface charge, electric field, and permittivity ε_o within the conductor; the pressure does not otherwise depend on ε of adjacent rigid materials.

We can infer from (5.2.4) the intuitively useful result that the average electric field pulling on the charge Q is E/2 since the total pulling force f = - P_eA, where A is the area of the plate:

$$f = -P_e A = A\varepsilon_o E^2/2 = AD(E/2) = Q(E/2) \tag{5.2.6}$$

If the two plates were both charged the same instead of oppositely, the surface charges would repel each other and move to the outer surfaces of the two plates, away from each other. Since there would now be no E between the plates, it could apply no force. However, the charges Q on the outside are associated with the same electric field strength as before, $E = Q/\varepsilon_o A$. These electric fields outside the plates therefore pull them apart with the same force density as before, $P_e = -\varepsilon_o E^2/2$, and the force between the two plates is now repulsive instead of attractive. In both the attractive and repulsive cases we have assumed the plate width and length are sufficiently large compared to the plate separation that fringing fields can be neglected.

Example 5.2A
Some copy machines leave the paper electrically charged. What is the electric field E between two adjacent sheets of paper if they cling together electrically with a force density of 0.01 oz. $\cong$ 0.0025 N per square centimeter = 25 Nm^{-2}? If we slightly separate two such sheets of paper by 4 cm, what is the voltage V between them?

Solution: Electric pressure is $P_e = -\varepsilon_o E^2/2$ [N m^{-2}], so $E = (-2P_e/\varepsilon_o)^{0.5} = (2\times25/8.8\times10^{-12})^{0.5} = 2.4$ [MV/m]. At 4 cm distance this field yields ~95 kV potential difference between the sheets. The tiny charge involved renders this voltage harmless.

5.2.2   Magnetic Lorentz forces on currents in conductors

The Lorentz force law can also be used to compute forces on electrons moving within conductors for which $\mu = \mu_o$. Computation of forces for the case $\mu \neq \mu_o$ is treated in Sections 5.3.3 and 5.4. If there is no net charge and no current flowing in a wire, the forces on the positive and negative charges all cancel because the charges comprising matter are bound together by strong inter- and intra-atomic forces.
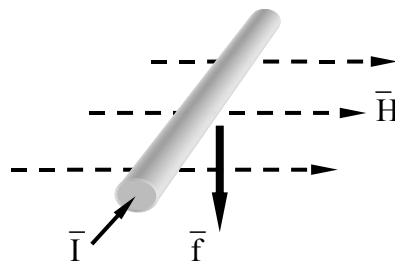
Figure 5.2.2   Magnetic force on a current-carrying wire.

However, if n_1 carriers per meter of charge q are flowing in a wire[15], as illustrated in Figure 5.2.2, then the total force density $\overline{F} = n_1 q \overline{v} \times \mu_o \overline{H} = \overline{I} \times \mu_o \overline{H}$ [N m^{-1}] exerted by a static magnetic field $\overline{H}$ acting on the static current $\overline{I}$ flowing in the wire is:

$$\overline{F} = n_1 q \overline{v} \times \mu_o \overline{H} = \overline{I} \times \mu_o \overline{H} \quad \left[\text{N m}^{-1} \right] \qquad \text{(magnetic force density on a wire)} \qquad (5.2.7)$$

where $\overline{I} = n_1 q \overline{v}$. If $\overline{H}$ is uniform, this force is not a function of the cross-section of the wire, which could be a flat plate, for example.
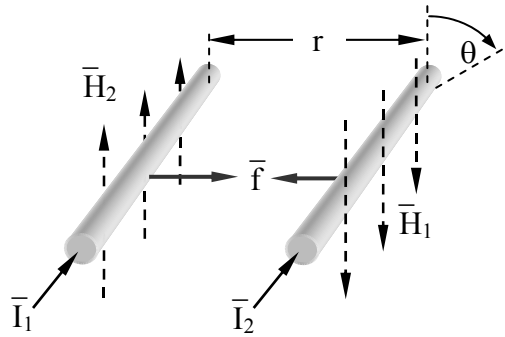
Figure 5.2.3   Magnetic forces attracting parallel currents.

We can easily extend the result of (5.2.7) to the case of two parallel wires carrying the same current I in the same $+\hat{z}$ direction and separated by distance r, as illustrated in Figure 5.2.3. Ampere's law with cylindrical symmetry readily yields $\overline{H}(r) = \hat{\theta} H(r)$:

$$\oint_c \overline{H} \bullet d\overline{s} = I = 2\pi r H \Rightarrow H = I/2\pi r \qquad (5.2.8)$$

The force density $\overline{F}$ pulling the two parallel wires together is then found from (5.2.7) and (5.2.8) to be:

$$|\overline{F}| = \mu_o I^2 / 2\pi r \quad \left[\text{Nm}^{-1} \right] \qquad (5.2.9)$$

The simplicity of this equation and the ease of measurement of F, I, and r led to its use in defining the permeability of free space, $\mu_o = 4\pi \times 10^{-7}$ Henries/meter, and hence the definition of a *Henry* (the unit of inductance). If the two currents are in opposite directions, the force acting on the wires is repulsive. For example, if I = 10 amperes and r = 2 millimeters, then (5.2.9) yields F = $4\pi \times 10^{-7} \times 10^2/2\pi 2 \times 10^{-3}$ = 0.01 Newtons/meter; this is approximately the average repulsive force between the two wires in a 120-volt AC lamp cord delivering one kilowatt. These forces are attractive when the currents are parallel, so if we consider a single wire as consisting of

---

[15] The notation n_j signifies number density [m^{-j}], so n_1 and n_3 indicate numbers per meter and per cubic meter, respectively.

parallel strands, they will squeeze together due to this *pinch effect*. At extreme currents, these forces can actually crush wires, so the maximum achievable instantaneous current density in wires is partly limited by their mechanical strength. The same effect can pinch electron beams flowing in charge-neutral plasmas.

The magnetic fields associated with surface currents on flat conductors generally exert a pressure $\overline{P}$ [N m^{-2}] that is simply related to the instantaneous field strength $\left|\overline{H}_s\right|$ at the conductor surface. First we can use the magnetic term in the Lorentz force law (5.2.7) to compute the force density $\overline{F}$ [N m^{-3}] on the surface current $\overline{J}_s$ [A m^{-1}]:

$$\overline{F} = nq\overline{v} \times \mu_o \overline{H} = \overline{J} \times \mu_o \overline{H} \quad \left[\text{N m}^{-3}\right]$$ (5.2.10)

where n is the number of charges q per cubic meter. To find the *magnetic pressure* P_m [N m^{-2}] on the conductor we must integrate the force density $\overline{F}$ over depth z, where both $\overline{J}$ and $\overline{H}$ are functions of z, as governed by Ampere's law in the static limit:

$$\nabla \times \overline{H} = \overline{J}$$ (5.2.11)

If we assume $\overline{H} = \hat{y}H_y(z)$ then $\overline{J}$ is in the x direction and $\partial H_y/\partial x = 0$, so that:

$$\nabla \times \overline{H} = \hat{x}\left(\partial H_z/\partial y - \partial H_y/\partial z\right) + \hat{y}\left(\partial H_x/\partial z - \partial H_z/\partial x\right) + \hat{z}\left(\partial H_y/\partial x - \partial H_x/\partial y\right)$$
$$= -\hat{x}dH_y/dz = \hat{x}J_x(z)$$ (5.2.12)

The instantaneous magnetic pressure P_m exerted by H can now be found by integrating the force density equation (5.2.10) over depth z to yield:

$$\overline{P}_m = \int_0^\infty \overline{F}\ dz = \int_0^\infty \overline{J}(z) \times \mu_o \overline{H}(z)\ dz = \int_0^\infty \left[-\hat{x}dH_y/dz\right] \times \left[\hat{y}\mu_o H_y(z)\right]dz$$
$$\overline{P}_m = -\hat{z}\mu_o \int_H^0 H_y\ dH_y = \hat{z}\mu_o H^2/2 \quad \left[\text{Nm}^{-2}\right] \qquad \text{(magnetic pressure)} \quad (5.2.13)$$

We have assumed $\overline{H}$ decays to zero somewhere inside the conductor. As in the case of the electrostatic pull of an electric field on a charged conductor, the average field strength experienced by the surface charges or currents is half that at the surface because the fields inside the conductor are partially shielded by any overlying charges or currents. The time average magnetic pressure for sinusoidal H is $\langle P_m \rangle = \mu_o \left|\overline{H}\right|^2/4$.

5.3   *Forces on bound charges within materials*

5.3.1   Introduction

Forces on materials can be calculated in three different ways: 1) via the Lorentz force law, as illustrated in Section 5.2 for free charges within materials, 2) via energy methods, as illustrated in Section 5.4, and 3) via photonic forces, as discussed in Section 5.6. When polarized or magnetized materials are present, as discussed here in Section 5.3, the Lorentz force law must be applied not only to the free charges within the materials, i.e. the surface charges and currents discussed earlier, but also to the orbiting and spinning charges bound within atoms. When the Lorenz force equation is applied to these bound charges, the result is the Kelvin polarization and magnetization force densities. Under the paradigm developed in this chapter these Kelvin forces must be added to the Lorentz forces on the free charges[16]. The Kelvin force densities are non-zero only when inhomogeneous fields are present, as discussed below in Sections 5.3.2 and 5.3.3. But before discussing Kelvin forces it is useful to review the relationship between the Lorentz force law and matter.

The Lorentz force law is complete and exact if we ignore relativistic issues associated with either extremely high velocities or field strengths; neither circumstance is relevant to current commercial products. To compute all the Lorentz forces on matter we must recognize that classical matter is composed of atoms comprised of positive and negative charges, some of which are moving and exhibit magnetic moments due to their spin or orbital motions. Because these charges are trapped in the matter, any forces on them are transferred to that matter, as assumed in Section 5.2 for electric forces on surface charges and for magnetic forces on surface currents.

When applying the Lorentz force law within matter under our paradigm it is important to use the expression:

$$\overline{f} = q\left(\overline{E} + \overline{v} \times \mu_o \overline{H}\right) \quad [\text{Newtons}] \tag{5.3.1}$$

without substituting $\mu \overline{H}$ for the last term when $\mu \neq \mu_o$. A simple example illustrates the dangers of this common notational shortcut. Consider the instantaneous magnetic pressure (5.2.13) derived using the Lorentz force law for a uniform plane wave normally incident on a conducting plate having $\mu \neq \mu_o$. The same force is also found later in (5.6.5) using photon momentum. If we incorrectly use:

$$\overline{f} = q\left(\overline{E} + \overline{v} \times \mu \overline{H}\right) \quad [\text{Newtons}] \qquad \text{(incorrect for this example)} \tag{5.3.2}$$

---

[16] The division here between Lorentz forces acting on free charges and the Lorentz forces acting on bound charges (often called Kelvin forces) is complete and accurate, but not unique, for these forces can be grouped and labeled differently, leading to slightly different expressions that are also correct.

because $\overline{v}$ occurs within μ, then the computed wave pressure would increase with μ, whereas the photon model has no such dependence and yields $P_m = \mu_o H^2/2$, the same answer as does (5.2.13). The photon model depends purely on the input and output photon momentum fluxes observed some distance from the mirror, and thus the details of the mirror construction are irrelevant once the fraction of photons reflected is known.

This independence of the Lorentz force from μ can also be seen directly from the Lorentz force calculation that led to (5.2.13). In this case the total surface current is not a function of μ for a perfect reflector, and neither is $\overline{H}$ just below the surface; they depend only upon the incident wave and the fact that the mirror is nearly perfect. $\overline{H}$ does decay faster with depth when μ is large, as discussed in Section 9.3, but the average $|\overline{H}|$ experienced by the surface-current electrons is still half the value of $\overline{H}$ at the surface, so $\overline{f}$ is unchanged as μ varies. The form of the Lorentz force law presented in (5.3.2) can therefore be safely used under our force paradigm only when $\mu = \mu_o$, although the magnetic term is often written as $\overline{v} \times \overline{B}$.

There are alternate correct paradigms that use μ in the Lorentz law rather than μ_o, but they interpret Maxwell's equations slightly differently. These alternative approaches are not discussed here.

The Lorentz force law can also be applied to those cases where non-uniform fields pull on dielectrics or permeable materials, as suggested by Figure 5.3.1. These problems are often more easily solved, however, using energy (Section 5.4) or pressure (Section 5.5) methods. To compute in general the forces on matter exerted by non-uniform electric or magnetic fields we can derive the Kelvin polarization and magnetization force density expressions from the Lorentz equation, as shown in Sections 5.3.2 and 5.3.3, respectively.
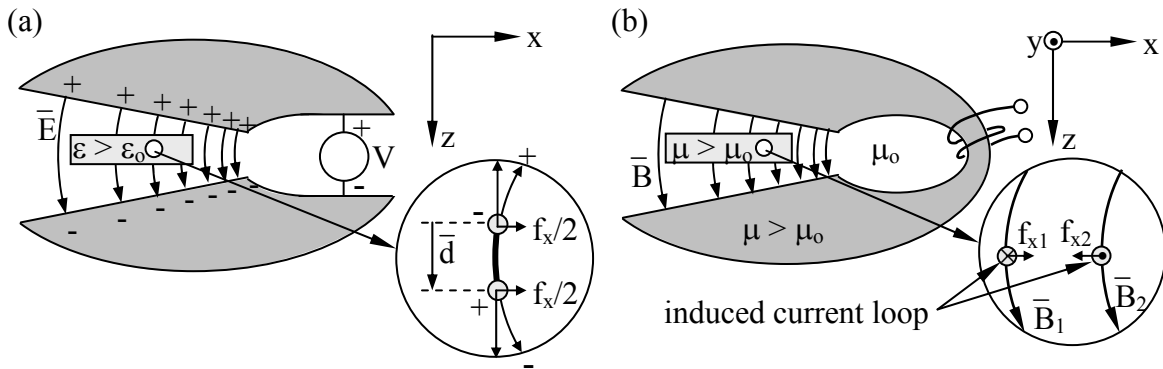
Figure 5.3.1 Kelvin polarization and magnetization forces on materials.

The derivations of the Kelvin force density expressions are based on the following simple models for charges in matter. Electric Lorentz forces act on atomic nuclei and the surrounding electron clouds that are bound together, and on any free charges. The effect of $\overline{E}$ on positive and negative charges bound within an atom is to displace their centers slightly, inducing a small electric dipole. The resulting atomic electric dipole moment is:

$$\overline{p} = \overline{d}q \qquad \text{(Coulomb meters)} \qquad (5.3.3)$$

where $\overline{d}$ is the displacement vector [m] pointing from the negative charge center to the positive charge center for each atom, and q is the atomic charge or atomic number. As discussed further in Section 5.3.2, Kelvin polarization forces result when the field gradients cause the electric field lines to curve slightly so that the directions of the electric Lorentz forces are slightly different for the two ends of the field-induced electric dipoles so they do not cancel exactly, leaving a net residual force.

The magnetic Lorentz forces act on electrons classically orbiting atomic nuclei with velocities $\overline{v}$, and act on electrons with classical charge densities spinning at velocity $\overline{v}$ about the electron spin axis. Protons also spin, and therefore both electrons and protons possess magnetic dipole moments; these spin moments are smaller than those due to electron orbital motion. If we consider these spin and orbital motions as being associated with current loops, then we can see that the net force on such a loop would be non-zero if the magnetic fields perpendicular to these currents were different on the two sides of the loop. Such differences exist when the magnetic field has a non-zero gradient and then Kelvin magnetization forces result, as discussed in Section 5.3.3. The electromagnetic properties of matter are discussed further in Sections 2.5 and 9.5.

5.3.2   Kelvin polarization force density

Kelvin polarization forces result when a non-zero electric field gradient causes the Lorentz electric forces on the two charge centers of each induced electric dipole in a dielectric to differ, as illustrated in Figure 5.3.1(a). The force density can be found by summing the force imbalance vectors for each dipole within a unit volume.

Assume the center of the negative charge -q for a particular atom is at $\overline{r}$, and the center of the positive charge +q is at $\overline{r} + \overline{d}$. Then the net electric Lorentz force on that atom in the x direction is:

$$f_x = q\left[E_x\left(\overline{r} + \overline{d}\right) - E_x\left(\overline{r}\right)\right] = q\left(\overline{d} \bullet \nabla E_x\right) \quad [N] \qquad (5.3.4)$$

Thus $\overline{f}_x$ is the projection of the charge offset $\overline{d}$ on the gradient of qE_x. We recall $\nabla \equiv \hat{x}\,\partial/\partial x + \hat{y}\,\partial/\partial y + \hat{z}\,\partial/\partial z.$

Equation (5.3.4) can easily be generalized to:

$$\overline{f} = \hat{x}\left(q\overline{d} \bullet \nabla E_x\right) + \hat{y}\left(q\overline{d} \bullet \nabla E_y\right) + \hat{z}\left(q\overline{d} \bullet \nabla E_z\right) \qquad (5.3.5)$$

$$= \hat{x}\left(\overline{p} \bullet \nabla E_x\right) + \hat{y}\left(\overline{p} \bullet \nabla E_y\right) + \hat{z}\left(\overline{p} \bullet \nabla E_z\right) \equiv \overline{p} \bullet \nabla \overline{E} \quad [N] \qquad (5.3.6)$$

where $\overline{p} = q\overline{d}$ and (5.3.6) defines the new compact notation $\overline{p} \bullet \nabla \overline{E}$. Previously we have defined only $\nabla \times \overline{E}$ and $\nabla \bullet \overline{E}$, and the notation $\overline{p} \bullet [.]$ would have implied a scalar, not a vector. Thus the new operator defined here is $[\bullet \nabla]$, and it operates on a pair of vectors to produce a vector.

Equation (5.3.6) then yields the Kelvin polarization force density $\overline{F}_p = n\overline{f}$, where n is the density of atomic dipoles $[m^{-3}]$, and the polarization density of the material $\overline{P}$ is $n\overline{p}$ $[C\ m^{-2}]$:

$$\overline{F}_p = \overline{P} \bullet \nabla \overline{E} \quad \left[N\ m^{-3} \right] \qquad \text{(Kelvin polarization force density)} \qquad (5.3.7)$$

Equation (5.3.7) states that electrically polarized materials are pulled into regions having stronger electric fields if there is polarization $\overline{P}$ in the direction of the gradient. Less obvious from (5.3.7) is the fact that there can be such a force even when the applied electric field $\overline{E}$ and $\overline{P}$ are orthogonal to the field gradient, as illustrated in Figure 5.3.1(a). In this example a z-polarized dielectric is drawn in the x direction into regions of stronger E_z. This happens in curl-free fields because then a non-zero $\partial E_z/\partial x$ implies a non-zero $\partial E_x/\partial z$ that contributes to $\overline{F}_p$. This relation between partial derivatives follows from the definition:

$$\nabla \times \overline{E} = 0 = \hat{x}\left(\partial E_z/\partial y - \partial E_y/\partial z\right) + \hat{y}\left(\partial E_x/\partial z - \partial E_z/\partial x\right) + \hat{z}\left(\partial E_y/\partial x - \partial E_x/\partial y\right) \qquad (5.3.8)$$

Since each cartesian component must equal zero, it follows that $\partial E_x/\partial z = \partial E_z/\partial x$ so both these derivatives are non-zero, as claimed. Note that if the field lines $\overline{E}$ were not curved, then $f_x = 0$ in Figure 5.3.1. But such fields with a gradient $\nabla E_z \neq 0$ would have non-zero curl, which would require current to flow in the insulating region.

The polarization $\overline{P} = \overline{D} - \varepsilon_o \overline{E} = (\varepsilon - \varepsilon_o)\overline{E}$, as discussed in Section 2.5.3. Thus, in free space, dielectrics with $\varepsilon > \varepsilon_o$ are always drawn into regions with higher field strengths while dielectrics with $\varepsilon < \varepsilon_o$ are always repulsed. The same result arises from energy considerations; the total energy w_e decreases as a dielectric with permittivity ε greater than that of its surrounding ε_o moves into regions having greater field strength $\overline{E}$.

Example 5.3A
What is the Kelvin polarization force density $\overline{F}_p$ $[N\ m^{-3}]$ on a dielectric of permittivity $\varepsilon = 3\varepsilon_o$ in a field $\overline{E} = \hat{z}E_o(1 + 5z)$?

Solution: (5.3.7) yields $F_{pz} = \overline{P} \bullet (\nabla E_z) = (\varepsilon - \varepsilon_o)\overline{E} \bullet \hat{z}5E_o = 10\varepsilon_o E_o{}^2$ $\left[N\ m^{-3} \right]$.

5.3.3   Kelvin magnetization force density

Magnetic dipoles are induced in permeable materials by magnetic fields. These induced magnetic dipoles arise when the applied magnetic field slightly realigns the randomly oriented

pre-existing magnetic dipoles associated with electron spins and electron orbits in atoms. Each such induced magnetic dipole can be modeled as a small current loop, such as the one pictured in Figure 5.3.1(b) in the x-y plane. The collective effect of these induced atomic magnetic dipoles is a permeability μ that differs from μ_o, as discussed further in Section 2.5.4. Prior to realignment of the magnetic dipoles in a magnetizable medium by an externally applied $\overline{H}$, their orientations are generally random so that their effects cancel and can be neglected.

Kelvin magnetization forces on materials result when a non-zero magnetic field gradient causes the Lorentz magnetic forces on the two current centers of each induced magnetic dipole to differ so they no longer cancel, as illustrated in Figure 5.3.1(b). The magnified portion of the figure shows a typical current loop in cross-section where the magnetic Lorentz forces f_{x1} and f_{x2} are unbalanced because $\overline{B}_1$ and $\overline{B}_2$ differ. The magnetic flux density $\overline{B}_1$ acts on the current flowing in the -y direction, and the magnetic field $\overline{B}_2$ acts on the equal and opposite current flowing in the +y direction. The force density $\overline{F}_m$ can be found by summing the net force vectors for every such induced magnetic dipole within a unit volume. This net force density pulls a medium with $\mu > \mu_o$ into the high-field region.

The current loops induced in magnetic materials such as iron and nickel tend to increase the applied magnetic field $\overline{H}$, as illustrated, so that the permeable material in the figure has $\mu > \mu_o$ and experiences a net force that tends to move it toward more intense magnetic fields. That is why magnets attract iron and any *paramagnetic material* that has $\mu > \mu_o$, while repulsing any *diamagnetic material* for which the induced current loops have the opposite polarity so that $\mu < \mu_o$. Although most ordinary materials are either paramagnetic or diamagnetic with $\mu \cong \mu$, only ferromagnetic materials such as iron and nickel have $\mu >> \mu_o$ and are visibly affected by ordinary magnets.

An expression for the Kelvin magnetization force density $\overline{F}_m$ can be derived by calculating the forces on a square current loop of I amperes in the x-y plane, as illustrated. The Lorentz magnetic force on each of the four legs is:

$$\overline{f}_i = \overline{I} \times \mu_o \overline{H} w \quad [N] \tag{5.3.9}$$

where i = 1,2,3,4, and w is the length of each leg. The sum of these four forces is:

$$\begin{aligned} \overline{f} &= Iw^2 \mu_o \left[\left(\hat{y} \times \partial \overline{H}/\partial x \right) - \left(\hat{x} \times \partial \overline{H}/\partial y \right) \right] \quad [N] \\ &= Iw^2 \mu_o \left[-\hat{z} \left(\partial H_x/\partial x + \partial H_y/\partial y \right) + \hat{x} \left(\partial H_z/\partial x \right) + \hat{y} \left(\partial H_z/\partial y \right) \right] \end{aligned} \tag{5.3.10}$$

This expression can be simplified by noting that $\overline{m} = \hat{z} Iw^2$ is the magnetic dipole moment of this current loop, and that $\partial H_x/\partial x + \partial H_y/\partial y = - \partial H_z/\partial z$ because $\nabla \bullet \overline{H} = 0$, while $\partial H_z/\partial x = \partial H_x/\partial z$ and $\partial H_z/\partial y = \partial H_y/\partial z$ because $\nabla \times \overline{H} = 0$ in the absence of macroscopic currents. Thus for the geometry of Figure 5.3.1(b), where $\overline{m}$ is in the z direction, the magnetization force of Equation (5.3.10) becomes:

$$\overline{f}_m = \mu_o m_z \left(\hat{z} \partial H_z / \partial z + \hat{x} \partial H_x / \partial z + \hat{y} \partial H_y / \partial z \right) = \mu_o m_z \partial \overline{H} / \partial z \qquad (5.3.11)$$

This expression can be generalized to cases where $\overline{m}$ is in arbitrary directions:

$$\overline{f}_m = \mu_o \left(m_x \partial \overline{H} / \partial x + m_y \partial \overline{H} / \partial y + m_z \partial \overline{H} / \partial z \right) = \mu_o \overline{m} \bullet \nabla \overline{H} \;\; [\text{N}] \qquad (5.3.12)$$

where the novel notation $\overline{m} \bullet \nabla \overline{H}$ was defined in (5.3.6).

Equation (5.3.12) then yields the Kelvin magnetization force density $\overline{F}_m = n_3 \overline{f}$, where n_3 is the equivalent density of magnetic dipoles [m^{-3}], and the magnetization $\overline{M}$ of the material is $n_3 \overline{m}$ [A m^{-1}]:

$$\overline{F}_m = \mu_o \overline{M} \bullet \nabla \overline{H} \;\; \left[\text{N m}^{-3} \right] \qquad \text{(Kelvin magnetization force density)} \qquad (5.3.13)$$

Such forces exist even when the applied magnetic field $\overline{H}$ and the magnetization $\overline{M}$ are orthogonal to the field gradient, as illustrated in Figure 5.3.1(b). As in the case of Kelvin polarization forces, this happens in curl-free fields because then a non-zero $\partial H_z / \partial x$ implies a non-zero $\partial H_x / \partial z$ that contributes to $\overline{F}_m$.

5.4 Forces computed using energy methods

5.4.1 Relationship between force and energy

Mechanics teaches that a force f in the z direction pushing an object a distance dz expends energy dw = f dz [J], so:

$$f = dw/dz \qquad\qquad\qquad \textit{(force/energy equation)} \qquad (5.4.1)$$

Therefore the net force f_{be} applied <u>by</u> the <u>e</u>nvironment to any object in the z direction can be found simply by differentiating the total system energy w with respect to motion of that object in the direction z. The total force vector $\overline{f}_{be}$ is the sum of its x, y, and z components.

Care must be taken, however, to ensure that the <u>total</u> system energy is differentiated, which can include the energy in any connected power supplies, mechanical elements, etc. Care must also be taken to carefully distinguish between forces f_{be} exerted <u>by</u> the <u>e</u>nvironment, and forces f_{oe} exerted by objects <u>on</u> their <u>e</u>nvironment; otherwise sign errors are readily introduced. This simple powerful approach to finding forces is illustrated in Section 5.4.2 for electrostatic forces and in Section 5.6 for photonic forces. The energy approach to calculating magnetic forces uses (5.4.1) in a straightforward way, but examples are postponed to Chapter 6 when magnetic fields in structures will be better understood.

Example 5.4A

A certain perfectly conducting electromagnet carrying one ampere exerts an attractive 100-N force f on a piece of iron while it moves away from the magnet at velocity v = 1 [m s^{-1}]. What voltage V is induced across the terminals of the electromagnet as a result of this velocity v? Is this voltage V positive or negative on that terminal where the current enters the magnet? Use conservation of power.

Solution: Conservation of power requires fv = VI, so V = fv/I = 100×1/1 = 100 volts. The voltage is negative because the magnet is acting as a generator since the motion of the iron is opposite to the magnetic force acting on it.

5.4.2   Electrostatic forces on conductors and dielectrics

The energy method easily yields the force f_{be} needed to separate in the z direction the two isolated capacitor plates oppositely charged with Q in vacuum and illustrated in Figure 5.2.1(a). Since the plates are attracted to one another, separating them does work and increases the stored energy w. The force needed to hold the plates apart is easily found using the force/energy equation (5.4.1):

$$f_{be} = dw/dz = d\left(Q^2 s/2\varepsilon_o A\right)/ds = Q^2/2\varepsilon_o A \quad [\text{N}]$$ (5.4.2)

where the plate separation is s and the plate area is A. The electric energy w_e stored in a capacitor C is $CV^2/2 = Q^2/2C = Q^2 s/2\varepsilon A$, where Q = CV and C = εA/s, as shown in Section 3.1.3. Here we assumed ε = ε$_o$.

The derivative in (5.4.2) was easy to evaluate because Q remains constant as the disconnected plates are forced apart. It would be incorrect to use w = $CV^2/2$ when differentiating (5.4.2) unless we recognize that V increases as the plates separate because V = Q/C when C decreases. It is easier to express energy in terms of parameters that remain constant as z changes.

We can put (5.4.2) in the more familiar form (5.2.4) for the electric pressure P_e pushing on a conductor by noting that the force f_{be} needed to separate the plates is the same as the electric force attracting the oppositely charged plates. The force f_{be} thus balances the electric pressure on the same plates and $P_e = -f_{be}/A$. Since Q = ε$_o$EA here we find:

$$P_e = -Q^2/2\varepsilon_o A^2 = -\varepsilon_o E^2/2 \quad \left[\text{Nm}^{-2}\right]$$ (5.4.3)

This static attractive pressure of electric fields remains the same if the plates are connected to a battery of voltage V instead of being isolated; the Lorentz forces are the same in both cases. A more awkward way to calculate the same force (5.4.2) is to assume (unnecessarily) that a battery is connected and that V remains constant as s changes. In this case Q must vary with dz, and dQ flows into the battery, increasing its energy by VdQ. Since dw in the force/energy expression (5.4.2) is the change in total system energy, the changes in both battery and electric

field energy must be calculated to yield the correct energy; an example with a battery begins later with (5.4.5). As illustrated above, this complexity can be avoided by carefully restating the problem without the source, and by expressing w in terms of electrical variables (Q here) that do not vary with position (s here).

The power of the energy method (5.4.1) is much more evident when calculating the force $\overline{f}$ needed to pull two capacitor plates apart laterally, as illustrated in Figure 5.4.1(a). To use the Lorentz force law directly would require knowledge of the lateral components of $\overline{E}$ responsible for the lateral forces, but they are not readily determined. Since energy derivatives can often be computed accurately and easily (provided the fringing fields are relatively small), that is often the preferred method for computing electric and magnetic forces.
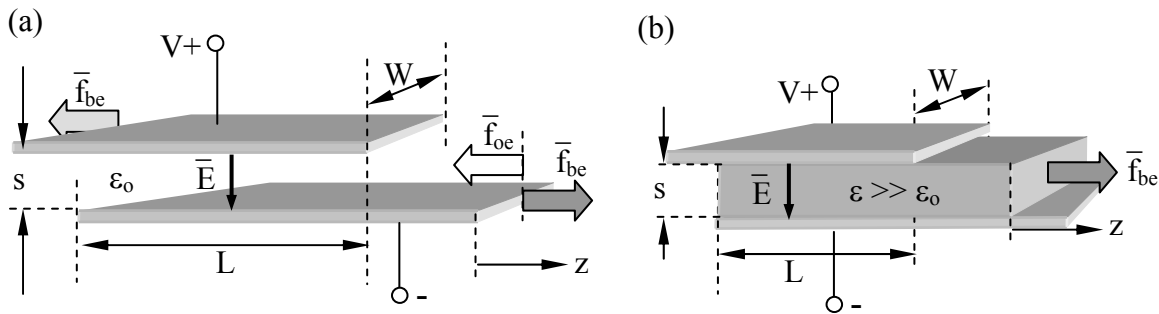
Figure 5.4.1 Capacitor plates and dielectrics being separated laterally.

The force/energy equation (5.4.1) can be expressed in terms of the area A = WL of the capacitor. Because L decreases as z increases, the sign of the derivative with respect to the plate overlap L is negative, and the force exerted on the plates by the environment is:

$$f_{be} = dw/dz = -d\left(Q^2 s/2\varepsilon_o WL\right)/dL = Q^2 s/2\varepsilon_o WL^2 \ [N] \tag{5.4.4}$$

where $dz = dL$ and $w_e = Q^2 s/2WL$. We again assumed that the plates were isolated in space so Q was constant, but the same force results when the plates are attached to a battery; in both cases the Lorentz forces arise from the very same charges so the two forces must be identical.

For purposes of illustration, let's solve the force/energy equation (5.4.1) for the same problem of Figure 5.4.1 the more difficult way by including the increase in battery energy as z increases. The incremental work $f_{be}dz$ involved in pulling the plates apart a distance dz is:

$$f_{be} = dw_T/dz = -d\left(\varepsilon_o WLV^2/2s\right)/dL - VdQ/dz \tag{5.4.5}$$

where w_T is the total energy and the two terms on the right-hand side of (5.4.5) reflect the energy changes in the capacitor and battery respectively. The first negative sign in (5.4.5) arises because the overlap distance L decreases as z increases, and the second negative sign arises because the battery energy increases as Q decreases.

Since only L and Q vary with L, where $Q = CV = \varepsilon_o WLV/s$, (5.4.5) becomes:

$$f_{be} = -\varepsilon_o WV^2/2s + \varepsilon_o WV^2/s = \varepsilon_o WV^2/2s \quad [N] \qquad (5.4.6)$$

where the sign of the second term ($\varepsilon_o WV^2/s$) reverses because Q decreases as z increases. This result when including the battery is the same as (5.4.4) without the battery, which can be seen by using $V = Q/C$ and $C = \varepsilon_o WL/s$:

$$f_{be} = \varepsilon_o WV^2/2s = Q^2 s/2\varepsilon_o WL^2 \quad [N] \qquad (5.4.7)$$

If the space between and surrounding the conducting plates were filled with a fluid having $\varepsilon > \varepsilon_o$, then for fixed V both the stored electric energy w_e and dw_e/dz, together with the force f_{be}, would obviously be increased by a factor of $\varepsilon/\varepsilon_o$ so that in this case the lateral force f_{be} would equal $\varepsilon WV^2/2s$.

Note that approximately the same force f_{be} is required to separate laterally two capacitor plates, one of which is coated with a dielectric having permittivity ε, as illustrated in Figure 5.4.1(b), because the force/energy equation (5.4.4) is largely unchanged except that $\varepsilon_o \rightarrow \varepsilon$:

$$f_{be} = dw/dz = -d\left(Q^2 s/2\varepsilon WL\right)/dL = Q^2 s/2\varepsilon WL^2 \ [N] \qquad (5.4.8)$$

5.5   *Electric and magnetic pressure*

5.5.1   Electromagnetic pressures acting on conductors

Forces on materials can be computed in several different ways, all of which can be derived using Maxwell's equations and the Lorentz force law. The pressure method for computing forces arising from static fields is useful because it expresses prior results in ways that are easy to evaluate and remember, and that have physical significance. The method simply notes that the electromagnetic force density (pressure) acting on the interface between two materials equals the difference in the electromagnetic energy densities on the two sides of the interface. Both energy density [J m^{-3}] and pressure [N m^{-2}] have identical units because [J] = [N m].

For example, both the Lorentz force law and the energy method yield the same expression, (5.2.4) and (5.4.3) respectively, for the *electric pressure* P_e due to a static electric field E pushing on a conductor:

$$P_e = -\varepsilon_o E^2/2 \left[N\ m^{-2}\right] \qquad \text{(electric pressure on conductors)} \qquad (5.5.1)$$

The Lorentz force law yields a similar expression (5.2.13) for the *magnetic pressure* pushing on a conductor:

$$P_m = \mu_o H^2 / 2 \; \left[\text{N m}^{-2} \right] \qquad \text{(magnetic pressure on conductors)} \qquad (5.5.2)$$

Thus motor and actuator forces are limited principally by the ability of material systems to sustain large static fields without breaking down in some way. Because large magnetic systems can sustain larger energy densities than comparable systems based on electric fields, essentially all large motors, generators, and actuators are magnetic. Only for devices with gaps on the order of a micron or less is the electrical breakdown field strength sufficiently high that electrostatic and magnetic motors compete more evenly with respect to power density, as discussed in Section 6.2.5.

5.5.2 Electromagnetic pressures acting on permeable and dielectric media

The Kelvin polarization and magnetization force densities, (5.3.7) and (5.3.13) respectively, can also be expressed in terms of energy densities and pressures. First we recall that $\overline{D} = \varepsilon \overline{E} = \varepsilon_o \overline{E} + \overline{P}$, so $\overline{P} = (\varepsilon - \varepsilon_o)\overline{E}$. Then it follows from (5.3.7) that the Kelvin polarization force density is:

$$\overline{F}_p = \overline{P} \bullet \nabla \overline{E} = (\varepsilon - \varepsilon_o)\overline{E} \bullet \nabla \overline{E} \; \left[\text{N m}^{-3} \right] \qquad (5.5.3)$$

The special operator $[\bullet \nabla]$ is defined in (5.3.6) and explained in (5.5.4). The x component of force density for a curl-free electric field $\overline{E}$ is:

$$F_{px} = \overline{P} \bullet (\nabla E_x) = (\varepsilon - \varepsilon_o)\overline{E} \bullet \nabla E_x = (\varepsilon - \varepsilon_o)(E_x \partial/\partial x + E_y \partial/\partial y + E_z \partial/\partial z)E_x \qquad (5.5.4)$$

$$= (\varepsilon - \varepsilon_o)(E_x \partial E_x / \partial x + E_y \partial E_y / \partial x + E_z \partial E_z / \partial x) \qquad (5.5.5)$$

$$= (\varepsilon - \varepsilon_o)\left(E_x \partial E_x^2 / \partial x + E_y \partial E_y^2 / \partial x + E_z \partial E_z^2 / \partial x \right) \Big/ 2 = (\varepsilon - \varepsilon_o)\left(\partial |\overline{E}|^2 / \partial x \right) \Big/ 2 \qquad (5.5.6)$$

In obtaining (5.5.5) we have used (5.3.8) for a curl-free electric field, for which $\partial E_x/\partial y = \partial E_y/\partial x$ and $\partial E_x/\partial z = \partial E_z/\partial x$.

Equations similar to (5.5.6) can be derived for the y and z components of the force density, which then add:

$$\overline{F}_p = (\varepsilon - \varepsilon_o)\nabla |\overline{E}|^2 \Big/ 2 \; \left[\text{N m}^{-3} \right] \qquad \textit{(Kelvin polarization force density)} \qquad (5.5.7)$$

A similar derivation applies to the Kelvin magnetization force density $\overline{F}_m$. We begin by recalling $\overline{B} = \mu_o \overline{M} \bullet \nabla \overline{H}$, so $\overline{M} = \left[(\mu/\mu_o) - 1 \right]\overline{H}$. Then it follows from (5.3.13) that the Kelvin magnetization force density is:

$$\overline{F}_m = \mu_o \overline{M} \bullet \nabla \overline{H} = (\mu - \mu_o)\overline{H} \bullet \nabla \overline{H} \ \left[N \ m^{-3}\right] \tag{5.5.8}$$

Repeating the steps of (5.5.4–7) yields for curl-free magnetic fields the parallel result:

$$\overline{F}_m = (\mu - \mu_o)\nabla|\overline{H}|^2\big/2 \ \left[N \ m^{-3}\right] \quad (\textit{Kelvin magnetization force density}) \tag{5.5.9}$$

Note that these force density expressions depend only on the field magnitudes $|\overline{E}|$ and $|\overline{H}|$, not on field directions.

Two examples treated in Chapter 6 using energy methods suggest the utility of simple pressure equations. Figure 6.2.4 shows a parallel-plate capacitor with a dielectric slab that fits snugly between the plates but that is only partially inserted in the z direction a distance D that is much less than the length L of both the slab and the capacitor plates. The electric field between the plates is $\overline{E}$, both inside and outside the dielectric slab. The total force on the dielectric slab is the integral of the Kelvin polarization force density (5.5.7) over the volume V of the slab, where V = LA and A is the area of the endface of the slab. We find from (5.5.7) that $\overline{F}_p$ is in the $\hat{z}$ direction and is non-zero only near the end of the capacitor plates where z = 0:

$$f_z = A \int_0^D F_{pz} \, dz = A\left[(\varepsilon - \varepsilon_o)/2\right]\int_0^D \left(d|\overline{E}|^2\big/dz\right)dz = A(\varepsilon - \varepsilon_o)|\overline{E}|^2\big/2 \ [N] \tag{5.5.10}$$

The integral is evaluated between the limit z = 0 where $E \cong 0$ outside the capacitor plates, and the maximum value z = D where the electric field between the plates is $\overline{E}$. Thus the pressure method yields the total force f_z on the dielectric slab; it is the area A of the end of the slab, times the electric pressure $(\varepsilon - \varepsilon_o)|\overline{E}|^2\big/2$ [N m^{-2}] at the end of the slab that is pulling the slab further between the plates. This pressure is zero at the other end of the slab because $\overline{E} \cong 0$ there. This pressure is the same as will be found in (6.2.21) using energy methods.

The second example is illustrated in Figure 6.4.1, where a snugly fitting cylindrical iron slug of area A has been pulled a distance D into a solenoidal coil that produces an axial magnetic field H. As in the case of the dielectric slab, one end of the slug protrudes sufficiently far from the coil that H at that end is approximately zero. The force pulling on the slug is easily found from (5.5.9):

$$f_z = A \int_0^D F_{mz} \, dz = A\left[(\mu - \mu_o)/2\right]\int_0^D \left(d|\overline{H}|^2\big/dz\right)dz = A(\mu - \mu_o)|\overline{H}|^2\big/2 \ [N] \tag{5.5.11}$$

This is more exact than the answer found in (6.4.10), where the μ_o term was omitted in (6.4.10) when the energy stored in the air was neglected.

To summarize, the static electromagnetic pressure [N m^{-2}] acting on a material interface with either free space or mobile liquids or gases is the difference between the two electromagnetic energy densities [J m^{-3}] on either side of that interface, provided that the relevant $\overline{E}$ and $\overline{H}$ are curl-free. In the case of dielectric or magnetic media, the pressure on the material is directed away from the greater energy density. In the case of conductors, external magnetic fields press on them while electric fields pull; the energy density inside the conductor is zero in both cases because $\overline{E}$ and $\overline{H}$ are presumed to be zero there.

Note that the pressure method for calculating forces on interfaces is numerically correct even when the true physical locus of the force may lie elsewhere. For example, the Kelvin polarization forces for a dielectric slab being pulled into a capacitor are concentrated at the edge of the capacitor plates at z = 0 in Figure 6.2.4, which is physically correct, whereas the pressure method implies incorrectly that the force on the slab is concentrated at its end between the plate where z = D. The energy method does not address this issue.

Example 5.5A
At what radius r from a 1-MV high voltage line does the electric force acting on a dust particle having $\varepsilon = 10\varepsilon_o$ exceed the gravitational force if its density ρ is 1 gram/cm^3? Assume the electric field around the line is the same as between concentric cylinders having radii a = 1 cm and b = 10 m.

Solution: The Kelvin polarization force density (5.5.7) can be integrated over the volume v of the particle and equated to the gravitational force $f_g = \rho vg = \sim 10^{-3}v10$ [N]. (5.5.7) yields the total Kelvin force: $\overline{f}_K = v(\varepsilon - \varepsilon_o)\nabla|\overline{E}|^2/2$ where $\overline{E}(r) = \hat{r}\,V/[r\,\ln(b/a)]$ [V m^{-1}]. $\nabla|\overline{E}|^2 = [V/\ln(b/a)]^2\nabla r^{-2} = -2\hat{r}\,[V/\ln(b/a)]^2 r^{-3}$, where the gradient here, $\nabla = \hat{r}\,\partial/\partial r$, was computed using cylindrical coordinates (see Appendix C). Thus $f_g = |\overline{f}_K|$ becomes $10^{-2}v = v9\varepsilon_o\,[V/\ln(b/a)]^2 r^{-3}$, so r = $\{900\varepsilon_o[V/\ln(b/a)]^2\}^{1/3}$ = $\{900\times8.85\times10^{-12}[10^6/\ln(1000)]^2\}^{1/3}$ = 5.5 meters, independent of the size of the particle. Thus high voltage lines make excellent dust catchers for dielectric particles.

5.6   *Photonic forces*

Photonic forces arise whenever electromagnetic waves are absorbed or reflected by objects, and can be found using either wave or photon paradigms. Section 5.2.2 derived the magnetic pressure $\overline{P}_m$ (5.2.13) applied by a surface magnetic field $H_s(t)$ that is parallel to a flat perfect conductor in the x-y plane:

$$\overline{P}_m = \hat{z}\mu_o H_s^2/2 \left[Nm^{-2}\right] \qquad \text{(magnetic pressure on perfect conductor)} \qquad (5.6.1)$$

Thus this instantaneous magnetic pressure perpendicular to the conductor surface equals the adjacent magnetic energy density ([N m^{-2}] = [J m^{-3}]).

In the sinusoidal steady state the time average pressure is half the peak instantaneous value given by (5.6.1), where $H_s(t) = H_s \cos \omega t$. This average pressure on a perfectly reflecting

conductor can also be expressed in terms of the time-average Poynting vector $\langle \overline{S}(t) \rangle$ of an incident wave characterized by $H_+ \cos \omega t$:

$$\langle \overline{P}_m(t) \rangle = \hat{z}\mu_o \langle H_s^2(t) \rangle / 2 = 2\langle \overline{S}(t) \rangle / c \quad [\text{Nm}^{-2}] \qquad (5.6.2)$$

where $H_s = 2H_+$ and $\langle S(t) \rangle = \eta_o H_+^2/2$; the impedance of free space $\eta_o = \mu_o/c$.

It is now easy to relate $\langle S(t) \rangle$ to the photon momentum flux, which also yields pressure. We recall[17] that:

$$photon\ momentum\ M = hf/c \quad [\text{Nm s}^{-1}] \qquad (5.6.3)$$

The momentum transferred to a mirror upon perfect reflection of a single photon at normal incidence is therefore $2hf/c$.

We recall from mechanics that the force f required to change momentum mv is:

$$f = d(mv)/dt \quad [\text{N}] \qquad (5.6.4)$$

so that the total *radiation pressure* on a perfect mirror reflecting directly backwards n photons $[\text{s}^{-1}\text{m}^{-2}]$ is:

$$\langle P_r \rangle = n2hf/c = 2\langle S(t) \rangle / c \quad [\text{Nm}^{-2}] \qquad \text{(radiation pressure on a mirror)} \qquad (5.6.5)$$

consistent with (5.6.2). Thus we have shown that both the Lorentz force method and the photonic force method yield the same pressure on perfectly reflecting mirrors; $P_m = P_r$. The factor of two in (5.6.5) arises because photon momentum is not zeroed but reversed by a mirror. If these photons were absorbed rather than reflected, the rate of momentum transfer to the absorber would be halved. In general if the incident and normally reflected power densities are $\langle S_1 \rangle$ and $\langle S_2 \rangle$, respectively, then the average radiation pressure on the mirror is:

$$\langle P \rangle = \langle S_1 + S_2 \rangle / c \qquad (5.6.6)$$

If the photons are incident at an angle, the momentum transfer is reduced by the cosine of the angle of incidence and reflection. And if the mirror is partially transparent, the momentum transfer is reduced by that fraction of the photon momentum that passes through unaltered.

---

[17] A crude plausibility argument for (5.6.3) is the following. The energy of a photon is hf [J], half being magnetic and half being electric. We have seen in (5.2.1) and (5.2.13) that only the magnetic fields contribute to the Lorentz force on a normal reflecting conductor for which both $E_\perp$ and $H_\perp = 0$, so we might notionally associate hf/2 with the "kinetic energy of a photon", where kinetic energy is linked to momentum. If photons had mass m, this notional kinetic energy hf/2 would equal $mc^2/2$, and the notional associated momentum mc of a photon would then equal hf/c, its actual value.

Consider the simple example of a reflective *solar sail* blown by radiation pressure across the solar system, sailing from planet to planet. At earth the *solar radiation* intensity is ~1400 W/m², so (5.6.6) yields, for example, the total force f on a sail of projected area A intercepting one square kilometer of radiation:

$$f = A\langle P\rangle = A2\langle S(t)\rangle/c \leq 10^6 \times 2 \times 1400/(3\times10^8) \cong 9[N] \qquad (5.6.7)$$

A sail this size one micron thick and having the density of water would have a mass m of 1000 kg. Since the sail velocity v = at = (f/m)t, where a is acceleration and t is time, it follows that after one year the accumulated velocity of a sail facing such constant pressure in vacuum could be as much as $(9/1000)3\times10^7 \cong 3\times10^5$ ms^{-1} = c/1000. Of course the solar photon pressure declines as the square of the solar distance, and solar gravity would also act on such sails.

Example 5.6A
What force F [N] is exerted on a 3-watt flashlight ($\lambda \cong 0.5$ microns) as a result of the exiting photons?

Solution: E = hf and power P = Nhf = 3 watts, where N is the number of photons per second. The force F = Nhf/c, where hf/c is the momentum of a single photon, and N = 3/hf here. So F = 3/c = 10^{-8} Newtons. A Newton approximates the gravitational force on the quarter-pound package of fig newtons. This force pushes the flashlight in the direction opposite to that of the light beam.

Chapter 6: Actuators and Sensors, Motors and Generators

6.1   Force-induced electric and magnetic fields

6.1.1   Introduction

Chapter 5 explained how electric and magnetic fields could exert force on charges, currents, and media, and how electrical power into such devices could be transformed into mechanical power. Chapter 6 explores several types of practical motors and actuators built using these principles, where an actuator is typically a motor that throws a switch or performs some other brief task from time to time. Chapter 6 also explores the reverse transformation, where mechanical motion alters electric or magnetic fields and converts mechanical to electrical power. Absent losses, conversions to electrical power can be nearly perfect and find application in electrical generators and mechanical sensors.

Section 6.1.2 first explores how mechanical motion of conductors or charges through magnetic fields can generate voltages that can be tapped for power. Two charged objects can also be forcefully separated, lengthening the electric field lines connecting them and thereby increasing their voltage difference, where this increased voltage can be tapped for purposes of sensing or electrical power generation. Section 6.1.3 then shows in the context of a current-carrying wire in a magnetic field how power conversion can occur in either direction.

6.1.2   Motion-induced voltages

Any conductor moving across magnetic field lines acquires an open-circuit voltage that follows directly from the Lorentz force law $(6.1.1)$[18]:

$$\overline{f} = q(\overline{E} + \overline{v} \times \mu_o \overline{H}) \tag{6.1.1}$$

Consider the electron illustrated in Figure 6.1.1, which has charge $-e$ and velocity $\overline{v}$.

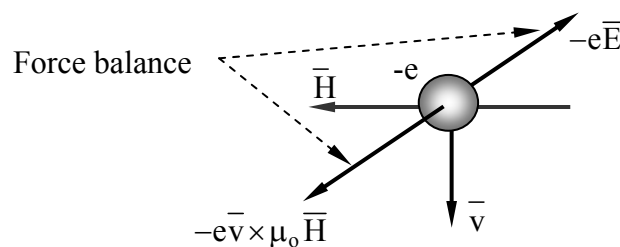Figure 6.1.1   Forces on an electron moving through electromagnetic fields.

---

[18] Some textbooks present alternative explanations that lead to the same results. The explanation here views matter as composed of charged particles governed electromagnetically solely by the Lorentz force law, and other forces, such as the Kelvin force densities acting on media discussed in Section 4.5, are derived from it.

It is moving perpendicular to $\overline{H}$ and therefore experiences a Lorentz force on it of $-e\overline{v}\times\mu_o\overline{H}$. It experiences that force even inside a moving wire and will accelerate in response to it. This force causes all free electrons inside the conductor to move until the resulting surface charges produce an equilibrium electric field distribution such that the net force on any electron is zero.

In the case of a moving open-circuited wire, the free charges (electrons) will move inside the wire and accumulate toward its ends until there is sufficient electric potential across the wire to halt their movement everywhere. Specifically, this Lorentz force balance requires that the force $-e\overline{E}_e$ on the electrons due to the resulting electric field $\overline{E}_e$ be equal and opposite to those due to the magnetic field $-e\overline{v}\times\mu_o\overline{H}$, that is:

$$-e\overline{v}\times\mu_o\overline{H} = e\overline{E}_e \qquad (6.1.2)$$

Therefore the equilibrium electric field inside the wire must be:

$$\overline{E}_e = -\overline{v}\times\mu_o\overline{H} \qquad (6.1.3)$$

There should be no confusion about $\overline{E}_e$ being non-zero inside a conductor. It is the net force on free electrons that must be zero in equilibrium, not the electric field $\overline{E}_e$. The electric Lorentz force $q\overline{E}_e$ must balance the magnetic Lorentz force or otherwise the charges will experience a net force that continues to move them until there is such balance.

Figure 6.1.2 illustrates such a wire of length W moving at velocity $\overline{v}$ perpendicular to $\overline{H}$.
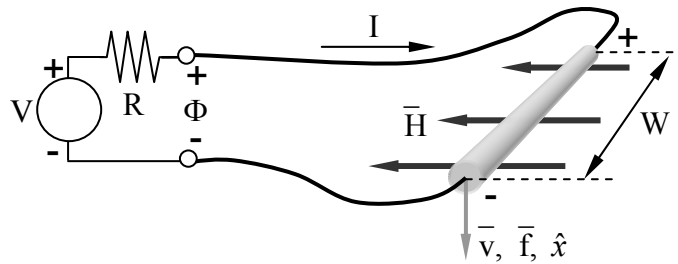
Figure 6.1.2   Forces and voltages on a wire moving in a magnetic field.

If the wire were open-circuited, the potential Φ across it would be the integral of the electric field necessary to cancel the magnetic forces on the electrons, where:

$$\Phi = v\mu_o HW \qquad (6.1.4)$$

and the signs and directions are as indicated in the figure. We assume that the fields, wires, and velocity $\overline{v}$ in the figure are all orthogonal so that $\overline{v}\times\mu_o\overline{H}$ contributes no potential differences except along the wire of length W.

Example 6.1A

A large metal airplane flies at 300 m s^{-1} relative to a vertical terrestrial magnetic field of ~10^{-4} Teslas (1 gauss). What is the open-circuit voltage V wingtip to wingtip if the wingspan W is ~40 meters? If $\overline{B}$ points upwards, is the right wing positive or negative?

Solution: The electric field induced inside the metal is $-\overline{v}\times\mu_o\overline{H}$ (4.3.2), so the induced voltage $V = Wv\mu_oH \cong 40\times300\times1.26\times10^{-6}\times10^{-4} \cong 1.5\times10^{-6}$ volts, and the right wingtip is positive.

6.1.3   Induced currents and back voltages

If the moving wire of Figure 6.1.2 is connected to a load R, then current I will flow as governed by Ohm's law. I depends on Φ, R, and the illustrated Thevenin voltage V:

$$I=(V-\Phi)/R=(V-v\mu_oHW)/R \qquad (6.1.5)$$

The current can be positive or negative, depending on the relative values of V and the motion-induced voltage Φ. From (5.2.7) we see that the magnetic force density on the wire is $\overline{F}=\overline{I}\times\mu_o\overline{H}$ [Nm^{-1}]. The associated total force $\overline{f}_{be}$ exerted on the wire by the environment and by $\overline{H}$ follows from (6.1.5) and is:

$$\overline{f}_{be}=\overline{I}\times\mu_o\overline{H}W=\hat{x}\mu_oHW(V-\Phi)/R \quad [N] \qquad (6.1.6)$$

where the unit vector $\hat{x}$ is parallel to $\overline{v}$.

Equation (6.1.6) enables us to compute the mechanical power delivered to the wire by the environment (P_{be}) or, in the reverse direction, by the wire to the environment (P_{oe}), where $P_{be} = -P_{oe}$. If the voltage source V is sufficiently great, then the system functions as a motor and the mechanical power P_{oe} delivered to the environment by the wire is:

$$P_{oe}=\overline{f}_{oe}\bullet\overline{v}=v\mu_oHW(V-\Phi)/R=\Phi(V-\Phi)/R \quad [W] \qquad (6.1.7)$$

The electrical power P_e delivered by the moving wire to the battery and resistor equals the mechanical power P_{be} delivered to the wire by the environment, where I is given by (6.1.5):

$$P_e=-VI+I^2R=-\left[V(V-\Phi)/R\right]+\left[(V-\Phi)^2/R\right]=\left[(V-\Phi)/R\right]\left[-V+(V-\Phi)\right]$$
$$=-\Phi(V-\Phi)/R=P_{be} \quad [W] \qquad (6.1.8)$$

The negative sign in the first term of (6.1.8) is associated with the direction of I defined in Figure 6.1.2; I flows out of the Thevenin circuit while P_e flows in. If V is zero, then the wire delivers maximum power, Φ^2/R. As V increases, this delivered power diminishes and then becomes negative as the system ceases to be an electrical generator and becomes a motor. As a motor the

mechanical power delivered to the wire by the environment becomes negative, and the electrical power delivered by the Thevenin source becomes positive. That is, we have a:

Motor: If mechanical power out $P_{oe} > 0$,
$V > \Phi = v\mu_o HW$, or $v < V/\mu_o HW$ (6.1.9)

Generator: If electrical power out $P_e > 0$,
$V < \Phi$, or $v > V/\mu_o HW$ (6.1.10)

We call Φ the "*back voltage*" of a motor; it increases as the motor velocity v increases until it equals the voltage V of the power source and $P_e = 0$. If the velocity increases further so that $\Phi > V$, the motor becomes a generator. When $V = \Phi$, then $I = 0$ and the motor moves freely without any electromagnetic forces.

This basic coupling mechanism between magnetic and mechanical forces and powers can be utilized in many configurations, as discussed further below.

Example 6.1B
A straight wire is drawn at velocity $v = \hat{x}\,10$ m s^{-1} between the poles of a 0.1-Tesla magnet; the velocity vector, wire direction, and field direction are all orthogonal to each other. The wire is externally connected to a resistor $R = 10^{-5}$ ohms. What mechanical force $\bar{f}$ is exerted on the wire by the magnetic field $\bar{B}$? The geometry is illustrated in Figure 6.1.2.

Solution: The force exerted on the wire by its magnetic environment (6.1.6) is
$\bar{f}_{be} = \bar{I} \times \bar{H}\mu_o W$ [N], where the induced current $I = -\Phi/R$ and the back voltage
$\Phi = v\mu_o HW$ [V]. Therefore:
$\bar{f}_{be} = -\hat{x}\,\mu_o HW\Phi/R = -\hat{x}\,v(\mu_o HW)^2/R = -\hat{x}\,10\times(0.1\times0.1)^2/10^{-5} = 1$ [N], opposite to $\bar{v}$.

6.2   Electrostatic actuators and motors

6.2.1   Introduction to Micro-Electromechanical Systems (MEMS)

Chapter 6 elaborates on Chapter 5 by exploring a variety of motors, generators, and sensors in both linear and rotary configurations. Electric examples are analyzed in Section 6.2, and magnetic examples in Section 6.3. Section 6.2.1 reviews the background, while Sections 6.2.2 and 6.2.3 explore parallel-capacitor-plate devices using linear and rotary motion respectively. Section 6.2.4 discusses electrostatic motors exerting forces on dielectrics, while Section 6.2.5 discusses the limits to power density posed by electrical breakdown of air or other media, which limits peak electric field strength.

Micro-electromechanical systems (MEMS) are commonly used as motors, generators, actuators, and sensors and underlie one of the major current revolutions in electrical engineering, namely the extension of integrated circuit fabrication technology to electromechanical systems on the same substrate as the circuits with which they interoperate. Such devices now function as

optical switches, radio-frequency switches, microphones, accelerometers, thermometers, pressure sensors, chemical sensors, micro-fluidic systems, electrostatic and magnetic motors, biological sensors, and other devices. They are used in systems as diverse as video projectors, automobile air bag triggers, and mechanical digital memories for hot environments.

Advantages of MEMS over their larger counterparts include size, weight, power consumption, and cost, and also much increased speed due to the extremely small masses and distances involved. For example, some MEMS electromechanical switches can operate at MHz frequencies, compared to typical speeds below ~1 kHz for most traditional mechanical devices. The feature size of MEMS ranges from sub-microns or microns up to one or more millimeters, although the basic electromagnetic principles apply to devices of any scale. Recent advances in micro-fabrication techniques, such as new lithography and etching techniques, precision micro-molds, and improved laser cutting and chipping tools, have simplified MEMS development and extended their capabilities.

The *Lorentz force law* (6.2.1) is fundamental to all electric and magnetic motors and generators and expresses the force vector $\bar{f}$ [Newtons] acting on a charge q [Coulombs] as a function of the local electric field $\bar{E}$, magnetic field $\bar{H}$, and charge velocity vector $\bar{v}$ [ms^{-1}]:

$$\bar{f} = q\left(\bar{E} + \bar{v} \times \mu_o \bar{H}\right) \text{ [Newtons]} \tag{6.2.1}$$

For the examples in Section 6.2 the velocities $\bar{v}$ and magnetic fields $\bar{H}$ are negligible, so the force is primarily electrostatic, $\bar{f} = q\bar{E}$, and can be readily found if $\bar{E}$ is known. When $\bar{E}$ is unknown, the energy method of Section 5.4.2 can often be used instead, as illustrated later. The power densities achievable in MEMS devices can be quite high, and are typically limited by materials failures, such as electrical breakdown or ohmic overheating.

6.2.2   Electrostatic actuators

The simplest MEMS actuators use the electric force between two capacitor plates to pull them together, as illustrated in Figure 6.2.1(a) for a cantilevered loudspeaker or switch. The Lorentz force density F [N m^{-2}] attracting the two plates is given by the $q\bar{E}$ term in (6.2.1). Although one might suppose the force density on the upper plate is simply $\rho_s E$, where ρ_s is the surface charge density [C m^{-2}] on that plate, the correct force is half this value because those charges nearer the surface screen those behind, as suggested in Figure 6.2.1(b); the charges furthest from the surface perceive almost no $\bar{E}$ at all. The figure shows a one-to-one correspondence between electric field lines and charges in a highly idealized distribution—reality is more random. The figure shows that the average field strength E perceived by the charges is half the surface field E_o, independent of their depth distribution $\rho(z)$. Therefore the total attractive electric pressure is:

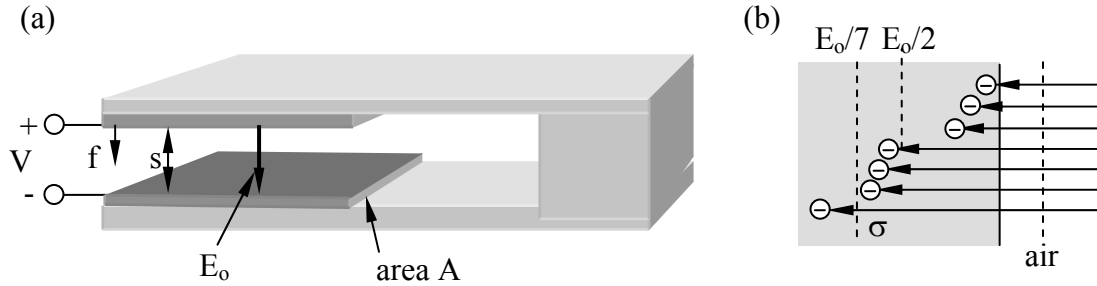$$P_e = \rho_s \left(E_o/2\right) \left[\text{Nm}^{-2}\right] \tag{6.2.2}$$

Figure 6.2.1   Electrostatic MEMS switch and forces on a charged conductor.

But the boundary condition at a conductor (2.6.15) is $\hat{n} \bullet \overline{D} = \rho_s$, so:

$$\rho_s = \varepsilon_o E_o \tag{6.2.3}$$

$$P_e = \varepsilon_o E_o{}^2 \big/ 2 \ \left[Nm^{-2} \right] \quad \textit{(electric pressure} \text{ attracting capacitor plate)} \tag{6.2.4}$$

This is the same pressure derived more rigorously in (5.2.4) and (5.4.3).

If E_o is near its breakdown value $E_B \cong 10^8$ [V m^{-1}] for gaps less than ~10^{-6} meters, then the pressure $P = \varepsilon_o E_o{}^2/2 \cong 8.8 \times 10^{-12} \times 10^{16}/2 = 4.4 \times 10^4$ [N m^{-2}].   A *Newton* is approximately the gravitational force on the apple that fell on Newton's head (prompting his theory of gravity), or on a quarter-pound of butter. Therefore this maximum electrostatic force density is about one pound per square centimeter, comparable to that of a strong magnet.

The cantilever acts like a spring with a *spring constant* k, so the total force f is simply related to the deflection x: f = kx = PA, where A is the area of the capacitor plate. Thus the deflection is:

$$x = PA/k = \varepsilon_o E_o{}^2 A \big/ 2k \ [m] \tag{6.2.5}$$

The ratio A/k is controlled by the composition, thickness, and length of the cantilever, and the desired deflection is controlled by the application. For example, k must be adequate to overcome stiction[19] in switches that make and break contact, and x must be adequate to ensure that the voltage between the capacitor plates does not cause arcing when the switch is open.

Alternatively both capacitor plates could be charged positive or negative so they repel each other. In this case the charge Q moves to the outside surfaces and connects to the very same field strengths as before due to boundary conditions (E = Q/ε_oA) , except that the negative pressure $\varepsilon_o E^2/2$ on the two plates acts to pull them apart rather than together. The field between the plates is then zero.

---

[19] Stiction is the force that must be overcome when separating two contacting surfaces. These forces often become important for micron-sized objects, particularly for good conductors in contact for long periods.

Even with extreme electric field strengths the power density [$W\ m^{-3}$] available with linear motion MEMS actuators may be insufficient. Power equals force times velocity, and rotary velocities can be much greater than linear velocities in systems with limited stroke, such as the cantilever of Figure 6.2.1(a) or the lateral-displacement systems illustrated in Figure 6.2.2. Since it is difficult to compute the lateral electric fields responsible for the lateral forces in rotary or linear systems [e.g., the z components in Figure 6.2.2(a)], the energy methods described below are generally used instead.

(a) (b)
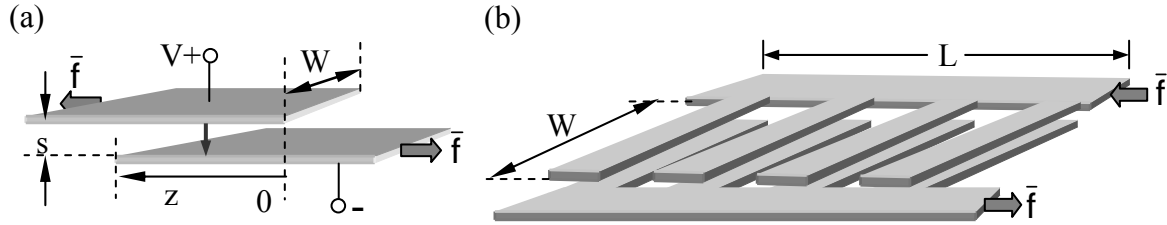
Figure 6.2.2   Electrostatic actuators comprising partially overlapping capacitor plates.

The two charged parallel plates illustrated in Figure 6.2.2(a) are pulled laterally toward one another (z increases) because opposite charges attract. The force $\bar{f}$ required to pull the plates apart depends only on their electric charge q and the plate geometry, independent of any attached circuit. This force $\bar{f}$ in the -z direction can be found by noting that f does work on the capacitor/circuit system, increasing its total energy w_T if f is positive:

$$f = - dw_T/dz = - dw_e/dz - V\ dq/dz\ \ [N] \qquad \text{(energy-force equation)} \qquad (6.2.6)$$

where w_e is the electric energy stored in the capacitor, V is the capacitor voltage[20], and dq is incremental charge flowing from any attached circuit into the positive terminal of the capacitor. The negative sign in (6.2.6) results because f is in the -z direction. Since this energy-force equation is correct regardless of any attached circuit, we can evaluate it for an attached open circuit, battery, or arbitrary Thevenin equivalent, provided it results in the given capacitor voltage V and charge q.

The force computed using (6.2.6) is the same for any attached circuit and any form of the energy expression (3.1.16):

$$w_e = CV^2/2 = q^2/2C\ \ [J] \qquad \text{(electric energy in a capacitor)} \qquad (6.2.7)$$

The algebra is minimized, however, if we assume the capacitor is open-circuit so that q is constant and dq/dz = 0 in (6.2.6). Because V depends on z in this case, it is simpler to use $w_e = q^2/2C$ to evaluate (6.2.6), where: 1) $C = \varepsilon_o Wz/s$ [F], 2) the overlap area of the capacitor is Wz, 3) the plate separation is s << W, and 4) we neglect fringing fields. Thus (6.2.6) becomes:

$$f = - (q^2/2)\ (dC^{-1}/dz) = - (q^2/2)(s/\varepsilon_o W)dz^{-1}/dz = (q^2/2)(s/\varepsilon_o Wz^2)\ \ [N] \qquad (6.2.8)$$

---

[20] For convenience, V represents voltage and v represents velocity in this section.

The rapid increase in force as $z \rightarrow 0$ results because q is constant and concentrates at the ends of the plates as the overlap approaches zero; $z \rightarrow 0$ also violates the assumption that fringing fields can be neglected.

It is interesting to relate the force f of (6.2.8) to the electric field strength E, where:

$$E = \rho_s/\varepsilon_o = q/Wz\varepsilon_o \ [V \ m^{-1}] \tag{6.2.9}$$

$$q = Wz\varepsilon_o E = Wz\varepsilon_o V/s \ [C] \tag{6.2.10}$$

$$f = q^2 s/2\varepsilon_o Wz^2 = Ws\varepsilon_o E^2/2 = A'P_e \ [N] \qquad \text{(lateral electric force)} \tag{6.2.11}$$

where A' = Ws is the cross-sectional area of the gap perpendicular to $\overline{f}$, and $P_e = \Delta W_e = \varepsilon_o E^2/2 - 0$ is the electric pressure difference acting at the end of the capacitor. Note that this pressure is perpendicular to $\overline{E}$ and is "pushing" into the adjacent field-free region where $W_e = 0$; in contrast, the pressure parallel to $\overline{E}$ always "pulls". Later we shall find that "magnetic pressure" $P_m = \Delta W_m$ is similarly attractive parallel to $\overline{H}$ and pushes in directions orthogonal to $\overline{H}$.

Note that if V is constant, then the force f (6.2.11) does not depend on z and is maximized as $s \rightarrow 0$. For a fixed V, the minimum practical plate separation s corresponds to E near the threshold of electrical breakdown, which is discussed further in Section 6.2.5. Also note that the force f is proportional to W, which can be maximized using multiple fingers similar to those illustrated in Figure 6.2.2(b). Actuator and motor designs generally maximize f and W while preserving the desired stroke[21].

Example 6.2A

Design a small electrostatic overlapping plate linear actuator that opens a latch by moving 1 mm with a force of 10^{-2} Newtons.

Solution: The two-plate actuator illustrated in Figure 6.2.2(a) exerts a force $f = Ws\varepsilon_o E^2$ (6.2.11). If E is near the maximum dry-gas value of $\sim 3.2 \times 10^6$ V m^{-1}, the gap s = 1 mm, and W = 1 cm, then $f = 10^{-2} \times 10^{-3} \times 8.85 \times 10^{-12} \times 10^{13} = 8.85 \times 10^{-4}$ [N]. By using M fingers, each wider than the 1-mm stroke, the force can be increased by M [see Figure 6.2.2(b)]. If we let M = 12 the device yields $f = 1.06 \times 10^{-2}$, but its length L must be greater than 12 times twice the finger width (see figure), where the finger width G must exceed not only the stroke but also several times s, in order to make fringing fields negligible. If $G \cong 4$ mm, then the actuator length is $12 \times 2 \times 4$ mm = 9.6 cm, large compared to the width. A three-plate actuator with two grounded plates on the outside and one charged plate inside would double the force, halve the length L, protect users from electrocution, and simplify sealing the actuator against moisture

---

[21] The "stroke" of an actuator is its range of positions; in Figure 6.2.2(a) it would be the maximum minus the minimum value of z. Although the force (6.2.11) becomes infinite as the minimum $z \rightarrow 0$ for constant q, this would violate the assumption z >> d and can cause $V \rightarrow \infty$; V is usually held constant, however.

that could short-circuit the plates. The plate voltage V = Es = 3200 volts. This design is not unique, of course.

6.2.3   Rotary electrostatic motors

Because forces (6.2.4) or (6.2.11) in electrostatic motors are limited by the maximum electric field strength E possible without electric arcing, higher power densities [W m^{-3}] require higher speeds since the power P = fv [watts], where f is force [N], and v is velocity [m s^{-1}]. Figure 6.2.3 pictures an ideal 4-segment rotary *electrostatic motor* for which v and the resulting centrifugal forces are ultimately limited by the tensile strength of the rotor. For both materials and aerodynamic reasons the maximum v at the rotor tip is usually somewhat less than the speed of sound, ~340 m/s. Some rotors spin much faster in vacuum if the material can withstand the centrifugal force.
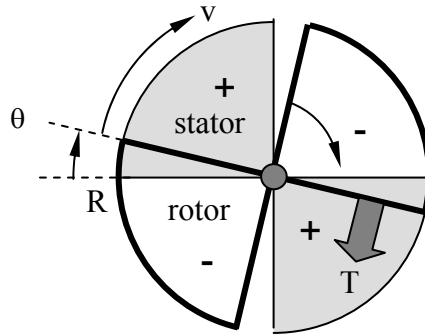
Figure 6.2.3   Four-segment rotary electrostatic motor.

This motor has radius R, plate separation s, and operating voltage V. Stationary "*stator*" plates occupy two quadrants of the motor and a second pair of quadrant plates (the "*rotor*") can rotate to yield an overlap area A = R$^2\theta$ [m^2] that varies from zero to $\pi R^2/2$ as θ increases from zero to $\pi/2$. If the voltage V is applied across the plates, a *torque* T is produced[22], where:

$$T = -dw_T/d\theta \ \ [\text{N m}] \tag{6.2.12}$$

and dw$_T$ is the increment by which the total system energy (fields plus battery) is <u>increased</u> as a result of the motion dθ. The negative sign in (6.2.12) reflects the fact that the torque T is applied by the motor to the environment. If we replace the overlap area of Wz in (6.2.8) by its equivalent R$^2\theta$, then (6.2.8) and (6.2.12) become:

$$w_e = Q^2/2C = Q^2 s/2\varepsilon_o R^2\theta \tag{6.2.13}$$

---

[22] Torque T [Nm] equals the force f on a lever times its length L. Therefore the mechanical work performed by the torque is w$_m$ = fx = fL(x/L) = Tθ, where θ = x/L is the angle (radians) through which the lever rotates about its pivot at one end. Power is Tdθ/dt = Tω [W].

$$T = -dw_T/d\theta = Q^2 s/2\varepsilon_o R^2 \theta^2 = \varepsilon_o R^2 V^2/s \ \text{[N m]} \tag{6.2.14}$$

where $Q = \varepsilon_o R^2 \theta V/s$ [C], which follows from (6.2.10) where $Wz \to R^2\theta$ [m^2].

If we assume $R = 10^{-3}$, $s = 10^{-6}$, and $V = 3$ volts (corresponding to 3×10^6 Vm^{-1}, below the breakdown limit discussed in Section 6.2.5; then (6.2.14) yields:

$$T = 8.8\times10^{-12}\times(10^{-3})^2 3^2/10^{-6} \cong 7.9\times10^{-11} \ \text{[N m]} \tag{6.2.15}$$

This torque exists only until the plates fully overlap, at which time the voltage V is switched to zero until the plates coast another 90° and V is restored. The duty cycle of this motor is thus 0.5 because $T \neq 0$ only half of the time.

A single such ideal motor can then deliver an average of $T\omega/2$ watts, where the factor ½ reflects the duty cycle, and $T\omega$ is the mechanical power associated with torque T on a shaft rotating at ω radians s^{-1}. If the tip velocity v of this rotor is 300 ms^{-1}, slightly less than the speed of sound so as to reduce drag losses while maximizing ω, then the corresponding angular velocity ω is $v/R = 300/10^{-3} = 3\times10^5$ radians s^{-1} or ~3×10^6 rpm, and the available power $T\omega/2 \cong 7.9\times10^{-11} \times 3\times10^5/2 \cong 1.2\times10^{-5}$ watts if we neglect all losses. In principle one might pack ~25,000 motors into one cubic centimeter if each motor were 10 microns thick, yielding ~0.3 W/cm^3. By using a motor with N segments instead of 4 this power density and torque could be increased by a factor of N/4. The small micron-sized gap s would permit values of N as high as ~500 before the fringing fields become important, and power densities of ~40 W/cm^3.

This 40-W/cm^3 power density can be compared to that of a 200-hp automobile engine that delivers 200×746 watts[23] and occupies 0.1 m^3, yielding only ~1.5 W/cm^3. Extremely high power densities are practical only in tiny MEMS devices because heat and torque are then easier to remove, and because only micron-scale gaps permit the highest field strengths, as explained in Section 6.2.5. Rotary MEMS motors have great potential for extremely low power applications where torque extraction can be efficient; examples include drivers for micro-gas-turbines and pumps. The field of MEMS motors is still young, so their full potential remains unknown.

6.2.4   Dielectric actuators and motors

One difficulty with the rotary motor of Figure 6.2.3 is that voltage must be applied to the moving vanes across a sliding mechanical boundary. One alternative is to use a dielectric rotor driven by voltages applied only to the stator. The configuration could be similar to that of Figure 6.2.3 but the rotor would be dielectric and mounted between identical conducting stators with a potential V between them that is turned on and off at times so as to produce an average torque as the rotor rotates. Figure 6.2.4 illustrates the concept in terms of a linear actuator for which the force f can more easily be found. We again assume that fringing fields can be neglected because W >> d.

---

[23] There are 746 watts per horsepower.
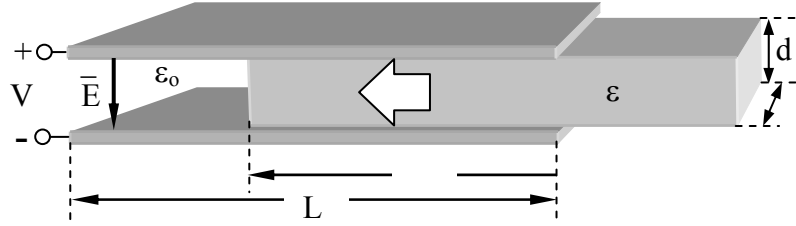
Figure 6.2.4   Linear dielectric slab actuator.

The force f can be found by differentiating the total stored electric energy w_e with respect to motion z, where C is the effective capacitance of this structure, and:

$$w_e = CV^2/2 = Q^2/2C \quad [J] \tag{6.2.16}$$

To simplify differentiating w_e with respect to z, it is easier to use the expression $w_e = Q^2/2C$ because is this case Q is independent of z whereas C is not.

For two capacitors in parallel $C = C_o + C_\varepsilon$ (3.1.14), where C_o and C_ε are the capacitances associated with the air and dielectric halves of the actuator, respectively. Capacitance C was defined in (3.1.8), and equals $\varepsilon A/s$ where A is the plate area and s is the plate separation. It follows that:

$$C = C_\varepsilon + C_o = \varepsilon z W/s + \varepsilon_o(L - z)W/s = [z(\varepsilon - \varepsilon_o) + \varepsilon_o L]W/s \tag{6.2.17}$$

The force f pulling the dielectric slab between the charged plates is given by the force-energy relation (6.2.6) and can be combined with (6.2.16) and (6.2.17) to yield:

$$
\begin{aligned}
f &\cong -dw_e/dz = -d\left(Q^2/2C\right)\!/dz = -\left(Q^2 s/2W\right) d\left[z\left(\varepsilon-\varepsilon_o\right)+\varepsilon_o L\right]^{-1}\!/dz \\
&= \left(Q^2 s/2W\right)\left[z\left(\varepsilon-\varepsilon_o\right)+\varepsilon_o L\right]^{-2}\left(\varepsilon-\varepsilon_o\right) = \left(Q^2 W/2sC^2\right)\left(\varepsilon-\varepsilon_o\right)[N]
\end{aligned}
\tag{6.2.18}
$$

This force can be expressed in terms of the electric field strength E between the two plates by substituting into (6.2.18) the expressions $Q = CV$ and $V = Es$:

$$f \cong (E^2 sW/2)(\varepsilon - \varepsilon_o) = [(\varepsilon - \varepsilon_o)E^2/2]Ws = \Delta P_e A \quad [N] \tag{6.2.19}$$

where $A = Ws$ is the area of the endface of the dielectric slab, and the differential electric pressure pulling the slab between the charged plates is:

$$\Delta P_e = (\varepsilon - \varepsilon_o)E^2/2 \quad [Nm^{-2}] \tag{6.2.20}$$

The differential pressure ΔP_e pushing the interface into the capacitor is thus the difference between the electric pressure on one side of the dielectric interface and that on the other, where the pressure P_e on each side is simply the electric energy density there:

$$P_e = \varepsilon E^2/2 \quad [Nm^{-2}], \ [Jm^{-3}] \tag{6.2.21}$$

Because the electric field at the right-hand end of the slab approaches zero, it exerts no additional force. Electric pressure is discussed further in Section 5.5.2.

Applying these ideas to the rotary motor of Figure 6.2.3 simply involves replacing the rotor by its dielectric equivalent and situating it between conducting stator plates that are excited by V volts so as to pull each dielectric quadrant into the space between them. Then V is switched to zero as the dielectric exits that space so the rotor can coast unpowered until the dielectric quadrants start entering the next pair of stator plates. Thus the drive voltage V is non-zero half the time, with two voltage pulses per revolution of this two-quadrant rotor. The timing of the voltages must be responsive to the exact position of the rotor, which is often determined by a separate rotor angular position sensor. Start-up can fail if the rotor is in exactly the wrong position where $f = 0$ regardless of V, and the rotor will spin backwards if it starts from the wrong position. Figure 6.3.6 suggests how multiple segments and excitation phases can avoid this problem in the context of magnetic motors.

Example 6.2B

Design a maximum-power-density rotary electrostatic motor that delivers 10 W power at $\omega \cong 10^6$ r s^{-1} without make/break or sliding electrical contacts.

Solution: A segmented dielectric rotor sandwiched between charged conducting plates avoids sliding electrical contacts. Assume the rotor has radius R, thickness s, and is made of two electrically insulated dielectrics having permittivities $\varepsilon = 10\varepsilon_o$ and ε_o, and that they are radially segmented as is the rotor in Figure 6.2.3, but with M segments rather than 4. The maximum pressure on the edges of the rotor dielectric boundaries between ε and ε_o is $\Delta P_e = (\varepsilon - \varepsilon_o)E^2/2$ [N m^{-2}]. The mechanical power delivered during the half cycle the voltages are applied to the plate is $T\omega = 20 = \Delta P_e(R/2)sM\omega$. Let's arbitrarily set $s = 10^{-6}$, $E = 10^6$ [V m^{-1}], and M = 800. Therefore $R = 2 \times 20/(sM\omega\Delta P_e) = 40/[10^{-6} \times 800 \times 10^6 \times 9 \times 8.85 \times 10^{-12} \times (10^6)^2/2] = 1.3 \times 10^{-3}$ [m]. The operating voltage is Es $\cong$ 1 volt and the power density is $\sim 10^5$ W/cm^3.

6.2.5   Electrical breakdown

In every case the torque or force produced by an electrostatic MEMS actuator or motor is limited by the *breakdown field* $E_B = V_B/d$, where V_B is the *breakdown voltage*, and the dependence of E_B on d is non-linear. Electric breakdown of a gas occurs when stray free electrons accelerated by E acquire enough velocity and energy (a few electron volts[24]) to knock additional free electrons off gas molecules when they collide, thus triggering a chain reaction that leads to arcing and potentially destructive currents. Water molecules shed electrons much more easily in collisions than do nitrogen or oxygen molecules, and so E_B is much lower in moist air. This is why it is easier to draw visible sparks in cold dry winter air than it is in summer, because in winter the

---

[24] An electron volt is the energy acquired by an electron or other equally charged particle as it accelerates through a potential difference of one volt. It is equivalent to e = 1.6021×10^{-19} Joules.

field strengths can be much greater before breakdown occurs, and such high-voltage breakdowns are more visible.

If, however, the gap between the two electrodes is sufficiently small, the probability diminishes that an ionizing collision will occur between any free electron and a gas atom before the electron hits the positively charged electrode. This *mean-free-path,* or average distance before a "collision", for free electrons is on the order of one micron in air, so breakdown is inhibited for gaps less than the mean free path. However, even when the gap is so narrow that gas breakdown is unlikely, if the field strength E is increased to $\sim 3 \times 10^8$ [V m^{-1}], or two orders of magnitude beyond typical values for E_B in dry gas, any free electrons can then acquire enough energy to knock an ion loose from the positively charged wall. Such a positive ion can then acquire enough energy to release multiple electrons when it impacts the negatively charged wall, producing another form of chain reaction, electrical arcing, and breakdown.

The reasons electric actuators and motors are so attractive on the scale of MEMS, but almost never used at larger scales, are therefore that: 1) the breakdown field strength E_B increases approximately two orders of magnitude for micron-sized gaps, enabling force densities up to four orders of magnitude greater than usual, and 2) enormous values for E and pressure can be achieved with reasonable voltages across micron or sub-micron gaps ($\Rightarrow \sim 3 \times 10^8$ [V m^{-1}] and ~ 10 lb/cm^2).

The breakdown fields for materials are problematic because any local defect can concentrate field strengths locally, exceeding the threshold. Fields of $\sim 10^6$ Vm^{-1} are a nominal upper bound, although somewhat higher values are obtained in integrated circuits.

6.3   Rotary magnetic motors

6.3.1   Commutated rotary magnetic motors

Most *electric motors* and generators are rotary because their motion can then be continuous and high velocity, which improves power density and efficiency while prolonging equipment life. Figure 6.3.1 illustrates an idealized motor with a rotor comprising a single loop of wire carrying current I in the uniform magnetic field $\overline{H}$. The magnetic field can originate from permanent magnets in the stationary *stator*, which is the magnetic structure within which the rotor rotates, or from currents flowing in wires wrapped around the stator. The rotor typically has many turns of wire, often wrapped around a steel core with poles that nearly contact the stator along a cylindrical surface.

The total torque (force times radius) on the motor axle is found by adding the contributions from each of the four sides of the current loop; only the longitudinal elements of length W at radius $\overline{r}$ contribute, however. This total *torque vector* $\overline{T} = \overline{f} \times \overline{r}$ is the integral of the torque contributions from the force density F acting on each incremental length ds of the wire along its entire contour C:

$$\overline{T} = \oint_C \overline{r} \times \overline{F} \, ds \qquad \qquad \text{(torque on rotor)} \qquad (6.3.1)$$

- 163 -
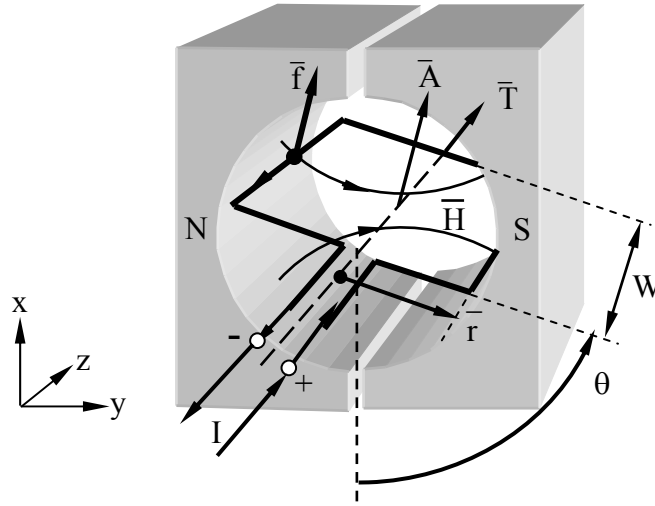
Figure 6.3.1   Rotary single-turn magnetic motor.

The force density $\overline{F}$ [N m^{-1}] on a wire conveying current $\overline{I}$ in a magnetic field $\overline{H}$ follows from the Lorentz force equation (5.1.1) and was given by (5.2.7):

$$\overline{F} = \overline{I} \times \mu_o \overline{H} \left[N\,m^{-1} \right] \qquad \text{(force density on wire)} \qquad (6.3.2)$$

Thus the torque for this motor at the pictured instant is clockwise and equals:

$$\overline{T} = \hat{z}\, 2 r I \mu_o H W \left[N\,m \right] \qquad (6.3.3)$$

In the special case where $\overline{H}$ is uniform over the coil area $A_o = 2rW$, we can define the *magnetic moment* $\overline{M}$ of the coil, where $|\overline{M}| = I A_o$ and where the vector $\overline{M}$ is defined in a right hand sense relative to the current loop $\overline{I}$. Then:

$$\overline{T} = \overline{M} \times \mu_o \overline{H} \qquad (6.3.4)$$

Because the current flows only in the given direction, $\overline{H}$ and the torque reverse as the wire loop passes through vertical ($\theta = n\pi$) and have zero average value over a full rotation. To achieve positive average torque, a *commutator* can be added, which is a mechanical switch on the rotor that connects one or more rotor windings with one or more stationary current sources in the desired sequence and polarity. The commutator reverses the direction of current at times chosen so as to maximize the average positive torque. A typical configuration is suggested in Figure 6.3.2(a) where two spring-loaded carbon brushes pass the current I to the commutator contacts, which are rigidly attached to the rotor so as to reverse the current polarity twice per revolution. This yields the more nearly constant torque history T(θ) illustrated by the dashed line in Figure 6.3.2(b). In this approximate analysis of a DC motor we assume that the time

constant L/R associated with the rotor inductance L and circuit resistance R is short compared to the torque reversals illustrated in Figure 6.3.2(b).
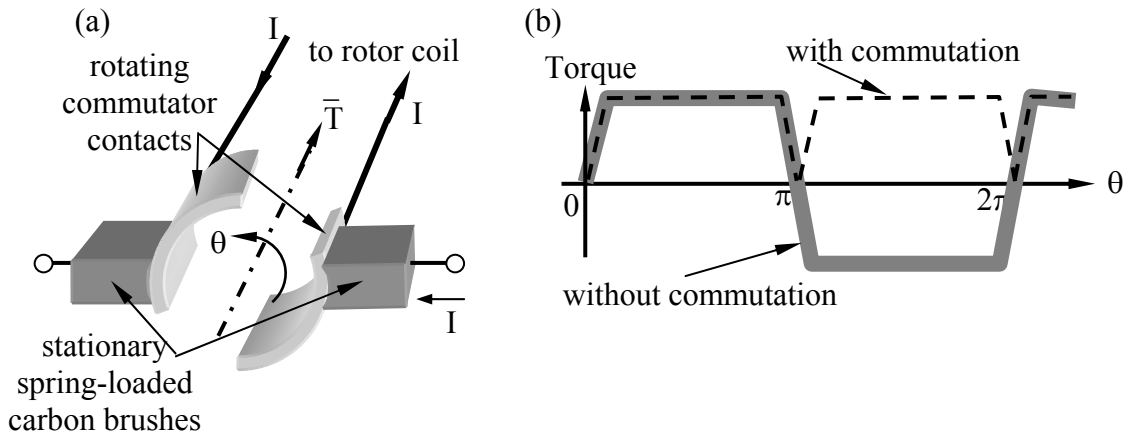
Figure 6.3.2   Commutator motor torque history and contact configuration.

Power is conserved, so if the windings are lossless then the average electrical power delivered to the motor, $P_e = \langle VI \rangle$, equals the average mechanical power delivered to the environment:

$$P_m = f_m v_m = f_m r_m \omega = T\omega \tag{6.3.5}$$

where v_m is the velocity applied to the motor load by force f_m at radius r_m. If the motor is driven by a current source I, then the voltage across the rotor windings in this lossless case is:

$$V = P_e/I = P_m/I = T\omega/I \tag{6.3.6}$$

This same voltage V across the rotor windings can also be deduced from the Lorentz force $\bar{f} = q(\bar{E} + \bar{v} \times \mu_o \bar{H})$, (6.1.1), acting on free conduction electrons within the wire windings as they move through $\bar{H}$. For example, if the motor is open circuit ($I \equiv 0$), these electrons spinning about the rotor axis at velocity $\bar{v}$ will move along the wire due to the "$q\bar{v} \times \mu_o \bar{H}$" force on them until they have charged parts of that wire relative to other parts so as to produce a "$q\bar{E}$" force that balances the local magnetic force, producing equilibrium and zero additional current. Free electrons in equilibrium have repositioned themselves so they experience no net Lorentz force. Therefore:

$$\bar{E} = -\bar{v} \times \mu_o \bar{H} \quad \left[V\,m^{-1} \right] \qquad \text{(electric field inside moving conductor)} \tag{6.3.7}$$

The integral of $\bar{E}$ from one end of the conducting wire to the other yields the open-circuit voltage Φ, which is the Thevenin voltage for this moving wire and often called the *motor back-*

voltage. Φ varies only with rotor velocity and H, independent of any load. For the motor of Figure 6.3.1, Equation (6.3.7) yields the open-circuit voltage for a one-turn coil:

$$\Phi = 2EW = 2v\mu_o HW = 2\omega r\mu_o HW = \omega A_o \mu_o H \quad [V] \qquad \text{(motor back-voltage)} \qquad (6.3.8)$$

where the single-turn coil area is $A_o = 2rW$. If the coil has N turns, then A_o is replaced by NA_o in (6.3.8).

The Thevenin equivalent circuits representing the motor and its external circuit determine the current I, as illustrated in Figure 6.3.3.
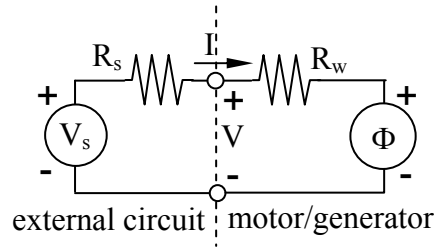
Figure 6.3.3   Equivalent circuit for a driven motor/generator.

R_w is the winding resistance of the motor, where:

$$I = (V_s - \Phi)/(R_s + R_w) \qquad \text{(motor current)} \qquad (6.3.9)$$

When the motor is first starting, $\omega = \Phi = 0$ and the current and the torque are maximum, where $I_{max} = V_s/(R_s + R_w)$. The maximum torque, or "starting torque" from (6.3.3), where $A_o = 2rW$ and there are N turns, is:

$$\overline{T}_{max} = \hat{z}2WrNI_{max}\mu_o H = \hat{z}NA_o I_{max}\mu_o H \quad [Nm] \qquad (6.3.10)$$

Since $\Phi = 0$ when $\overline{v} = 0$, $I\Phi = 0$ and no power is converted then. As the motor accelerates toward its maximum ω, the back-voltage Φ steadily increases until it equals the source voltage V_s so that the net driving voltage, torque T, and current $I \to 0$ at $\omega = \omega_{max}$. Since (6.3.8) says $\Phi = \omega NA_o\mu_o H$, it follows that if $\Phi = V_s$, then:

$$\omega_{max} = \frac{V_s}{NA_o\mu_o H} = \frac{V_s I_{max}}{T_{max}} \qquad (6.3.11)$$

where the relation to T_{max} comes from (6.3.10), and ω_{max} occurs at $T_{min.} = 0$. At ω_{max} no power is being converted, so the maximum motor power output P_{max} occurs at an intermediate speed ω_p, as illustrated in Figure 6.3.4.
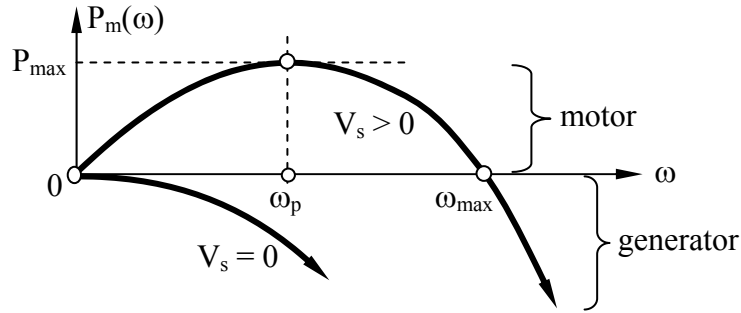
Figure 6.3.4 Mechanical power output $P_m(\omega)$ from a magnetic motor.

An expression for the mechanical output power $P_m(\omega)$ follows from (6.3.9):

$$P_m = T\omega = I\Phi = \left(V_s\Phi - \Phi^2\right)\Big/\left(R_s + R_w\right) \qquad \text{(mechanical power out)} \qquad (6.3.12)$$

where (6.3.8) says $\Phi = \omega N A_o \mu_o H$, so $P_m \propto (V_s\omega - N A_o \mu_o H \omega^2)$.

Equation (6.3.12) says that if $V_s \gg \Phi$, which occurs for modest values of ω, then the motor power increases linearly with Φ and ω. Also, if $V_s = 0$, then P_m is negative and the device acts as a *generator* and transfers electrical power to $R_s + R_w$ proportional to Φ^2 and therefore ω^2. Moreover, if we differentiate P_m with respect to Φ and set the result to zero, we find that the mechanical power is greatest when $\Phi = V_s/2$, which implies $\omega_p = \omega_{max}/2$. In either the motor or generator case, the maximum power transfer is usually limited by currents overheating the insulation or by high voltages causing breakdown. Even when no power is transferred, the back-voltage Φ could cause breakdown if the device spins too fast. Semiconductor switches that may fail before the motor insulation are increasingly replacing commutators so the risk of excessive ω is often a design issue. In an optimum motor design, all failure types typically occur near the same loading levels or levels of likelihood.

Typical parameters for a commutated 2-inch motor of this type might be: 1) $B = \mu_o H = 0.4$ Tesla (4000 gauss) provided by permanent magnets in the stator, 2) an $N = 50$-turn coil on the rotor with effective area $A = 10^{-3}N$ [m^2], 3) $V_s = 24$ volts, and 4) $R_s + R_w = 0.1$ ohm. Then it follows from (6.3.11), (6.3.12) for $\Phi = V_s/2$, and (6.3.10), respectively, that:

$$\omega_{max} = V_s/\mu_o HAN = 24\Big/\left(0.4\times10^{-3}\times50\right) = 1200 \ \left[rs^{-1}\right] \Rightarrow 11,460 \ [rpm]^{\,25} \qquad (6.3.13)$$

$$P_{max} = \left(V_s\Phi - \Phi^2\right)\Big/\left(R_s + R_w\right) = V_s^2\Big/\left[4\left(R_s + R_w\right)\right][W] = 24^2/0.4 \cong 1.4\,[kW] \qquad (6.3.14)$$

$$T_{max} = AN\mu_o HI_{max} = A\mu_o HV_s\Big/\left(R_s + R_w\right) = 0.05\times0.4\times24/0.1 = 4.8 \ [Nm] \qquad (6.3.15)$$

---

[25] The abbreviation "rpm" means revolutions per minute.

In practice, most motors like that of Figure 6.3.1 wrap the rotor windings around a high permeability core with a thin gap between rotor and stator; this maximizes H near the current I. Also, if the unit is used as an AC generator, then there may be no need for the polarity-switching commutator if the desired output frequency is simply the frequency of rotor rotation.

Example 6.3A

Design a commutated DC magnetic motor that delivers maximum mechanical power of 1 kW at 600 rpm. Assume B = 0.2 Tesla and that the source voltage V_s = 50 volts.

Solution: Maximum mechanical power is delivered at $\omega_p = \omega_{max}/2$ (see Figure 6.3.4). Solving (6.3.13) yields $NA_o = V_s/(\omega_{max}\mu_o H) = 50/(2\times600\times60\times2\pi\times0.2) = 5.53\times10^{-4}$, where ω_{max} corresponds to 1200 rpm. If N = 6, then the winding area $2rW = A_o \cong$ 1 cm^2. To find I_{max} we use (6.3.14) to find the maximum allowed value of $R_s + R_w = (V_s\Phi - \Phi^2)/P_m$. But when the delivered mechanical power P_{mech} is maximum, $\Phi = V_s/2$, so $R_s + R_w = (50\times25 - 25^2)/10^3 = 0.63\Omega$, which could limit N if the wire is too thin. $I_{max} = V_s/(R_s + R_w) = 50/0.63 = 80$ [A]. The starting torque $T_{max} = I_{max}(NA_o)\mu_o H = 80(5.53\times10^{-4})0.2 = 0.22$ [N m]. This kilowatt motor occupies a fraction of a cubic inch and may therefore overheat because the rotor is small and its thermal connection with the external world is poor except through the axle. It is probably best used in short bursts between cooling-off periods. The I^2R_w thermal power dissipated in the rotor depends on the wire design.

6.3.2   Reluctance motors

Reluctance motors combine the advantages of rotary motion with the absence of rotor currents and the associated rotary contacts. Figure 6.3.5 suggests a simple idealized configuration with only a single drive coil.
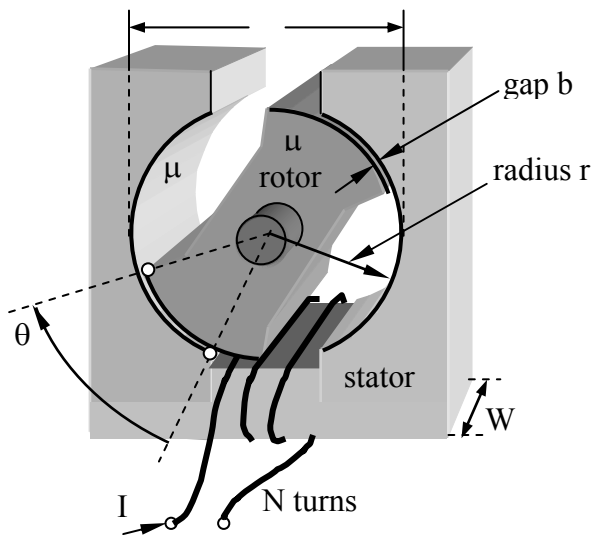

Figure 6.3.5   Two-pole single-winding reluctance motor.

When the coil is energized the rotor is pulled by the magnetic fields into alignment with the magnetic fields linking the two poles of the stator, where $\mu \gg \mu_o$ in both the rotor and stator. Reluctance motors must sense the angular position of the rotor, however, so the stator winding(s) can be excited at the right times so as to pull the passive high-μ rotor toward its next rotary position, and then not retard it as it moves on toward the following attractive position. For example, the current I in the figure will pull the rotor so as to increase θ, which is the overlap angle between the rotor and the stator poles. Once the overlap is complete the current I would be set to zero as the rotor coasts until the poles again have $\theta \cong 0$ and are in position to be pulled forward again by I. Such motors are efficient if hysteresis losses in the stator and rotor are modest and the stator windings are nearly lossless.

The torque on such a reluctance motor can be readily calculated using (6.2.12):

$$T = -dw_T/d\theta \ [\text{N m}] \tag{6.3.16}$$

The total magnetic energy w_T includes w_μ within the rotor and stator, w_g in the air gaps between them, and any energy in the power supply driving the motor. Fortunately we can simplify the problem by noting that w_g generally dominates, and that by short-circuiting the stator the same torque exists without any power source if I remains unchanged.

The circumstances for which the gap energy dominates the total energy w_T are easily found from the static integral form of Ampere's law (1.4.1):

$$NI = \oint_C \left(\overline{H}_{gap} + \overline{H}_{stator} + \overline{H}_{rotor} \right) \bullet d\overline{s} \cong 2bH_{gap} \tag{6.3.17}$$

To derive an approximate result we may assume the coil has N turns, the width of each gap is b, and the contour C threads the coil and the rest of the motor over a distance $\sim 2D$, and through an approximately constant cross-section A; D is the rotor diameter. Since the boundary conditions in each gap require $\overline{B}_\perp$ be continuous, $\mu H_\mu \cong \mu_o H_g$, where $H_\mu \cong H_{stator} \cong H_{rotor}$ and $H_g \equiv H_{gap}$. The relative energies stored in the two gaps and the rotor/stator are:

$$w_g \cong 2bA \left(\mu_o H_g^2 \big/ 2 \right) \tag{6.3.18}$$

$$w_{r/s} \cong 2DA \left(\mu H_\mu^2 \big/ 2 \right) \tag{6.3.19}$$

Their ratio is:

$$w_g / w_{r/s} = 2b \left(\mu_o/\mu \right) \left(H_g/H_\mu \right)^2 \big/ 2D = b \left(\mu/\mu_o \right) \big/ D \tag{6.3.20}$$

Thus $w_g \gg w_\mu$ if $b/D \gg \mu_o/\mu$. Since gaps are commonly $b \cong 100$ microns, and iron or steel is often used in reluctance motors, $\mu \cong 3000$, so gap energy w_g dominates if the motor diameter D $\ll 0.3$ meters. If this approximation doesn't apply then the analysis becomes somewhat more

- 169 -

complex because both energies must be considered; reluctance motors can be much larger than 0.3 meters and still function.

Under the approximations $w_T \cong w_g$ and $A = $ gap area $= r\theta W$, we may compute the torque T using (6.3.16) and (6.3.18):[26]

$$T = -dw_g/d\theta = -b\mu_o d\left(A\left|H_g\right|^2\right)/d\theta \qquad (6.3.21)$$

The θ dependence of H_g can be found from Faraday's law by integrating $\overline{E}$ around the short-circuited coil:

$$\oint_{c \text{ coil}} \overline{E} \bullet d\overline{s} = -N \oiint_A \left(d\overline{B}/dt\right) \bullet d\overline{a} = -d\Lambda/dt = 0 \qquad (6.3.22)$$

The flux linkage Λ is independent of θ and constant around the motor [contour C of (6.3.17)], so Λ, w_g and T are easily evaluated at the gap where the area is $A = r\theta W$:

$$\Lambda = N\int_A \overline{B} \bullet d\overline{a} = NB_g A = N\mu_o H_g r\theta W \qquad (6.3.23)$$

$$w_g = 2\left(\mu_o H_g{}^2/2\right)bA = b\Lambda^2/\left(N^2\mu_o r\theta W\right) \qquad (6.3.24)$$

$$T = -dw_g/d\theta = b\Lambda^2/\left(N^2\mu_o rW\theta^2\right) = r2\left(\mu_o H_g{}^2/2\right)Wb \; [\text{N m}] \qquad (6.3.25)$$

The resulting torque T in (6.3.25) can be interpreted as being the product of radius r and twice the force exerted at the leading edge of each gap (twice, because there are two gaps), where this force is the *magnetic pressure* $\mu_o H_g{}^2/2$ [N m^{-2}] times the gap area Wb projected on the direction of motion. Because the magnetic field lines are perpendicular to the direction of force, the magnetic pressure pushes rather than pulls, as it would if the magnetic field were parallel to the direction of force. Unfortunately, increasing the gap b does not increase the force, because it weakens H_g proportionately, and therefore weakens $T \propto H^2$. In general, b is designed to be minimum and is typically limited to roughly 25-100 microns by thermal variations and bearing and manufacturing tolerances. The magnetic field in the gap is limited by the saturation field of the magnetic material, as discussed in Section 2.5.4.

The drive circuits initiate the current I in the reluctance motor of Figure 6.3.5 when the gap area $r\theta W$ is minimum, and terminate it when that area becomes maximum. The rotor then coasts with $I = 0$ and zero torque until the area is again minimum, when the cycle repeats. Configurations that deliver continuous torque are more commonly used instead because of their smoother performance.

---

[26] The approximate dependence (6.3.19) of $w_{r/s}$ upon $A = r\theta W$ breaks down when $\theta \rightarrow 0$, since w_g doesn't dominate then and (6.3.19) becomes approximate.
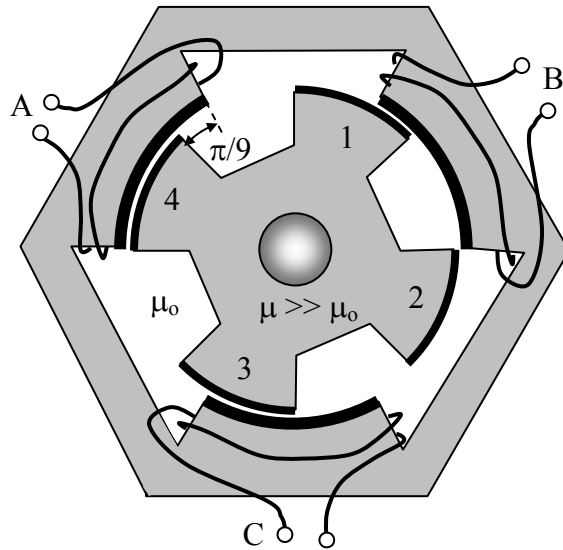
Figure 6.3.6   Reluctance motor with 3 stator and 4 rotor poles.

Figure 6.3.6 illustrates a reluctance motor that provides continuous torque using three stator poles (A, B, C) and four rotor poles (1, 2, 3, 4). When windings A and B are excited, rotor pole 1 is pulled clockwise into stator pole B. The gap area for stator pole A is temporarily constant and contributes no additional torque.   After the rotor moves $\pi/9$ radians, the currents are switched to poles B and C so as to pull rotor pole 2 into stator pole C, while rotor pole 1 contributes no torque. Next C and A are excited, and this excitation cycle (A/B, B/C, C/A) is repeated six times per revolution.   Counter-clockwise torque is obtained by reversing the excitation sequence. Many pole combinations are possible, and those with more poles yield higher torques because torque is proportional to the number of active poles. In this case only one pole is providing torque at once, so the constant torque $T = bW(\mu_o H_g^2/2)$ [N m].

A calculation very similar to that above also applies to *relays* such as that illustrated in Figure 6.3.7, where a coil magnetizes a flexible or hinged bar, drawing it downward to open and/or close one or more electrical contacts.
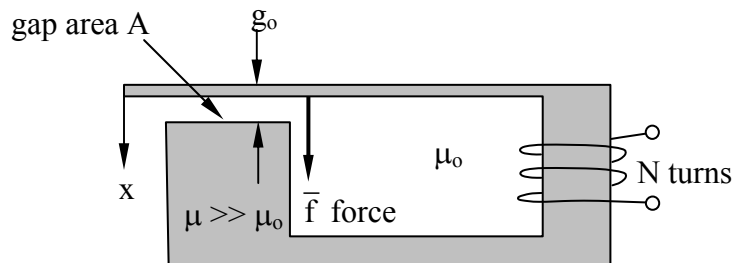
Figure 6.3.7   Magnetic relay.

We can find the force f, flux linkage Λ, and gap energy w_g using a short-circuited N-turn coil to render Λ constant, as before:

$$f = -dw_g/dx \qquad\qquad\qquad\text{(force closing the gap)} \quad (6.3.26)$$

$$\Lambda = N\mu_o H_g A \qquad\qquad\qquad\text{(flux linkage)} \quad (6.3.27)$$

$$w_g = \left(\mu_o H_g^2/2\right)A(g_o - x) = (g_o - x)\Lambda^2\Big/\left(2N^2\mu_o A\right) \ [J] \qquad\text{(gap energy)} \quad (6.3.28)$$

$$f = -dw_g/dx = \Lambda^2\Big/\left(2N^2\mu_o A\right) = \left(\mu_o H_g^2/2\right)A \ [N] \qquad\qquad\text{(force)} \quad (6.3.29)$$

This force can also be interpreted as the gap area A times a *magnetic pressure* P_m, where:

$$P_m = \mu_o H_g^2/2 \ [N/m^2] \qquad\qquad\qquad\text{(magnetic pressure)} \quad (6.3.30)$$

The magnetic pressure is attractive parallel to the field lines, tending to close the gap. The units N/m^2 are identical to J/m^3. Note that the minus sign is used in (6.3.29) because f is the magnetic force closing the gap, which equals the mechanical force required to hold it apart; motion in the x direction reduces w_g.

Magnetic micro-rotary motors are difficult to build without using magnetic materials or induction[27] because it is difficult to provide reliable sliding electrical contacts to convey currents to the rotor. One form of rotary magnetic motor is similar to that of Figures 6.2.3 and 6.2.4, except that the motor pulls into the segmented gaps a rotating high-permeability material instead of a dielectric, where the gaps would have high magnetic fields induced by stator currents like those in Figure 6.3.6. As in the case of the rotary dielectric MEMS motors discussed in Section 6.2, the timing of the currents must be synchronized with the angular position of the rotor. The force on a magnetic slab moving into a region of strong magnetic field can be shown to approximate $A\mu H_\mu^2/2$ [N], where A is the area of the moving face parallel to H_μ, which is the field within the moving slab, and $\mu \gg \mu_o$. The rotor can also be made permanently magnetic so it is attracted or repelled by the synchronously switched stator fields; permanent magnet motors are discussed later in Section 6.5.2.

Example 6.3B
A relay like that of Figure 6.3.7 is driven by a current source I [A] and has a gap of width g. What is the force f(g) acting to close the gap? Assume the cross-sectional area A of the gap and metal is constant around the device, and note the force is depends on whether the gap is open or closed.

Solution: This force is the pressure $\mu_o H_g^2/2$ times the area A (6.3.29), assuming $\mu \gg \mu_o$. Since $\nabla \times \overline{H} = \overline{J}$, therefore $NI = \oint H(s) \ ds = H_g g + H_\mu S$, where S is the path length around the loop having permeability μ. When $H_g g \gg H_\mu S$, then $H_g \cong NI/g$ and $f \cong$

---

[27] Induction motors, not discussed in this text, are driven by the magnetic forces produced by a combination of rotor and stator currents, where the rotor currents are induced by the time-varying magnetic fields they experience, much like a transformer. This avoids the need for direct electrical contact with the rotor.

- 172 -

$\mu_o(NI/g)^2A/2$ for the open relay. When the relay is closed and $g \cong 0$, then $H_g \cong \mu H_u/\mu_o$, where $H_\mu \cong NI/S$; then $f \cong (\mu NI/S)^2 A/2\mu_o$. The ratio of forces when the relay is closed to that when it is open is $(\mu g/\mu_o S)^2$, provided $H_g g >> H_\mu S$ and this ratio $>> 1$.

6.4   Linear magnetic motors and actuators

6.4.1   Solenoid actuators

Compact actuators that flip latches or switches, increment a positioner, or impact a target are often implemented using solenoids. *Solenoid actuators* are usually cylindrical coils with a slideably disposed high-permeability cylindrical core that is partially inserted at rest, and is drawn into the solenoid when current flows, as illustrated in Figure 6.4.1. A spring (not illustrated) often holds the core near its partially inserted rest position.
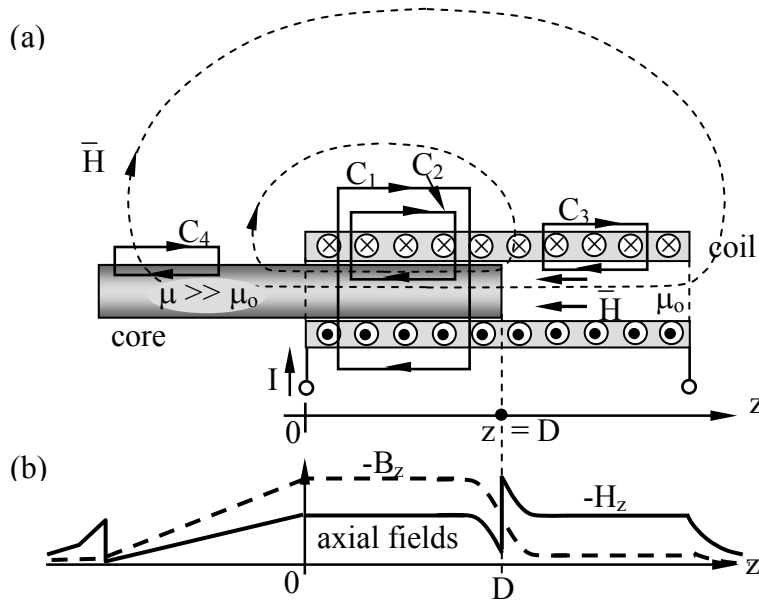
Figure 6.4.1   Solenoid actuator and fields (B and H are plotted on different scales).

If we assume the diameter of the solenoid is small compared to its length, then the fringing fields at the ends of the coil and core can be neglected relative to the field energy stored elsewhere along the solenoid. If we integrate $\overline{H}$ along contour C_1 (see figure) we obtain zero from Ampere's law because no net current flows through C_1 and $\partial\overline{D}/\partial t \cong 0$:

$$\oint_C \overline{H} \bullet d\overline{s} = \oiint_A \left(\overline{J} + \partial\overline{D}/\partial t\right) \bullet \hat{n} \, da = 0 \qquad (6.4.1)$$

This implies $\overline{H} \cong 0$ outside the solenoid unless H_z is approximately uniform outside, a possibility that is energetically disfavored relative to H being purely internal to the coil. Direct evaluation

of $\overline{H}$ using the Biot-Savart law (1.4.6) also yields $\overline{H} \cong 0$ outside. If we integrate $\overline{H}$ along contour C_2, which passes along the axis of the solenoid for unit distance, we obtain:

$$\oint_{C2} \overline{H} \bullet d\overline{s} = N_o I = -H_z \tag{6.4.2}$$

where N_o is defined as the number of turns of wire per meter of solenoid length. We obtain the same answer (6.4.2) regardless of the permeability along the contour C_2, provided we are not near the ends of the solenoid or its moveable core. For example, (6.4.2) also applies to contour C_3, while the integral of $\overline{H}$ around C_4 is zero because the encircled current there is zero.

Since (6.4.2) requires that H_z along the solenoid axis be approximately constant, B_z must be a factor of μ/μ_o greater in the permeable core than it is in the air-filled portions of the solenoid. Because boundary conditions require $\overline{B}_\perp$ to be continuous at the core-air boundary, $\overline{H}_\perp$ must be discontinuous there so that $\mu H_\mu = \mu_o H_o$, where H_μ and H_o are the axial values of H in the core and air, respectively. This appears to conflict with (6.4.2), which suggests $\overline{H}$ inside the solenoid is independent of μ, but this applies only if we neglect fringing fields at the ends of the solenoid or near boundaries where μ changes. Thus the axial H varies approximately as suggested in Figure 6.4.1(b): it has a discontinuity at the boundary that relaxes toward constant H $= N_o I$ away from the boundary over a distance comparable to the solenoid diameter. Two representative field lines in Figure 6.4.1(a) suggest how $\overline{B}$ diverges strongly at the end of the magnetic core within the solenoid while other field lines remain roughly constant until they diverge at the right end of the solenoid. The transition region between the two values of B_z at the end of the solenoid occurs over a distance roughly equal to the solenoid diameter, as suggested in Figure 6.4.1(b). The magnetic field lines $\overline{B}$ and $\overline{H}$ "repel" each other along the protruding end of the high permeability core on the left side of the figure, resulting in a nearly linear decline in magnetic field within the core there; at the left end of the core there is again a discontinuity in $|H_z|$ because $\overline{B}_\perp$ must be continuous.

Having approximated the field distribution we can now calculate energies and forces using the expression for magnetic energy density, $W_m = \mu H^2/2$ [J m^{-3}]. Except in the negligible fringing field regions at the ends of the solenoid and at the ends of its core, $|H| \cong N_o I$ (6.4.2) and $\mu H^2 >> \mu_o H^2$, so to simplify the solution we neglect the energy stored in air as we compute the magnetic force f_z pulling on the core in the +z direction:

$$f_z = -dw_T/dz \ \ [N] \tag{6.4.3}$$

The energy in the core is confined largely to the length z within the solenoid, which has a cross-sectional area A [m^2]. The total magnetic energy w_m thus approximates:

$$w_m \cong Az\mu H^2\big/2 \ \ [J] \tag{6.4.4}$$

If we assume $w_T = w_m$ and differentiate (6.4.4) assuming H is independent of z, we find the magnetic force expels the core from the solenoid, the reverse of the truth. To obtain the correct answer we must differentiate the total energy w_T in the system, which includes any energy in the

power source supplying the current I. To avoid considering a power supply we may alternatively assume the coil is short-circuited and carrying the same I as before.   Since the instantaneous force on the core depends on the instantaneous I and is the same whether it is short-circuited or connected to a power source, we may set:

$$v = 0 = d\Lambda/dt \tag{6.4.5}$$

where:

$$\Lambda \cong N\psi_m = N\iint_A \mu\overline{H}_\mu \bullet d\overline{a} = N_o z\ \mu H_\mu A \tag{6.4.6}$$

H_μ is the value of H inside the core (μ) and $N_o z$ is the number of turns of wire circling the core, where N_o is the number of turns per meter of coil length. But $H_\mu = J_s\ [A\ m^{-1}] = N_o I$, so:

$$\Lambda = N_o^2 Iz\mu A \tag{6.4.7}$$

$$I = \Lambda / \left(N_o^2 z\mu A \right) \tag{6.4.8}$$

We now can compute w_T using only w_m because we have replaced the power source with a short circuit that stores no energy:

$$w_T \cong \mu H_\mu^2 Az/2 = \mu \left(N_o I \right)^2 Az/2 = \mu \left(\Lambda/\mu N_o Az \right)^2 Az/2 = \Lambda^2 / \left(\mu N_o^2 Az2 \right) \tag{6.4.9}$$

So (6.4.9) and (6.4.6) yield the force pulling the core into the solenoid:

$$f_z = -\frac{dw_T}{dz} = -\frac{d}{dz}\left[\frac{\Lambda^2}{\mu N_o^2 2Az} \right] = \frac{(\Lambda/N_o z)^2}{2A\mu} = \frac{\mu H_\mu^2 A}{2} \quad [N] \tag{6.4.10}$$

where $H_\mu = H$. This force is exactly the area A of the end of the core times the same magnetic pressure $\mu H^2/2\ [Nm^{-2}]$ we saw in (6.3.25), but this time the magnetic field is pulling on the core in the direction of the magnetic field lines, whereas before the magnetic field was pushing perpendicular to the field lines. This pressure equals the magnetic energy density W_m, as before. A slight correction for the non-zero influence of μ_o and associated small pressure from the air side could be made here, but more exact answers to this problem generally also require consideration of the fringing fields and use of computer tools.

It is interesting to note how electric and magnetic pressure [N/m^2] approximates the energy density [J m^{-3}] stored in the fields, where we have neglected the pressures applied from the low-field side of the boundary when $\varepsilon \gg \varepsilon_o$ or $\mu \gg \mu_o$. We have now seen examples where $\overline{E}$ and $\overline{H}$ both push or pull on boundaries from the high-field (usually air) side of a boundary, where both $\overline{E}$ and $\overline{H}$ pull in the direction of their field lines, and push perpendicular to them.

6.4.2   MEMS magnetic actuators

One form of magnetic MEMS switch is illustrated in Figure 6.4.2. A control current I_2 deflects a beam carrying current I_1. When the beam is pulled down toward the substrate, the switch (not shown) will close, and when the beam is repelled upward the switch will open. The Lorentz force law (1.2.1) states that the magnetic force $\bar{f}$ on a charge q is $q\bar{v}\times\mu_o\bar{H}$, and therefore the force density per unit length $\bar{F}$ [N m^{-1}] on a current $\bar{I}_1 = Nq\bar{v}$ induced by the magnetic field $\bar{H}_{12}$ at position 1 produced by I_2 is:

$$\bar{F} = Nq\bar{v}\times\mu_o\bar{H}_{12} = \bar{I}_1\times\mu_o\bar{H}_{12} \ \left[Nm^{-1}\right] \tag{6.4.11}$$

N is the number of moving charges per meter of conductor length, and we assume that all forces on these charges are conveyed directly to the body of the conductor.
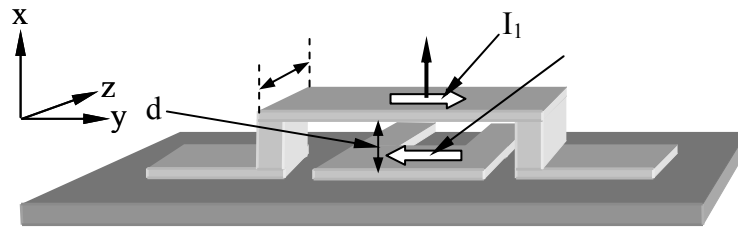
Figure 6.4.2   Magnetic MEMS switch.

If the plate separation d $<<$ W, then fringing fields can be neglected and the I_2-induced magnetic field affecting current I_1 is $\bar{H}_{12}$, which can be found from Ampere's law (1.4.1) computed for a contour C circling I_2 in a right-hand sense:

$$\oint_C \bar{H}\bullet d\bar{s} \cong H_{12}2W = \oiint_A \bar{J}\bullet\hat{n}\,da = I_2 \tag{6.4.12}$$

Thus $\bar{H}_{12} \cong \hat{z}I_2/2W$. The upward pressure on the upper beam found from (6.4.11) and (6.4.12) is then:

$$\bar{P} = \bar{F}/W \cong \hat{x}\mu_o I_1 I_2/2W^2 \ \left[Nm^{-2}\right] \tag{6.4.13}$$

If $I_1 = -I_2$ then the magnetic field between the two closely spaced currents is $H_o{'} = I_1/W$ and (6.4.13) becomes $\bar{p} = \hat{x}\mu_o H_o^{'2}/2$ [N m^{-2}]; this expression for magnetic pressure is derived differently in (6.4.15).

This pressure on the top is downward if both currents flow in the same direction, upward if they are opposite, and zero if either is zero. This device therefore can perform a variety of logic functions. For example, if a switch is arranged so its contacts are closed in state "1" when the

- 176 -

beam is forced upward by both I_1 and I_2 being positive (these currents were defined in the figure as flowing in opposite directions), and not otherwise, this is an "and" gate.

An alternate way to derive magnetic pressure (6.4.13) is to note that if the two currents I_1 and I_2 are anti-parallel, equal, and close together (d $\ll$ W), then $\overline{H} = 0$ outside the two conductors and H_o' is doubled in the gap between them so $WH_o' = I_1$. That is, if the integration contour C circles either current alone then (6.4.12) becomes:

$$\oint_C \overline{H} \bullet d\overline{s} \cong H_o'W = \oiint_A \overline{J} \bullet \hat{n}\, da = I_1 = I_2 \qquad (6.4.14)$$

But not all electrons comprising these currents see the same magnetic field because the currents closer to the two innermost conductor surfaces screen the outer currents, causing the magnetic field to approach zero inside the conductors, as suggested in Figure 6.4.3.
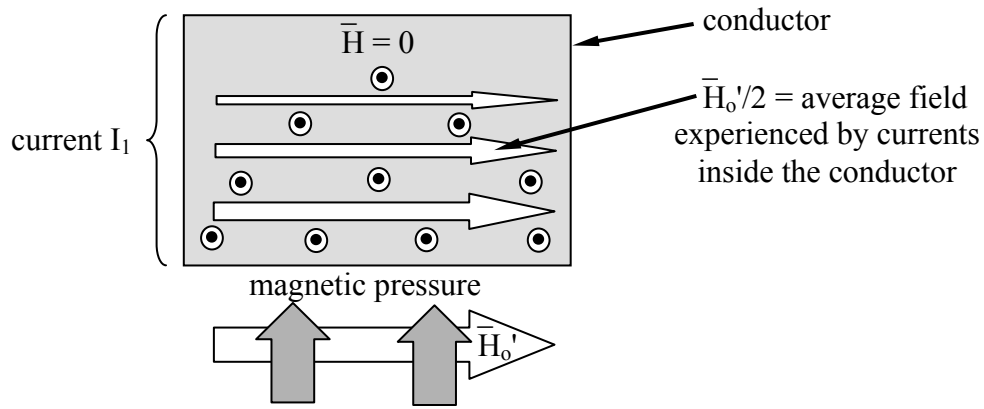
Figure 6.4.3   Surface current and force distribution in a conductor.

Therefore the average moving electron sees a magnetic field $H_o'/2$, half that at the surface[28]. Thus the total *magnetic pressure* upward on the upper beam given by (6.4.13) and (6.4.14) is:

$$\overline{P} = \overline{F}/W = \overline{I}_1 \times \mu_o \overline{H}_o'\Big/2W = \hat{x}\left(H_o'W\right)\left(\mu_o H_o'\big/2W\right) \qquad \text{(magnetic pressure)} \quad (6.4.15)$$
$$= \hat{x}\mu_o H_o'^2\big/2 \ \left[\mathrm{N\,m}^{-2}\right]$$

where H_o' is the total magnetic field magnitude between the two conductors, and there is no magnetic field on the top of the upper beam to press in the opposite direction. This magnetic pressure [N m^{-2}] equals the magnetic energy density [J m^{-3}] stored in the magnetic field adjacent to the conductor (2.7.8).

---

[28] A simple integral of the form used in (5.2.4) yields this same result for pressure.

6.5   *Permanent magnet devices*

6.5.1   <u>Introduction</u>

A *permanent magnet* (Section 2.5.4) has a *residual flux density* $\overline{B}_r$ when $\overline{H} = 0$ inside it, and this is the rest state of an isolated permanent magnet. In this case the magnetic energy density inside is $W_m = \overline{\underline{B}} \bullet \overline{H}/2 = 0$, and that outside, $W_m = \mu_o |\overline{H}|^2 \neq 0$. Boundary conditions (2.6.5) say $\overline{B}_{r\perp} = \mu_o \overline{H}_{o\perp}$, where $\overline{H}_{o\perp}$ is the boundary value in air. Since $\overline{H}_{//}$ is continuous across an insulating boundary and $\overline{H}_r = 0$ inside a resting permanent magnet, $\overline{H}_{o//} = 0$ too. If an external $\overline{H}$ is applied to a permanent magnet, then $\overline{B}$ within that magnet is altered as suggested by the hysteresis diagram in Figure 2.5.3(b).

The force f [N] attracting a permanent magnet to a high-permeability material can be found using:

$$f = dw_m/dx \qquad\qquad\qquad (6.5.1)$$

where x is the separation between the two, as illustrated in Figure 6.5.1, and w_m is the total energy in the magnetic fields [J]. The changing magnetic energy in the high-permeability material is negligible compared to that in air because: 1) boundary conditions require continuity in $\overline{B}_\perp$ across the boundary so that $\overline{B}_\perp = \mu \overline{H}_\perp = \mu_o \overline{H}_{o\perp}$, and therefore $H_\perp/H_{o\perp} = \mu_o/\mu \ll 1$, and 2) W_m [Jm^{-3}] $= \mu |\overline{H}|^2/2$ where $\mu \gg \mu_o$; thus the energy density in air is greater by $\sim\mu/\mu_o \gg 1$.
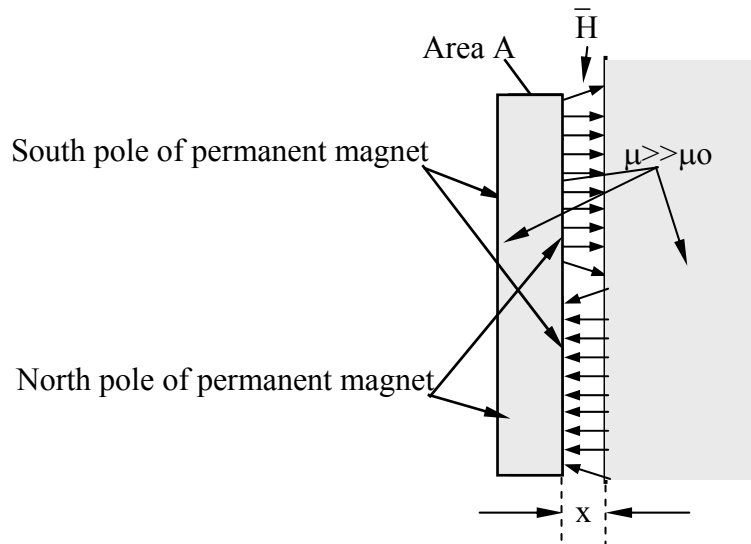

Figure 6.5.1   Permanent magnet adhering to a permeable surface.

The variable magnetic energy is dominated by the energy w_m in the gap, which is the energy density, $W_m = \mu_o H_{gap}^2/2$, times the volume of the gap Ax, where A is the area of the magnet face and x is the gap width. Thus:

$$w_m \cong \mu_o H_{gap}^2 Ax \big/ 2 \ [\text{J}]$$ (6.5.2)

Differentiating w_m with respect to x yields the attractive force $f \cong \mu_o H_{gap}^2 A/2 \ [\text{N}]$, and the force density:

$$F \cong \mu_o H_{gap}^2 \big/ 2 = W_{gap} = B_{gap}^2 \big/ 2\mu_o \ \left[\text{J m}^{-3} \right]$$ (6.5.3)

This can be expressed in terms of B: $F = B_{gap}^2/2\mu_o \ [\text{Nm}^{-2}]$. Note that the rest energy density inside the permanent magnet is zero, so it exerts no pressure. Most permanent magnets have magnetic flux densities B less than one Tesla (10^4 gauss), so a magnet this powerful with an area $A = 10 \ \text{cm}^2$ (~the size of a silver dollar) would therefore apply an attractive force $AF = 0.001 \times 1^2 / 2 \times 4\pi 10^{-7} \cong 400\text{N}$ (~100 pound force). A more typical permanent magnet the same size might attract a steel surface with only a 10–20 pound force.

If two equal coin-shaped permanent magnets are stacked so they stick together, then they experience primarily the attractive magnetic pressure $B_{gap}^2/2\mu_o \ [\text{Nm}^{-2}]$ associated with the gap between them, and are bonded with approximately the same force density as if one of them were merely a high-permeability sheet. In this case $B_{gap} \cong B_r$, as shown in Figure 2.5.3(b).

This simple gap-based magnetic pressure model does not explain the repulsive force between two such coin magnets when one is flipped, however, for then $\overline{H}_g \cong 0$ and $w_{gap} \cong 0$ for all small values of x, and dw_g/dx is also ~0. In this case the energy of interest w_T lies largely inside the magnets. This special case illustrates the risks of casually substituting simple models for the underlying physical reality captured in Maxwell's equations, the Lorentz force law, and material characteristics.

Permanent magnets fail above their *Curie temperature* when the magnetic domains become scrambled. Cooling overheated permanent magnets in a strong external magnetic field usually restores them. Some types of permanent magnets can also fail at very low temperatures, and should not be used where that is a risk.

6.5.2   Permanent magnet motors

Compact high-power-density motors often incorporate permanent magnets so current is not wasted on maintenance of $\overline{\overline{H}}$. For example, the stator for the rotary single-turn coil motor of Figure 6.3.1 could easily contain permanent magnets, avoiding the need for current excitation. Moreover, modern permanent magnets can provide quite intense fields, above 0.5 Tesla. In this case we should also consider the effect of the rotor currents on the stator permanent magnets, whereas in the earlier example we considered the stator fields and rotor currents as given. The

incremental permeability of a permanent magnet varies with the applied H. If H is oriented to attract the stator pole and $\mu_oH > B_r$, then B in the permanent magnet will increase above B_r (see Figure 2.5.3), where the incremental permeability approaches μ_o. To the extent the incremental $\mu > \mu_o$, some reluctance-motor torque will supplement the dominant torque studied earlier.

The permanent magnets can alternatively be placed on the rotor, avoiding the need for rotor currents or a commutator, provided the stator currents are synchronously switched instead. Clever electronics can detect the voltage fluctuations in the stator induced by the rotor and thus deduce its position, potentially avoiding the need for a separate expensive angle encoder for stator current synchronization.

Because different parts of permanent magnets see different B/H histories, and these depend in part on B/H histories elsewhere in the device, modern design of such motors or generators relies extensively on complex software tools for modeling support.

Example 6.5A

Two identical coin-shaped permanent magnets of 12-cm diameter produce 0.05 Tesla field perpendicular to their flat faces; one side is the north pole of the magnet and the other is south. What is the maximum force f attracting the magnets when placed face to face?

Solution: Using (6.5.3) yields $f = AB_{gap}^2/2\mu_o = \pi(0.06)^2(0.05)^2/(2\times1.26\times10^{-6}) = 11.2$ [N].

6.6   Electric and magnetic sensors

6.6.1   Electrostatic MEMS sensors

Sensors are devices that respond to their environment. Some sensors alter their properties as a function of the chemical, thermal, radiation, or other properties of the environment, where a separate active circuit probes these properties. The conductivity, permeability, and permittivity of materials are typically sensitive to multiple environmental parameters. Other sensors directly generate voltages in response to the environment that can be amplified and measured. One common *MEMS sensor* measures small displacements of cantilevered arms due to temperature, pressure, acceleration, chemistry, or other changes. For example, temperature changes can curl a thin cantilever due to differences in thermal expansion coefficient across its thickness, and chemical reactions on the surface of a cantilever can change its mass and mechanical resonance frequency. Microphones can detect vibrations in such cantilevers, or accelerations along specific axes.

Figure 6.6.1 portrays a standard capacitive MEMS sensor that illustrates the basic principles, where the capacitor plates of area A are separated by the distance d, and the voltage V is determined in part by the voltage divider formed by the source resistance R_s and the amplifier input resistance R. V_s is the source voltage.
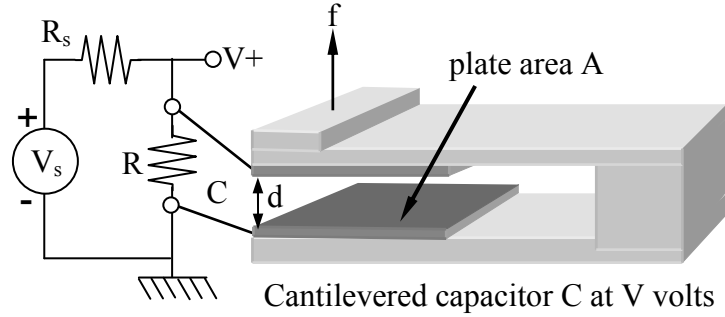
Figure 6.6.1   Capacitive MEMS sensor.

The instantaneous circuit response to an increase δ in the plate separation d is an increase in capacitor voltage V above its normal equilibrium value V_e determined by the voltage divider, where $V_e = V_sR/(R+R_s)$. The capacitor then discharges exponentially toward V_e with a time constant $\tau = (R//R_s)C$.[29] See Section 3.5.1 for further discussions of RC circuit behavior. If R_s >> R then $\tau \cong RC$. If R_s >> R and R represents the input resistance of a high-performance sensor amplifier, then that sensor can detect as little as $\Delta w_B \cong 10^{-20}$ joules per "*bit of information*"[30]. This can be compared to the incremental increase Δw_c in capacitor energy due to the displacement δ << d as C decreases to C':

$$\Delta w_c = (C - C')V^2/2 = V^2\varepsilon_oA\left(d^{-1} - [d+\delta]^{-1}\right)/2 \cong V^2\varepsilon_oA\delta/2d^2 \text{ [J]} \qquad (6.6.1)$$

A simple example illustrates the extreme potential sensitivity of such a sensor. Assume the plate separation d is one micron, the plates are 1-mm square ($A = 10^{-6}$), and V = 300. Then the minimum detectable δ given by (6.6.1) for $\Delta w_c = \Delta w_B = 10^{-20}$ Joules is:

$$\delta_{min} = \Delta w_B \times 2d^2/V^2\varepsilon_oA \cong 10^{-20}2(10^{-6})^2/(300^2 \times 8.8\times10^{-12} \times 10^{-6})$$
$$\cong 2\times10^{-20} \text{ [m]} \qquad (6.6.2)$$

At this potential level of sensitivity we are limited instead by thermal and mechanical noise due to the Brownian motion of air molecules and conduction electrons. A more practical set of parameters might involve a less sensitive detector ($\Delta_B \cong 10^{-14}$) and lower voltages (V $\cong$ 5); then $\delta_{min} \cong 10^{-10}$ meters $\cong$ 1 angstrom (very roughly an atomic diameter). The dynamic range of such a sensor would be enormously greater, of course. This one-angstrom sensitivity is comparable to that of the human eardrum at ~1kHz.

---

[29] The resistance R of two resistors in parallel is R = (R_a || R_b) = $R_aR_b/(R_a + R_b)$.

[30] Most good communications systems can operate with acceptable probabilities of error if E_b/N_o >~10, where E_b is the energy per bit and $N_o = kT$ is the *noise power density* [W Hz^{-1}] = [J]. A bit is a single yes-no piece of information. Boltzmann's constant k $\cong$ 1.38×10^{-23} [J $^oK^{-1}$], and T is the *system noise temperature*, which might approximate 100K in a good system at RF frequencies. Thus the minimum energy required to detect each bit of information is ~10N$_o$ = 10 kT $\cong$ 10^{-20} [J].

An alternative to such observations of MEMS sensor voltage transients is to observe changes in resonant frequency of an LC resonator that includes the sensor capacitance; this approach can reduce the effects of low-frequency interference.

6.6.2 Magnetic MEMS sensors

Microscopic magnetic sensors are less common than electrostatic ones because of the difficulty of providing strong inexpensive reliable magnetic fields at microscopic scales. High magnetic fields require high currents or strong permanent magnets. If such fields are present, however, mechanical motion of a probe wire or cantilever across the magnetic field lines could produce fluctuating voltages, as given by (6.1.4).

6.6.3 Hall effect sensors

Hall effect sensors are semiconductor devices that produce an output voltage V_{Hall} proportional to magnetic field $\overline{\overline{H}}$, where the voltage is produced as a result of magnetic forces on charge carriers moving at velocity $\overline{v}$ within the semiconductor. They can measure magnetic fields or, if the magnetic field is known, can determine the average velocity and type (hole or electron) of the charge carriers conveying current. A typical configuration appears in Figure 6.6.2, for which the Hall-effect voltage V_{Hall} is proportional to the current I and to the perpendicular magnetic field $\overline{H}$.
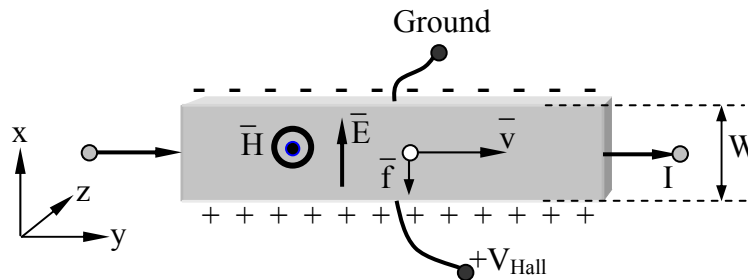
Figure 6.6.2   Hall effect sensor.

The operation of a Hall-effect sensor follows directly from the Lorentz force law:

$$\overline{f} = q\left(\overline{E} + \overline{v} \times \mu_o \overline{H}\right) \quad [\text{Newtons}] \tag{6.6.3}$$

Positively charged carriers moving at velocity $\overline{v}$ would be forced downward by $\overline{f}$, as shown in the figure, where they would accumulate until the resulting electric field $\overline{E}$ provided a sufficiently strong balancing force qE in the opposite direction to produce equilibrium. In equilibrium the net force and the right-hand side of (6.6.3) must be zero, so $\overline{E} = -\overline{v} \times \mu_o \overline{H}$ and the resulting V_{Hall} is:

$$V_{Hall} = \hat{x} \bullet \overline{E} W = v \mu_o H W \quad [V] \tag{6.6.4}$$

For charge carrier velocities of 30 m s^{-1} and fields $\mu_o H$ of 0.1 Tesla, V_{Hall} would be 3 millivolts across a width W of one millimeter, which is easily detected.

If the charge carriers are electrons so q < 0, then the sign of the Hall voltage is reversed. Since the voltage depends on the velocity v of the carriers rather than on their number, their average number density N [m^{-1}] can be determined using I = Nqv. That is, for positive carriers:

$$v = V_H/W\mu_o H \; \left[ms^{-1}\right] \tag{6.6.5}$$

$$N = I/qv \tag{6.6.6}$$

Thus the Hall effect is useful for understanding carrier behavior (N,v) as a function of semiconductor composition.

Chapter 7: TEM Transmission Lines

7.1   TEM waves on structures

7.1.1   Introduction

Transmission lines typically convey electrical signals and power from point to point along arbitrary paths with high efficiency, and can also serve as circuit elements. In most transmission lines, the electric and magnetic fields point purely transverse to the direction of propagation; such waves are called transverse electromagnetic or *TEM waves*, and such transmission lines are called *TEM lines*. The basic character of TEM waves is discussed in Section 7.1, the effects of junctions are introduced in Section 7.2, and the uses and analysis of TEM lines with junctions are treated in Section 7.3. Section 7.4 concludes by discussing TEM lines that are terminated at both ends so as to form resonators.

Transmission lines in communications systems usually exhibit frequency-dependent behavior, so complex notation is commonly used. Such lines are the subject of this chapter. For broadband signals such as those propagating in computers, complex notation can be awkward and the physics obscure. In this case the signals are often analyzed in the time domain, as introduced in Section 7.1.2 and discussed further in Section 8.1. Non-TEM transmission lines are commonly called waveguides; usually the waves propagate inside some conducting envelope, as discussed in Section 9.3, although sometimes they propagate partly outside their guiding structure in an "open" waveguide such as an optical fiber, as discussed in Section 12.2.

7.1.2   TEM waves between parallel conducting plates

The sinusoidal uniform plane wave of equations (7.1.1) and (7.1.2) is consistent with the presence of thin parallel conducting plates orthogonal to the electric field $\overline{E}(z,t)$, as illustrated in Figure 7.1.1(a)[31].

$$\overline{E}(z,t) = \hat{x}E_o \cos(\omega t - kz) \ \text{[V/m]} \tag{7.1.1}$$

$$\overline{H}(z,t) = \hat{y}\frac{E_o}{\eta_o}\cos(\omega t - kz) \ \text{[A/m]} \tag{7.1.2}$$

Although perfect consistency requires that the plates be infinite, there is approximate consistency so long as the plate separation d is small compared to the plate width W and the fringing fields outside the structure are negligible.   The more general wave $\overline{E}(z,t) = \hat{x}E_x(z-ct)$, $\overline{H}(z,t) = \hat{z} \times \overline{E}(z,t)/\eta_o$ is also consistent [see (2.2.13), (2.2.18)], since any arbitrary waveform E(z - ct) can be expressed as the superposition of sinusoidal waves at all frequencies. In both cases all boundary conditions of Section 2.6 are satisfied because $\overline{E}_{//} = \overline{H}_\perp = 0$ at the

---

[31] See Section 2.3.1 for an introduction to uniform sinusoidal electromagnetic plane waves.

conductors. The voltage between two plates v(z,t) for this sinusoidal wave can be found by integrating $\overline{E}(z,t)$ over the distance d from the lower plate, which we associate here with the voltage +v, to the upper plate:

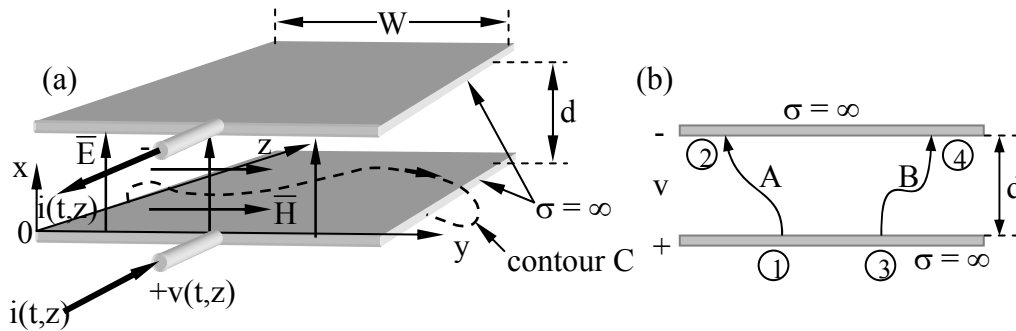$$v(t,z) = \hat{x}\cdot\overline{E}(z,t)d = E_o d\cos(\omega t - kz) \quad [V] \tag{7.1.3}$$

Figure 7.1.1   Parallel-plate TEM transmission line.

Although this computed voltage v(t,z) does not depend on the path of integration connecting the two plates, provided it is at constant z, it does depend on z itself. Thus there can be two different voltages between the same pair of plates at different positions z. Kirchoff's voltage law says that the sum of voltage drops around a loop is zero; this law is violated here because such a loop in the x-z plane encircles time varying magnetic fields, $\overline{H}(z,t)$, as illustrated. In contrast, the sum of voltage drops around a loop confined to constant z is zero because it circles no $\partial\overline{H}/\partial t$; therefore the voltage v(z,t), computed by integrating $\overline{E}(z)$ between the two plates, does not depend on the path of integration at constant z. For example, the integrals of $\overline{E}\bullet d\overline{s}$ along contours A and B in Figure 7.1.1(b) must be equal because the integral around the loop 1, 2, 4, 3, 1 is zero and the path integrals within the perfect conductors both yield zero.

If the electric and magnetic fields are zero outside the two plates and uniform between them, then equal and opposite currents i(t,z) flow in the two plates in the ±z direction. The surface current is determined by the boundary condition (2.6.17): $\overline{J}_S = \hat{n}\times\overline{H}$ [A m^{-1}]. If the two conducting plates are spaced close together compared to their widths W so that d << W, then the fringing fields at the plate edges can be neglected and the total current flowing in the plates can be found from the given magnetic field $\overline{H}(z,t) = \hat{y}(E_o/\eta_o)\cos(\omega t - kz)$, and the integral form of Ampere's law:

$$\int_C \overline{H}\bullet d\overline{s} = \iint_A \left[\overline{J} + (\partial\overline{D}/\partial t)\right]\bullet\hat{n}\ da \tag{7.1.4}$$

If the integration contour C encircles the lower plate and surface A at constant z in a clockwise (right-hand) sense with respect to the +z axis as illustrated in Figure 7.1.1, then $\overline{D}\bullet\hat{n} = 0$ and the current flowing in the +z direction in the lower plate is simply:

$$i(z,t) = W\, J_{sz}(z,t) = W\, H_y(z,t) = \left(WE_o/\eta_o\right)\cos\left(\omega t - kz\right)\ [A] \qquad (7.1.5)$$

An equal and opposite current flows in the upper plate.

Note that the computed current does not depend on the integration contour C chosen so long as C circles the plate at constant z. Also, the current flowing into a section of conducting plate at z_1 does not generally equal the current flowing out at z_2, seemingly violating Kirchoff's current law (the sum of currents flowing into a node is zero). This inequality exists because any section of parallel plates exhibits capacitance that conveys a displacement current $\partial\overline{D}/\partial t$ between the two plates; the right-hand side of Equation (2.1.6) suggests the equivalent nature of the conduction current density $\overline{J}$ and the displacement current density $\partial\overline{D}/\partial t$.

Such a two-conductor structure conveying waves that are purely transverse to the direction of propagation, i.e., $E_z = H_z = 0$, is called a *TEM transmission line* because it is propagating transverse electromagnetic waves (*TEM waves*). Such lines generally have a physical cross-section that is independent of z. This particular TEM transmission line is called a *parallel-plate TEM line*.

Because there are no restrictions on the time structure of a plane wave, any v(t) can propagate between parallel conducting plates. The ratio between v(z,t) and i(z,t) for this or any other sinusoidal or non-sinusoidal forward traveling wave is the *characteristic impedance Z_o* of the TEM structure:

$$v(z,t)/i(z,t) = \eta_o d/W = Z_o\ [\text{ohms}] \qquad \text{(characteristic impedance)} \qquad (7.1.6)$$

In the special case $d = W$, Z_o equals the characteristic impedance η_o of free space, 377 ohms. Usually $W \gg d$ in order to minimize fringing fields, yielding $Z_o \ll 377$.

Since the two parallel plates can be perfectly conducting and lossless, the physical significance of Z_o ohms may be unclear. Z_o is defined as the ratio of line voltage to line current for a forward wave only, and is non-zero because the plates have inductance L per meter associated with the magnetic fields within the line. The value of Z_o also depends on the capacitance C per meter of this structure. Section 7.1.3 shows (7.1.59) that $Z_o = (L/C)^{0.5}$ for any lossless TEM line and (7.1.19) shows it for a parallel-plate line. The product of voltage and current v(z,t)i(z,t) represents power P(z,t) flowing past any point z toward infinity; this power is not being converted to heat by resistive losses, it is simply propagating away without reflections.

It is easy to demonstrate that the power P(z,t) carried by this forward traveling wave is the same whether it is computed by multiplying v and i, or by integrating the Poynting vector $\overline{S} = \overline{E} \times \overline{H}$ [W m^{-2}] over the cross-sectional area Wd of the TEM line:

$$P(z,t) = v(z,t)i(z,t) = \left[E(z,t)d\right]\left[H(z,t)W\right] = \left[E(z,t)H(z,t)\right]Wd = S\,Wd \qquad (7.1.7)$$

The differential equations governing v and i on TEM lines are easily derived from Faraday's and Ampere's laws for the fields between the plates of this line:

$$\nabla \times \overline{E} = -(\partial/\partial t)\mu\overline{H} = \hat{y}(\partial/\partial z)E_x(z,t) \tag{7.1.8}$$

$$\nabla \times \overline{H} = (\partial/\partial t)\varepsilon\overline{E} = -\hat{x}(\partial/\partial z)H_y(z,t) \tag{7.1.9}$$

Because all but one term in the curl expressions are zero, these two equations are quite simple. By substituting $v = E_x d$ (7.1.3) and $i = H_y W$ (7.1.5), (7.1.8) and (7.1.9) become:

$$dv/dz = -(\mu d/W)(di/dt) = -L\,di/dt \tag{7.1.10}$$

$$di/dz = -(\varepsilon W/d)(dv/dt) = -C\,dv/dt \tag{7.1.11}$$

where we have used the expressions for *inductance per meter* L [Hy m^{-1}] and *capacitance per meter* C [F m^{-1}] of a parallel-plate TEM line [see (3.2.11)[32] and (3.1.10)]. This form of the differential equations in terms of L and C applies to any lossless TEM line, as shown in Section 7.1.3.

These two differential equations can be solved for v by eliminating i. The current i can be eliminated by differentiating (7.1.10) with respect to z, and (7.1.11) with respect to t, thus introducing $d^2i/(dt\ dz)$ into both expressions permitting its substitution. That is:

$$d^2v/dz^2 = -L\,d^2i/(dt\ dz) \tag{7.1.12}$$

$$d^2i/(dz\ dt) = -C\,d^2v/dt^2 \tag{7.1.13}$$

Combining these two equations by eliminating $d^2i/(dt\ dz)$ yields the wave equation:

$$d^2v/dz^2 = LC\,d^2v/dt^2 = \mu\varepsilon\,d^2v/dt^2 \qquad \text{(wave equation)} \tag{7.1.14}$$

Wave equations relate the second spatial derivative to the second time derivative of the same variable, and the solution therefore can be any arbitrary function of an argument that has the same dependence on space as on time, except for a constant multiplier. That is, one solution to (7.1.14) is:

$$v(z,t) = v_+(z-ct) \tag{7.1.15}$$

where v_+ is an arbitrary function of the argument $(z-ct)$ and is associated with waves propagating in the +z direction at velocity c. This is directly analogous to the propagating waves

---

[32] Note: (3.2.11) gives the total inductance L for a length D of line, where area A = Dd. The inductance per unit length L = μd/W in both cases.

characterized in Figure 2.2.1 and in Equation (2.2.9). Demonstration that (7.1.15) satisfies (7.1.14) for $c = (\mu\epsilon)^{-0.5}$ follows the same proof provided for (2.2.9) in (2.2.10–12).

The general solution to (7.1.14) is any arbitrary waveform of the form (7.1.15) plus an independent arbitrary waveform propagating in the -z direction:

$$v(z,t) = v_+(z - ct) + v_-(z + ct) \qquad (7.1.16)$$

The general expression for current i(z,t) on a TEM line can be found, for example, by substituting (7.1.16) into the differential equation (7.1.11) and integrating over z. Thus, using the notation that $v'(q) \equiv dv(q)/dq$:

$$di/dz = -Cdv/dt = cC\left[v'_+(z - ct) - v'_-(z + ct)\right] \qquad (7.1.17)$$

$$i(z,t) = cC\left[v_+(z - ct) - v_-(z + ct)\right] = Z_o^{-1}\left[v_+(z - ct) - v_-(z + ct)\right] \qquad (7.1.18)$$

Equation (7.1.18) defines the characteristic impedance $Z_o = (cC)^{-1} = \sqrt{L/C}$ for the TEM line. Both the forward and backward waves alone have the ratio Z_o between v and i, although the sign of i is reversed for the negative-propagating wave because a positive voltage then corresponds to a negative current. These same TEM results are derived differently in Sections 7.1.3 and 8.1.1.

The characteristic impedance Z_o of a parallel-plate line can be usefully related using (7.1.18) to the capacitance C and inductance L per meter, where $C = \epsilon W/d$ and $L = \mu d/W$ for parallel-plate structures (7.1.10–11):

$$Z_o = \sqrt{\frac{L}{C}} \text{ [ohms]} = \frac{d}{c\epsilon W} = \sqrt{\frac{\mu}{\epsilon}}\frac{d}{W} \qquad \text{(characteristic impedance)} \qquad (7.1.19)$$

All lossless TEM lines have this simple relationship, as seen in (8.3.9) for $R = G = 0$. It is also consistent with (7.1.6), where $\eta_o = 1/c\epsilon = (\mu_o/\epsilon_o)^{0.5}$.

The electric and magnetic energies per meter on a parallel-plate TEM line of plate separation d and plate width W are:[33]

$$W_e(t,z) = \frac{1}{2}\epsilon\left|\overline{E}(t,z)\right|^2 = \frac{1}{2}\epsilon\left(\frac{v(t,z)}{d}\right)^2 Wd \;\; \left[\text{J m}^{-1}\right] \qquad (7.1.20)$$

$$W_m(t,z) = \frac{1}{2}\mu\left|\overline{H}(t,z)\right|^2 = \frac{1}{2}\mu\left(\frac{i(t,z)}{d}\right)^2 Wd \;\; \left[\text{J m}^{-1}\right] \qquad (7.1.21)$$

---

[33] Italicized symbols for W_e and W_m [J m^{-1}] distinguish them from W_e and W_m [J m^{-3}].

Substituting C = cW/d and L = μd/W into (7.1.20) and (7.1.21) yields:

$$W_e(t,z) = \frac{1}{2}Cv^2 \left[Jm^{-1} \right] \qquad\qquad \text{(TEM electric energy density)} \qquad (7.1.22)$$

$$W_m(t,z) = \frac{1}{2}Li^2 \left[Jm^{-1} \right] \qquad\qquad \text{(TEM magnetic energy density)} \qquad (7.1.23)$$

If there is only a forward-moving wave, then $v(t,z) = Z_o i(t,z)$ and so:

$$W_e(t,z) = \frac{1}{2}Cv^2 = \frac{1}{2}CZ_o^2 i^2 = \frac{1}{2}Li^2 = W_m(t,z) \qquad\qquad (7.1.24)$$

These relations (7.1.22) to (7.1.24) are true for any TEM line.

The same derivations can be performed using complex notation. Thus (7.1.10) and (7.1.11) can be written:

$$\frac{d\underline{V}(z)}{dz} = -\frac{\mu d}{W} j\omega \underline{I}(z) = -j\omega L \underline{I}(z) \qquad\qquad (7.1.25)$$

$$\frac{d\underline{I}(z)}{dz} = -\frac{\varepsilon W}{d} j\omega \underline{V}(z) = -j\omega C \underline{V}(z) \qquad\qquad (7.1.26)$$

Eliminating I(z) from this pair of equations yields the wave equation:

$$\left(\frac{d^2}{dz^2} + \omega^2 LC \right) \underline{V}(z) = 0 \qquad\qquad \text{(wave equation)} \qquad (7.1.27)$$

The solution to the wave equation (7.1.27) is the sum of forward and backward propagating waves with complex magnitudes that indicate phase:

$$\underline{V}(z) = \underline{V}_+ e^{-jkz} + \underline{V}_- e^{+jkz} \qquad\qquad (7.1.28)$$

$$\underline{I}(z) = Y_o (\underline{V}_+ e^{-jkz} - \underline{V}_- e^{+jkz}) \qquad\qquad (7.1.29)$$

where the *wavenumber* k follows from $k^2 = \omega^2 LC$, which is obtained by substituting (7.1.28) into (7.1.27):

$$k = \omega\sqrt{LC} = \frac{\omega}{c} = \frac{2\pi}{\lambda} \qquad\qquad (7.1.30)$$

The characteristic impedance of the line, as seen in (7.1.19) is:

$$Z_o = \sqrt{\frac{L}{C}} = \frac{1}{Y_o} \quad \text{[ohms]} \tag{7.1.31}$$

and the time average stored electric and magnetic energy densities are:

$$W_e = \frac{1}{4}C\,|\,\underline{V}\,|^2 \quad \text{[J/m]}, \qquad W_m = \frac{1}{4}L\,|\,\underline{I}\,|^2 \quad \text{[J/m]} \tag{7.1.32}$$

The behavior of these arbitrary waveforms at TEM junctions is discussed in the next section and the practical application of these general solutions for arbitrary waveforms is discussed further in Section 8.1. Their practical application to sinusoidal waveforms is discussed in Sections 7.2–4.

Example 7.1A

A certain TEM line consists of two parallel metal plates that are 10 cm wide, separated in air by d = 1 cm, and extremely long. A voltage v(t) = 10 cos ωt volts is applied to the plates at one end (z = 0). What currents i(t,z) flow? What power P(t) is being fed to the line? If the plate resistance is zero, where is the power going? What is the inductance L per unit length for this line?

Solution: In a TEM line the ratio $v/i = Z_o$ for a single wave, where $Z_o = \eta_o d/W$ [see (7.1.6)], and $\eta_o = (\mu/\varepsilon)^{0.5} \cong 377$ ohms in air. Therefore $i(t,z) = Z_o^{-1}v(t,z) = (W/d\eta_o)10\cos(\omega t - kz) \cong [0.1/(0.01 \times 377)]\, 10\cos(\omega t - kz) \cong 0.27\cos[\omega(t - z/c)]$ [A]. $P = vi = v^2/Z_o \cong 2.65\cos^2[\omega(t - z/c)]$ [W]. The power is simply propagating losslessly along the line toward infinity. Since $c = (LC)^{-0.5} = 3 \times 10^8$, and $Z_o = (L/C)^{0.5} \cong 37.7$, therefore $L = Z_o/c = 1.3 \times 10^{-7}$ [Henries m^{-1}].

7.1.3   TEM waves in non-planar transmission lines

TEM waves can propagate in any perfectly conducting structure having at least two non-contacting conductors with an arbitrary cross-section independent of z, as illustrated in Figure 7.1.2, if they are separated by a uniform medium characterized by ε, μ, and σ. The parallel plate TEM transmission line analyzed in Section 7.1.2 is a special case of this configuration, and we shall see that the behavior of non-planar TEM lines is characterized by the same differential equations for v(z,t) and i(z,t), (7.1.10) and (7.1.11), when expressed in terms of L and C. This result follows from the derivation below.

We first divide the del operator into its transverse and longitudinal (z-axis) components:

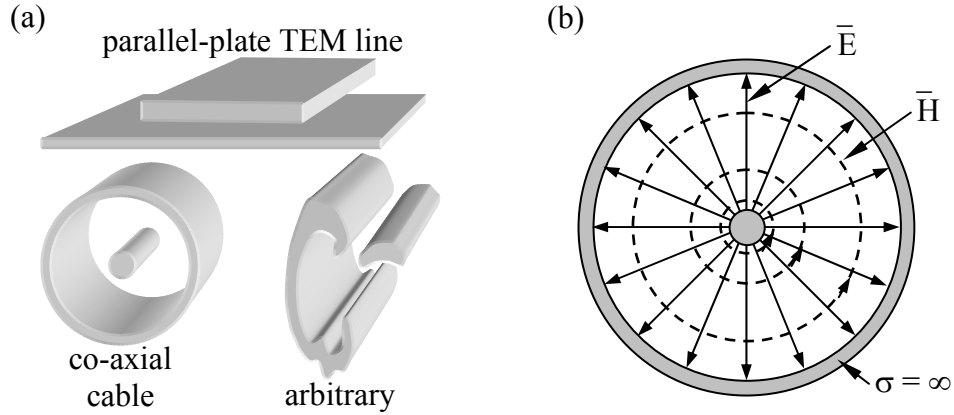$$\nabla = \nabla_T + \hat{z}\partial/\partial z \tag{7.1.33}$$

Figure 7.1.2   TEM lines with arbitrary cross-sections.

where $\nabla_T \equiv \hat{x}\partial/\partial x + \hat{y}\partial/\partial y$. Faraday's and Ampere's laws then become:

$$\nabla \times \overline{E} = \nabla_T \times \overline{E}_T + (\partial/\partial z)(\hat{z} \times \overline{E}_T) = -\mu \partial \overline{H}_T/\partial t \tag{7.1.34}$$

$$\nabla \times \overline{H} = \nabla_T \times \overline{H}_T + (\partial/\partial z)(\hat{z} \times \overline{H}_T) = \sigma \overline{E}_T + \varepsilon \partial \overline{E}_T/\partial t \tag{7.1.35}$$

The right-hand sides of these two equations have no $\hat{z}$ components, and therefore the transverse curl components on the left-hand side are zero because they lie only along the z axis:

$$\nabla_T \times \overline{E}_T = \nabla_T \times \overline{H}_T = 0 \tag{7.1.36}$$

Moreover, the divergences of $\overline{E}_T$ and $\overline{H}_T$ are also zero since $\hat{z} \bullet \overline{H}_T = \hat{z} \bullet \overline{E}_T = 0$, and:

$$\nabla \bullet \overline{H} = 0 = \nabla_T \bullet \overline{H}_T + (\partial/\partial z)(\hat{z} \bullet \overline{H}_T) \tag{7.1.37}$$

$$\nabla \bullet \overline{E} = \rho/\varepsilon = 0 = \nabla_T \bullet \overline{E}_T + (\partial/\partial z)(\hat{z} \bullet \overline{E}_T) \tag{7.1.38}$$

Since the curl and divergence of $\overline{E}_T$ and $\overline{H}_T$ are zero, both these fields must independently satisfy Laplace's equation (4.5.7), which governs electrostatics and magnetostatics; these field solutions will differ because their boundary conditions differ. Thus we can find the transverse electric and magnetic fields for TEM lines with arbitrary cross-sections using the equation-solving and field mapping methods described in Sections 4.5 and 4.6.

The behavior of $\overline{E}$ and $\overline{H}$ for an arbitrary TEM line can be expressed more simply if we first define the line's *capacitance per meter* C and the *inductance per meter* L. C is the charge Q' per unit length divided by the voltage v between the two conductors of interest, and L is the flux linkage Λ' per unit length divided by the current i. Capacitance, inductance, and flux linkage are discussed more fully in Sections 3.1 and 3.2.
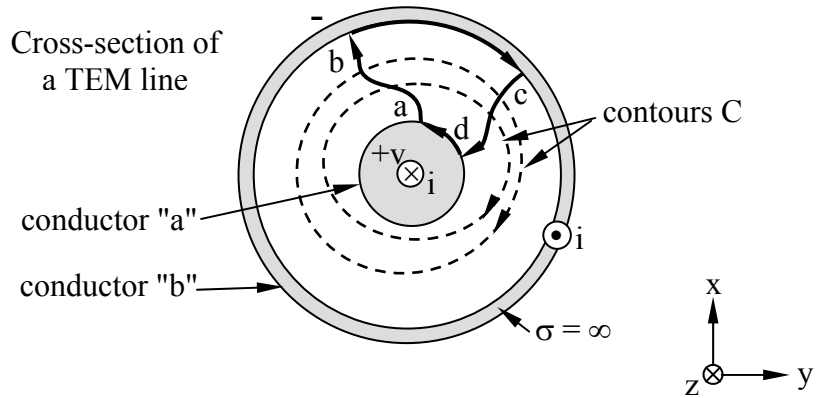
Figure 7.1.3   Integration paths for computing TEM line voltages and currents.

To compute Q' and Λ' we consider a differential element of length δ along the z axis of the TEM line illustrated in Figure 7.1.3, and then compute for Q' and Λ', respectively, surface and line integrals encircling the central positively charged conducting element "a" in a right-hand sense relative to $\hat{z}$. To compute the voltage v we integrate $\overline{E}_T$ from element a to element b, and to compute the current i we integrate $\overline{H}_T$ in a right-hand sense along the contour C circling conductor a:

$$C = Q'/v = \left(\delta^{-1} \iint_A \varepsilon \overline{E}_T \bullet \hat{n} \, da\right) \Big/ \left(\int_a^b \overline{E}_T \bullet d\overline{s}\right)$$
$$= \left[\oint_C \varepsilon \hat{z} \bullet \left(\overline{E}_T \times d\overline{s}\right)\right] \Big/ \left(\int_a^b \overline{E}_T \bullet d\overline{s}\right) \ \left[Fm^{-1}\right]$$
(capacitance/m)   (7.1.39)

$$L = \Lambda'/i = \left[-\int_a^b \mu \hat{z} \bullet \left(\overline{H}_T \times d\overline{s}\right)\right] \Big/ \left(\oint_C \overline{H}_T \bullet d\overline{s}\right)$$
$$= \left[\int_a^b \mu \overline{H}_T \bullet \left(\hat{z} \times d\overline{s}\right)\right] \Big/ \left(\oint_C \overline{H}_T \bullet d\overline{s}\right) \ \left[Hm^{-1}\right]$$
(inductance/m)   (7.1.40)

It is also useful to define G, the line *conductance per meter*, in terms of the leakage current density J_σ' [A m^{-1}] conveyed between the two conductors by the conductivity σ of the medium, where we can use (7.1.39) to show:

$$G = J_\sigma'/v = \left(\delta^{-1} \iint_A \sigma \overline{E}_T \bullet \hat{n} \, da\right) \Big/ \left(\int_a^b \overline{E}_T \bullet d\overline{s}\right) = C\sigma/\varepsilon$$
(7.1.41)

We can readily prove that the voltage and current computed using line integrals in (7.1.39–41) do not depend on the integration path. Figure 7.1.3 illustrates two possible paths of integration for computing v within a plane corresponding to a single value of z, the paths ab and dc. Since the curl of $\overline{E}_T$ is zero in the transverse plane we have:

- 193 -

$$\oint_C \overline{E}_T \bullet d\overline{s} = \int_a^b \overline{E}_T \bullet d\overline{s} + \int_b^c \overline{E}_T \bullet d\overline{s} + \int_c^d \overline{E}_T \bullet d\overline{s} + \int_d^a \overline{E}_T \bullet d\overline{s} = 0 \qquad (7.1.42)$$

The line integrals along the conductors are zero (paths bc and da), and the cd path is the reverse of the dc path. Therefore voltage is uniquely defined because for any path dc we have:

$$\int_a^b \overline{E}_T \bullet d\overline{s} = \int_d^c \overline{E}_T \bullet d\overline{s} = v(z, t) \qquad (7.1.43)$$

The current i(z,t) is also uniquely defined because all possible contours C in Figure 7.1.3 circle the same current flowing in conductor a:

$$i(z, t) = \oint_C \overline{H}_T \bullet d\overline{s} \qquad (7.1.44)$$

To derive the differential equations governing v(z,t) and i(z,t) we begin with (7.1.34) and (7.1.35), noting that $\nabla_T \times \overline{E}_T = \nabla_T \times \overline{H}_T = 0$:

$$(\partial/\partial z)(\hat{z} \times \overline{E}_T) = -\mu \partial \overline{H}_T / \partial t \qquad (7.1.45)$$

$$(\partial/\partial z)(\hat{z} \times \overline{H}_T) = (\sigma + \varepsilon \partial/\partial t) \overline{E}_T \qquad (7.1.46)$$

To convert (7.1.45) into an equation in terms of v we can compute the line integral of $\overline{E}_T$ from a to b: the first step is to use the identity $\overline{A} \times (\overline{B} \times \overline{C}) = \overline{B}(\overline{A} \bullet \overline{C}) - \overline{C}(\overline{A} \bullet \overline{B})$ to show $(\hat{z} \times \overline{E}_T) \times \hat{z} = \overline{E}_T$. Using this we operate on (7.1.45) to yield:

$$\begin{aligned} (\partial/\partial z) \int_a^b \left[(\hat{z} \times \overline{E}_T) \times \hat{z} \right] \bullet d\overline{s} &= (\partial/\partial z) \int_a^b \overline{E}_T \bullet d\overline{s} \\ &= \partial v(z, t)/\partial z \\ &= -\mu (\partial/\partial t) \int_a^b (\overline{H}_T \times \hat{z}) \bullet d\overline{s} \end{aligned} \qquad (7.1.47)$$

Then the right-hand integral in (7.1.47), in combination with (7.1.40) and (7.1.44), becomes:

$$\int_a^b (H_T \times \hat{z}) \bullet d\overline{s} = \int_a^b \overline{H}_T \bullet (\hat{z} \times d\overline{s}) = \mu^{-1} L \oint_C \overline{H}_T \bullet d\overline{s} = \mu^{-1} L i(z, t) \qquad (7.1.48)$$

Combining (7.1.47) and (7.1.48) yields:

$$\partial v(z, t)/\partial z = -L \partial i(z, t)/\partial t \qquad (7.1.49)$$

A similar contour integration of $\overline{H}_T$ to yield i(z,t) simplifies (7.1.46):

$$(\partial/\partial z)\int_C \left[(\hat{z}\times\overline{H}_T)\times\hat{z}\right]\bullet d\overline{s} = (\partial/\partial z)\oint_C \overline{H}_T \bullet d\overline{s} = \partial i/\partial z = (\sigma + \varepsilon\partial/\partial z)\oint_C (\overline{E}_T \times \hat{z})\bullet d\overline{s} \qquad (7.1.50)$$

The definitions of C (7.1.39) and G (7.1.41), combined with $(\overline{E}\times\hat{z})\bullet d\overline{s} = (\overline{E}_T \times d\overline{s})\bullet\hat{z}$ and the definition (7.1.43) of v, yields:

$$\partial i(z,t)/\partial z = -(G + C\partial/\partial t)v(z,t) \qquad (7.1.51)$$

This pair of equations, (7.1.49) and (7.1.51), can then be combined to yield a more complete description of wave propagation on general TEM lines.

Because the characteristic impedance and phase velocity for general TEM lines are frequency dependent, the simple solutions (7.1.49) and (7.1.51) are not convenient. Instead it is useful to express them as complex functions of ω:

$$\partial\underline{V}(z)/\partial z = -j\omega L\underline{I}(z) \qquad (7.1.52)$$

$$\partial\underline{I}(z)/\partial z = -(G + j\omega C)\underline{V}(z) \qquad (7.1.53)$$

Combining this pair of equations yields the wave equation:

$$\partial^2\underline{V}(z)/\partial z^2 = j\omega L(G + j\omega C)\underline{V}(z) \qquad \text{(TEM wave equation)} \qquad (7.1.54)$$

The solution to this *TEM wave equation* must be a function that equals a constant times its own second derivative, such as:

$$\underline{V}(z) = \underline{V}_+e^{-j\underline{k}z} + \underline{V}_-e^{+j\underline{k}z} \qquad \text{(wave equation solution)} \qquad (7.1.55)$$

Substituting this assumed solution into the wave equation yields the *dispersion relation* for general TEM lines made with perfect conductors:

$$\underline{k}^2 = -j\omega L(G + j\omega C) \qquad \text{(TEM dispersion relation)} \qquad (7.1.56)$$

This equation yields a complex value for the *TEM propagation constant* $\underline{k}$ = k' - jk", the significance of which is that the forward (V_+) and backward (V_-) propagating waves are exponentially attenuated with distance:

$$\underline{V}(z) = \underline{V}_+e^{-jk'z-k"z} + \underline{V}_-e^{+j\underline{k}'z+k"z} \qquad (7.1.57)$$

The current can be found by substituting (7.1.57) into (7.1.53) to yield:

$$\underline{I}(z) = (\underline{k}/j\omega L)\left(\underline{V}_+e^{-j\underline{k}z} - \underline{V}_-e^{+j\underline{k}z}\right) = \left(\underline{V}_+e^{-j\underline{k}z} - \underline{V}_-e^{+j\underline{k}z}\right)\Big/\underline{Z}_o \qquad (7.1.58)$$

$$\underline{Z}_o = \left[j\omega L/(G + j\omega C) \right]^{0.5} \tag{7.1.59}$$

These expressions reduce to those for lossless TEM lines as $G \rightarrow 0$.

Another consequence of this dispersion relation (7.1.56) is that the *TEM phase velocity* v_p is frequency dependent and thus most lossy lines are dispersive:

$$v_p = \omega/k' = (LC)^{-0.5} (1 - jG/\omega C)^{-0.5} \tag{7.1.60}$$

Although most TEM lines also have resistance R per unit length, this introduces $E_z \neq 0$, so analysis becomes much more complex. In this case the approximate Telegrapher's equations (8.3.3–4) are often used.

Example 7.1B

What is the characteristic impedance Z_o for the air-filled *co-axial cable* illustrated in Figure 7.1.3 if the relevant diameters for the inner and outer conductors are a and b, respectively, where b/a = e? "Co-axial" means cylinders a and b share the same axis of symmetry.

Solution: $Z_o = (L/C)^{0.5}$ from (7.1.59). Since $c = (LC)^{-0.5}$ it follows that $L = (c^{-2}C^{-1})$ and $Z_o = 1/cC$ ohms. C follows from (7.1.39), which requires knowledge of the transverse electric field $\overline{E}_T$ (for TEM waves, there are no non-transverse fields). Symmetry in this cylindrical geometry requires $\overline{E}_T = \hat{r}E_o/r$. Thus

$$C = Q'/v = \left[\oiint_A \varepsilon_o \overline{E}_T \bullet \hat{r} \, da \right] \Big/ \left[\int_a^b \overline{E}_T \bullet d\overline{s} \right] = a^{-1} \left[\varepsilon_o E_o 2\pi a \right] \Big/ \left[\int_a^b E_o r^{-1} dr \right]$$

$$= \varepsilon_o 2\pi \ln(b/a) = 2\pi\varepsilon_o = 56 \times 10^{-12} \text{ [F]. Therefore } Z_o = (56 \times 10^{-12} \times 3 \times 10^8)^{-1}$$

$$\cong 60 \text{ ohms, and } L \cong 2 \times 10^{-7} \text{ [H]}.$$

7.1.4   Loss in transmission lines

Transmission line losses can be computed in terms of the resistance R, Ohms per meter, of TEM line length, or conductance G, Siemens/m, of the medium separating the two conductors. As discussed in Section 8.3.1, the time average power P_d dissipated per meter of length is simply the sum of the two contributions from the series and parallel conductances:

$$P_d(z)[W/m] = |\underline{I}(z)|^2 R/2 + |\underline{V}(z)|^2 G/2 \tag{7.1.61}$$

When R and G are unknown, resistive losses in transmission lines can be estimated by integrating $|\overline{J}|^2/2\sigma$ [W m^{-3}] over the volume of interest, where σ is the material conductivity [S m^{-1}] and $\overline{J}$ is the current density [A m^{-2}]. This surface loss density P_d [W m^{-2}] is derived for good conductors in Section 9.2 and is shown in (9.2.61) to be equal to the power dissipated by

the same surface current $\underline{J}_s$ flowing uniformly through a slab of thickness δ, where $\delta = (2/\omega\mu\sigma)^{0.5}$ is the skin depth. The surface current $\bar{J}_s$ equals $|\underline{\bar{H}}_s|$, which is the magnetic field parallel to the conductor surface. Therefore:

$$P_d \cong |\underline{\bar{H}}_s|^2 \sqrt{\frac{\omega\mu}{8\sigma}} \; \left[\mathrm{W/m^2}\right] \qquad \text{(power dissipation in conductors)} \quad (7.1.62)$$

For example, it is easy to compute with (7.1.62) the power dissipated in a 50-ohm copper TEM coaxial cable carrying $P_o = 10$ watts of entertainment over a 500-MHz band with an inner conductor diameter of one millimeter. First we note that $|\underline{\bar{H}}_s| = \underline{I}/2\pi r$ [A/m] where $|\underline{I}^2|Z_o/2 = P_o = 10$, and $2r = 10^{-3}$ [m]. Therefore $|\underline{\bar{H}}_s| = (P_o/Z_o)^{0.5}/2\pi r$ [A/m] $\cong 142$. Also, since the diameter of the outer sheath is typically ~5 times that of the inner conductor, the surface current density there, J_s, is one fifth that for the inner conductor, and the power dissipation per meter length is also one fifth. Therefore the total power dissipated per meter, P_L, in both conductors is ~1.2 times that dissipated in the inner conductor alone. If we consider only the highest and most lossy frequency, and assume $\sigma = 5\times10^7$, then substituting $|\underline{\bar{H}}_s|$ into (7.1.62) and integrating over both conductors yields the power loss:

$$\begin{aligned} P_L &\cong 1.2 \times 2\pi r |\underline{\bar{H}}_s|^2 \left(\omega\mu_o/4\sigma\right)^{0.5} = 1.2 \times 2\pi r \left[\left(2P_o/Z_o\right)^{0.5}/2\pi r\right]^2 \left(\omega\mu_o/8\sigma\right)^{0.5} \\ &= 1.2 \times P_o \left(Z_o\pi r\right)^{-1} \left(\omega\mu_o/2\sigma\right)^{0.5} = 12\left(50\pi 10^{-3}\right)^{-1} \left(2\pi \times 5\times 10^8 \times 4\pi \times 10^{-7}/10^8\right)^{0.5} \\ &= 0.48 \; \text{watts / meter} \end{aligned} \qquad (7.1.63)$$

The loss L [dB m^{-1}] is proportional to the ratio of P_L [W m^{-1}] to P_o [W]:

$$L\left[\text{dB m}^{-1}\right] = 4.34 \; P_L/P_o \qquad (7.1.64)$$

Thus P_L is 0.48 watts/meter, a large fraction of the ten watts propagating on the line. This loss of 4.8 percent of the power per meter, including the outer conductor, corresponds to $10\log_{10}(1 - 0.048) \cong -0.21$dB per meter. If we would like amplifiers along a cable to provide no more than ~50 dB gain, we need amplifiers every ~234 meters. Dropping the top frequency to 100 MHz, or increasing the diameter of the central wire could reduce these losses by perhaps a factor of ~4. These loss issues and desires for broad bandwidth are motivating substitution of low-loss optical fiber over long cable lines, and use of co-axial cables only for short hops from a local fiber to the home or business.

Example 7.1C
A perfectly conducting 50-ohm coaxial cable is filled with slightly conducting dielectric that gives the line a shunt conductivity $G = 10^{-6}$ Siemens m^{-1} between the two conductors. What is the attenuation of this cable (dB m^{-1})?

Solution: The attenuation L[dB m^{-1}] = 4.34 P$_d$/P$_o$ (7.1.64), where the power on the line P$_o$ [W] = |$\underline{V}$|^2/2Z$_o$, and the dissipation here is P$_d$ [W m^{-1}] = |$\underline{V}$|^{2}G/2 (7.1.61); see Figure 8.3.1 for the incremental model of a lossy TEM transmission line. Therefore L = 4.34 GZ$_o$ = 2.2×10^{-4} dB m^{-1}. This is generally independent of frequency and therefore might dominate at lower frequencies if the frequency-dependent dissipative losses in the wires become sufficiently small.

7.2 *TEM lines with junctions*

7.2.1 Boundary value problems

A junction between two transmission lines forces the fields in the first line to conform to the fields at the second line at the boundary between the two. This is a simple example of a broad class of problems called boundary value problems. The general electromagnetic *boundary value problem* involves determining exactly which, if any, combination of waves matches any given set of *boundary conditions*, which generally includes both active and passive boundaries, the active boundaries usually being sources. Boundary conditions generally constrain $\overline{E}$ and/or $\overline{H}$ for all time on the boundary of the one-, two- or three-dimensional region of interest.

The uniqueness theorem presented in Section 2.8 states that only one solution satisfies all Maxwell's equations if the boundary conditions are sufficient. Therefore we may solve boundary value problems simply by hypothesizing the correct combination of waves and testing it against Maxwell's equations. That is, we leave undetermined the numerical constants that characterize the chosen combination of waves, and then determine which values of those constraints satisfy Maxwell's equations. This strategy eases the challenge of hypothesizing the final answer directly. Moreover, symmetry and other considerations often suggest the nature of the wave combination required by the problem, thus reducing the numbers of unknown constants that must be determined.

The four basic steps for solving boundary value problems are:

1) Determine the natural behavior of each homogeneous section of the system without the boundaries.

2) Express this general behavior as the superposition of waves or static fields characterized by unknown constants; symmetry and other considerations can minimize the number of waves required. Here our basic building blocks are TEM waves.

3) Write equations for the boundary conditions that must be satisfied by these sets of superimposed waves, and then solve for the unknown constants.

4) Test the resulting solution against any of Maxwell's equations that have not already been imposed.

Variations of this four-step procedure can be used to solve almost any problem by replacing Maxwell's equations with their approximate equivalent for the given problem domain[34]. For example, profitability, available capital, technological constraints, employee capabilities, and customer needs are often "boundary conditions" when deriving strategies for start-up enterprises, while "natural behavior" could include the probable family of behaviors of the entrepreneurial team and its customers, financiers, and suppliers.

7.2.2 Waves at TEM junctions in the time domain

The boundary value problem approach described in Section 7.2.1 can be used for waves at TEM junctions. We assume that an arbitrary incident wave will produce both reflected and transmitted waves. For this introductory problem we also assume that no waves are incident from the other direction, for their solution could be superimposed later. Section 7.2.3 treats the same problem in the complex domain. We represent TEM lines graphically by parallel lines and their characteristic impedance Z_o, as illustrated in Figure 7.2.1 for lines a and b.
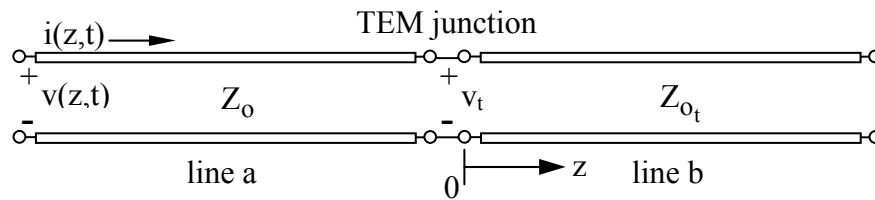

Figure 7.2.1   Junction of two TEM transmission lines.

Step one of the boundary value method involves characterizing the natural behavior of waves in the two media of interest, lines a and b. This follows from (7.1.16) for $v(z,t)$ and (7.1.18) for $i(z,t)$. Step two involves hypothesizing the form of the reflected and transmitted waves, $v_-(z,t)$ and $v_t(z,t)$. For simplicity we assume the source $v_+(z,t)$ is on the left, the TEM junction is at $z = 0$, and the line impedances Z_o are constants independent of time and frequency. Step three is to write the boundary conditions for the waves with unknown constants; v and i must both be constant across the junction at $z = 0$:

$$v(z,t) = v_+(z,t) + v_-(z,t) = v_t(z,t) \qquad \text{(at } z = 0) \qquad (7.2.1)$$

$$i(z,t) = Z_o^{-1}\left[v_+(z,t) - v_-(z,t)\right] = Z_t^{-1} v_t(t) \qquad \text{(at } z = 0) \qquad (7.2.2)$$

Step four involves solving (7.2.1) and (7.2.2) for the unknown waves $v_-(z,t)$ and $v_t(z,t)$. We can simplify the problem by taking the ratios of reflection and transmission relative to the incident wave and provide its amplitude later. If we regard the arguments (z=0, t) as understood, then (7.2.1) and (7.2.2) become:

---

[34] A key benefit of a technical education involves learning precise ways of thinking and solving problems; this procedure, when generalized, is an excellent example applicable to almost any career.

$$1 + \left(v_- / v_+ \right) = v_t / v_+ \tag{7.2.3}$$

$$1 - \left(v_- / v_+ \right) = \left(Z_o / Z_t \right) v_t / v_+ \tag{7.2.4}$$

To make the algebra for these two equations still more transparent it is customary to define v_-/v_+ as the *reflection coefficient* Γ, v_t/v_+ as the *transmission coefficient* T, and $Z_t/Z_o = Z_n$ as the *normalized impedance* for line b. Note that v_-, v_+, Z_o, and Z_t are real, and the fraction of incident power that is reflected from a junction is $|\Gamma|^2$. Equations (7.2.3) and (7.2.4) then become:

$$1 + \Gamma = T \tag{7.2.5}$$

$$1 - \Gamma = T/Z_n \tag{7.2.6}$$

Multiplying (7.2.6) by Z_n and subtracting the result from (7.2.5) eliminates T and yields:

$$\Gamma = \frac{v_-}{v_+} = \frac{Z_n - 1}{Z_n + 1} \tag{7.2.7}$$

$$v_- \left(0, t \right) = \left[\left(Z_n - 1 \right) / \left(Z_n + 1 \right) \right] v_+ \left(0, t \right) \tag{7.2.8}$$

$$v_- \left(0 + ct \right) = \left[\left(Z_n - 1 \right) / \left(Z_n + 1 \right) \right] v_+ \left(0 + ct \right) \tag{7.2.9}$$

$$v_- \left(z + ct \right) = \left[\left(Z_n - 1 \right) / \left(Z_n + 1 \right) \right] v_+ \left(z + ct \right) \tag{7.2.10}$$

The transitions to (7.2.9) and (7.2.10) utilized the fact that if two functions of two arguments are equal for all values of their arguments, then the functions remain equal as their arguments undergo the same numerical shifts. For example, if $X(a) = Y(b)$ where a and b have the same units, then $X(a + c) = Y(b + c)$. Combining (7.2.3) and (7.2.7) yields the transmitted voltage v_t in terms of the source voltage v_+:

$$v_t \left(z - ct \right) = \left[2 Z_n / \left(Z_n + 1 \right) \right] v_+ \left(z - ct \right) \tag{7.2.11}$$

This completes the solution for signal behavior at single TEM junctions.

Example 7.2A
Two parallel plates of width W and separation $d_1 = 1$ cm are connected at $z = D$ to a similar pair of plates spaced only $d_2 = 2$ mm apart. If the forward wave on the first line is $V_o \cos(\omega t - kz)$, what voltage $v_t(t,z)$ is transmitted beyond the junction at $z = D$?

Solution: $v_t(t,z) = Tv_+(t,z) = (1 + \Gamma)v_+(t,z) = 2Z_n\ v_+(t,z)/(Z_n + 1)$, where $Z_n = Z_t/Z_o = \eta_o d_2 W/\eta_o d_1 W = d_2/d_1 = 0.2$. Therefore for $z > D$, $v_t(t,z) = v_+(t,z)2\times0.2/(0.2 + 1) = (V_o/3)\cos(\omega t - kz)$ [V].

7.2.3   Sinusoidal waves on TEM transmission lines and at junctions

The basic equations characterizing lossless TEM lines in the sinusoidal steady state correspond to the pair of differential equations (7.1.25) and (7.1.26):

$$d\underline{V}(z)/dz = -j\omega L\underline{I}(z) \tag{7.2.12}$$

$$d\underline{I}(z)/dz = -j\omega C\underline{V}(\ ) \tag{7.2.13}$$

L and C are the inductance and capacitance of the line per meter, respectively.

This pair of equations leads easily to the *transmission line wave equation*:

$$d^2\underline{V}(z)/dz^2 = -\omega^2 LC\underline{V}(z) \qquad\text{(wave equation)}\quad (7.2.14)$$

The solution $\underline{V}(z)$ to this wave equation involves exponentials in z because the second derivative of $\underline{V}(z)$ equals a constant times $\underline{V}(z)$. The exponents can be + or -, so in general a sum of these two alternatives is possible, where $\underline{V}_+$ and $\underline{V}_-$ are complex constants determined later by boundary conditions and k is given by (7.1.30):

$$\underline{V}(z) = \underline{V}_+e^{-jkz} + \underline{V}_-e^{+jkz} \text{ [V]} \qquad\text{(TEM voltage)}\quad (7.2.15)$$

The corresponding current is readily found using (7.2.12):

$$\underline{I}(z) = (j/\omega L)d\underline{V}(z)/dz = (j/\omega L)\left(-jk\underline{V}_+e^{-jkz} + jk\underline{V}_-e^{+jkz}\right) \tag{7.2.16}$$

$$\underline{I}(z) = (1/Z_o)\left(\underline{V}_+{}^{-jkz} - \underline{V}_-e^{+jkz}\right) \qquad\text{(TEM current)}\quad (7.2.17)$$

where the *characteristic impedance* Z_o of the line is:

$$Z_o = Y_o^{-1} = \omega L/k = cL = (L/C)^{0.5} \text{ [ohms]} \quad\text{(characteristic impedance)}\quad (7.2.18)$$

The *characteristic admittance* Y_o of the line is the reciprocal of Z_o, and has units of Siemens or ohms[-1]. It is important to appreciate the physical significance of Z_o; it is simply the ratio of voltage to current for a wave propagating in one direction only on the line, e.g., for the + wave only. This ratio does not correspond to dissipative losses in the line, although it is related to the power traveling down the line for any given voltage across the line.
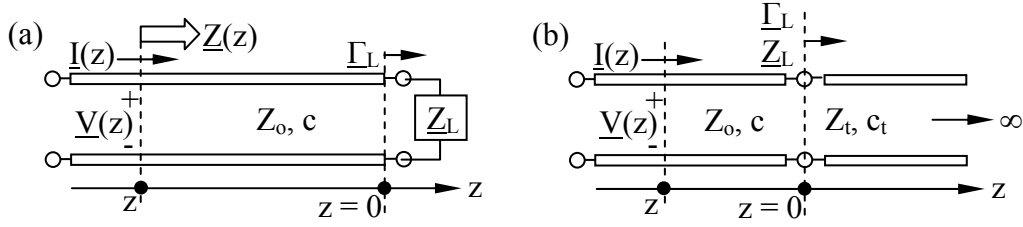
Figure 7.2.2   TEM transmission line impedances and coupling.

When there are both forward and backward waves on a line, the voltage/current ratio is called the complex impedance and varies with position, as suggested in Figure 7.2.2(a).   The *impedance* at any point along the line is defined as:

$$\underline{Z}(z) \equiv \underline{V}(z)/\underline{I}(z) = Z_0 \left[1+\underline{\Gamma}(z)\right]/\left[1-\underline{\Gamma}(z)\right]$$
$$= Z_0 \left[1+\underline{\Gamma}(z)\right]/\left[1-\underline{\Gamma}(z)\right] \text{ ohms}$$

(line impedance)   (7.2.19)

The complex *reflection coefficient* $\underline{\Gamma}(z)$ is defined as:

$$\underline{\Gamma}(z) \equiv \underline{V}_- e^{+jkz}/\underline{V}_+ e^{-jkz} = (\underline{V}_-/\underline{V}_+)e^{2jkz} = \underline{\Gamma}_L e^{2jkz}$$

(reflection coefficient)   (7.2.20)

When z = 0 at the load, then $\underline{V}_-/\underline{V}_+$ is defined at the load and $\underline{\Gamma}_L$ is the load reflection coefficient, denoted by the subscript L.

Equation (7.2.20) leads to a simple algorithm for relating impedances at different points along the line.   We first define normalized impedance $\underline{Z}_n$ and relate it to the reflection coefficient $\underline{\Gamma}(z)$ using (7.2.19); (7.2.22) follows from (7.2.21):

$$\underline{Z}_n(z) \equiv \frac{\underline{Z}(z)}{Z_0} = \frac{1+\underline{\Gamma}(z)}{1-\underline{\Gamma}(z)}$$

(normalized impedance)   (7.2.21)

$$\underline{\Gamma}(z) = \frac{\underline{Z}_n(z)-1}{\underline{Z}_n(z)+1}$$

(7.2.22)

For example, we can see the effect of the load impedance $\underline{Z}_L$ (z = 0) at some other point z on the line by using (7.2.20–22) in an appropriate sequence:

$$\underline{Z}_L \to \underline{Z}_{Ln} \to \underline{\Gamma}_L \to \underline{\Gamma}(z) \to \underline{Z}_n(z) \to \underline{Z}(z)$$ (impedance transformation) (7.2.23)

A simple example of the use of (7.2.23) is the transformation of a 50-ohm resistor by a 100-ohm line $\lambda/4$ long.   Using (7.2.23) in sequence, we see $\underline{Z}_L = 50$, $\underline{Z}_{Ln} = 50/100 = 0.5$, $\underline{\Gamma}_L = -1/3$ from (7.2.22), $\underline{\Gamma}(z = -\lambda/4) = +1/3$ from (7.2.20) where $e^{+2jkz} = e^{2j(2\pi/\lambda)(-\lambda/4)} = e^{-j\pi} = -1$, $\underline{Z}_n(-\lambda/4) = 2$ from (7.2.21), and therefore $\underline{Z}(-\lambda/4) = 200$ ohms.

Two other impedance transformation techniques are often used instead: a direct equation and the Smith chart (Section 7.3). The direct equation (7.2.24) can be derived by first substituting $\underline{\Gamma}_L = (\underline{Z}_L - Z_o)/(\underline{Z}_L + Z_o)$, i.e. (7.2.22), into $\underline{Z}(z) = \underline{V}(z)/\underline{I}(z)$, where $\underline{V}(z)$ and $\underline{I}(z)$ are given by (7.2.15) and (7.2.17), respectively, and $\underline{V}_-/\underline{V}_+ = \underline{\Gamma}_L$. The next step involves grouping the exponentials to yield sin kz and cos kz, and then dividing sin by cos to yield tan and the solution:

$$\underline{Z}(z) = Z_o \frac{\underline{Z}_L - jZ_o \tan kz}{Z_o - j\underline{Z}_L \tan kz} \qquad \text{(transformation equation)} \qquad (7.2.24)$$

A closely related problem is illustrated in Figure 7.2.2(b) where two transmission lines are connected together and the right-hand line presents the impedance $\underline{Z}_t$ at z = 0. To illustrate the general method for solving boundary value problems outlined in Section 7.2.1, we shall use it to compute the reflection and transmission coefficients at this junction. The expressions (7.2.15) and (7.2.17) nearly satisfy the first two steps of that method, which involve writing trial solutions composed of superimposed waves with unknown coefficients that satisfy the wave equation within each region of interest. The third step is to write equations for these waves that satisfy the boundary conditions, and then to solve for the unknown coefficients. Here the boundary conditions are that both $\underline{V}$ and $\underline{I}$ are continuous across the junction at z = 0; the subscript t corresponds to the transmitted wave. The two waves on the left-hand side have amplitudes $\underline{V}_+$ and $\underline{V}_-$, whereas the wave on the right-hand side has amplitude $\underline{V}_t$. We assume no energy enters from the right. Therefore:

$$\underline{V}(0) = \underline{V}_+ + \underline{V}_- = \underline{V}_t \qquad (7.2.25)$$

$$\underline{I}(0) = \left(\underline{V}_+ - \underline{V}_-\right)/Z_o = \underline{V}_t/\underline{Z}_t \qquad (7.2.26)$$

We define the complex reflection and transmission coefficients at the junction (z = 0) to be $\underline{\Gamma}$ and $\underline{T}$, respectively, where:

$$\underline{\Gamma} = \underline{V}_-/\underline{V}_+ \qquad \text{(complex reflection coefficient)} \qquad (7.2.27)$$

$$\underline{T} = \underline{V}_t/\underline{V}_+ \qquad \text{(complex transmission coefficient)} \qquad (7.2.28)$$

We may solve for $\underline{\Gamma}$ and $\underline{T}$ by first dividing (7.2.25) and (7.2.26) by $\underline{V}_+$:

$$1 + \underline{\Gamma} \quad \underline{T} \qquad (7.2.29)$$

$$1 - \underline{\Gamma} = \left(Z_o/\underline{Z}_t\right)\underline{T} \qquad (7.2.30)$$

This pair of equations is readily solved for $\underline{\Gamma}$ and $\underline{T}$:

$$\underline{\Gamma} = \frac{\underline{Z}_t - Z_o}{\underline{Z}_t + Z_o} = \frac{\underline{Z}_n - 1}{\underline{Z}_n + 1} \qquad\qquad (7.2.31)$$

$$\underline{T} = \underline{\Gamma} + 1 = \frac{2\underline{Z}_n}{\underline{Z}_n + 1} \qquad\qquad (7.2.32)$$

where normalized impedance was defined in (7.2.21) as $\underline{Z}_n \equiv \underline{Z}_t/Z_o$. For example, (7.2.31) says that the reflection coefficient $\underline{\Gamma}$ is zero when the normalized impedance is unity and the line impedance is matched, so $\underline{Z}_t = Z_o$; (7.2.32) then yields $\underline{T} = 1$.

The complex coefficients $\underline{\Gamma}$ and $\underline{T}$ refer to wave amplitudes, but often it is power that is of interest. In general the time-average power incident upon the junction is:

$$P_+ = \underline{V}_+ \underline{I}_+^{\,*}/2 = |\underline{V}_+|^2/2Z_o \ [\text{W}] \qquad\qquad \text{(incident power)} \quad (7.2.33)$$

Similarly the reflected and transmitted powers are P_- and P_t, where $P_- = |\underline{V}_-|^2/2Z_o$ and $P_t = |\underline{V}_t|^2/2Z_t \ [\text{W}]$.

Another consequence of having both forward and backward moving waves on a TEM line is that the magnitudes of the voltage and current vary along the length of the line. The expression for voltage given in (7.2.15) can be rearranged as:

$$|\underline{V}(z)| = |\underline{V}_+ e^{-jkz} + \underline{V}_- e^{+jkz}| = |\underline{V}_+ e^{-jkz}||1 + \underline{\Gamma}(z)| \qquad\qquad (7.2.34)$$

The magnitude of $|\underline{V}_+ e^{-jkz}|$ is independent of z, so the factor $|1 + \underline{\Gamma}(z)|$ controls the magnitude of voltage on the line, where $\underline{\Gamma}(z) = \underline{\Gamma}_L e^{2jkz}$ (7.2.20). Figure 7.2.3(a) illustrates the behavior of $|V(z)|$; it is quasi-sinusoidal with period $\lambda/2$ because of the 2jkz in the exponent. The maximum value $|V(z)|_{max} = |V_+| + |V_-|$ occurs when $\underline{\Gamma}(z) = |\underline{\Gamma}|$; the minimum occurs when $\underline{\Gamma}(z) = -|\underline{\Gamma}|$.
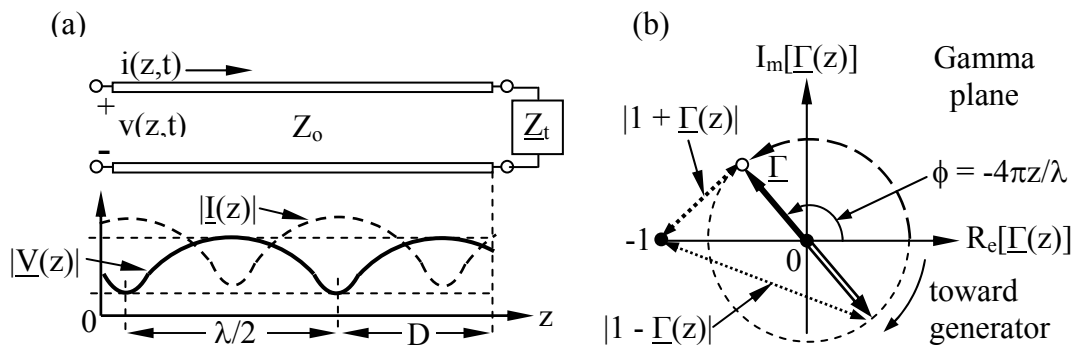

Figure 7.2.3   Standing waves on a TEM line and the Gamma plane.

The origins of this behavior of $|\underline{V}(z)|$ is suggested in Figure 7.2.3(b), which illustrates the z dependence of $\underline{\Gamma}(z)$ in the complex *gamma plane*, where the horizontal and vertical axes are the real and imaginary parts of $\underline{\Gamma}(z)$, respectively. Increases in z simply rotate the vector $\underline{\Gamma}(z)$ clockwise, preserving its magnitude [see (7.2.20) and Figure 7.2.3(b)].

The quasi-sinusoidal form of $|\underline{V}(z)|$ arises because $|\underline{V}(z)| \propto |1 + \underline{\Gamma}(z)|$, which is the length of the vector linking $\underline{\Gamma}(z)$ with the point -1 on the gamma plane, as illustrated in Figure 7.2.3(b). As the phase ϕ of $\underline{\Gamma}$ varies with z and circles the diagram, the vector $1 + \underline{\Gamma}(z)$ varies as might an arm turning a crank, and so it is sometimes called the "crank diagram". When $|\underline{\Gamma}| \ll 1$ then $|\underline{V}(z)|$ resembles a weak sinusoid oscillating about a mean value of $|\underline{V}_+|$, whereas when $|\underline{\Gamma}| \cong 1$ then $|\underline{V}(z)|$ resembles a fully rectified sinusoid. The voltage envelope $|\underline{V}(z)|$ is called the standing-wave pattern, and fields have a standing-wave component when $|\underline{\Gamma}| > 0$. The figure also illustrates how $|\underline{I}(z)| \propto |1 - \underline{\Gamma}(z)|$ exhibits the same quasi-sinusoidal variation as $|\underline{V}(z)|$, but 180 degrees out of phase.

Because $|V(z)|$ and $|I(z)|$ are generally easy to measure along any transmission line, it is useful to note that such measurements can be used to determine not only the fraction of power that has been reflected from any load, and thus the efficiency of any connection, but also the impedance of the load itself. First we define the *voltage standing wave ratio* or *VSWR* as:

$$\text{VSWR} \equiv |\underline{V}(z)|_{max} / |\underline{V}(z)|_{min} = (|\underline{V}_+| + |\underline{V}_-|)/(|\underline{V}_+| - |\underline{V}_-|) = (1 + |\underline{\Gamma}|)/(1 - |\underline{\Gamma}|) \qquad (7.2.35)$$

Therefore:

$$|\underline{\Gamma}| = (\text{VSWR} - 1)/(\text{VSWR} + 1) \qquad (7.2.36)$$

$$P_-/P_+ = |\underline{\Gamma}|^2 = [(\text{VSWR} - 1)/(\text{VSWR} + 1)]^2 \qquad (7.2.37)$$

This simple relation between VSWR and fractional power reflected (P_-/P_+) helped make VSWR a common specification for electronic equipment.

To find the load impedance $\underline{Z}_L$ from observations of $|\underline{V}(z)|$ such as those plotted in Figure 7.2.3(a) we first associate any voltage minimum with that point on the gamma plane that corresponds to $-|\underline{\Gamma}|$. Then we can rotate on the gamma plane counter-clockwise (toward the load) an angle $\phi = 2kD = 4\pi D/\lambda$ radians that corresponds to the distance D between that voltage minimum and the load, where a full revolution in the gamma plane corresponds to $D = \lambda/2$. Once $\underline{\Gamma}$ for the load is determined, it follows from (7.2.21) that:

$$\underline{Z}_L = Z_0[1 + \underline{\Gamma}]/[1 - \underline{\Gamma}] \qquad (7.2.38)$$

If more than two TEM lines join a single junction then their separate impedances combine in series or parallel, as suggested in Figure 7.2.4. The impedances add in parallel for Figure 7.2.4(a) so the impedance at the junction as seen from the left would be:

$$\underline{Z}_{parallel} = \underline{Z}_a \underline{Z}_b / (\underline{Z}_a + \underline{Z}_b) \qquad (7.2.39)$$

For Figure 7.2.4(b) the lines are connected in series so the impedance seen from the left would be $\underline{Z}_a + \underline{Z}_b$.
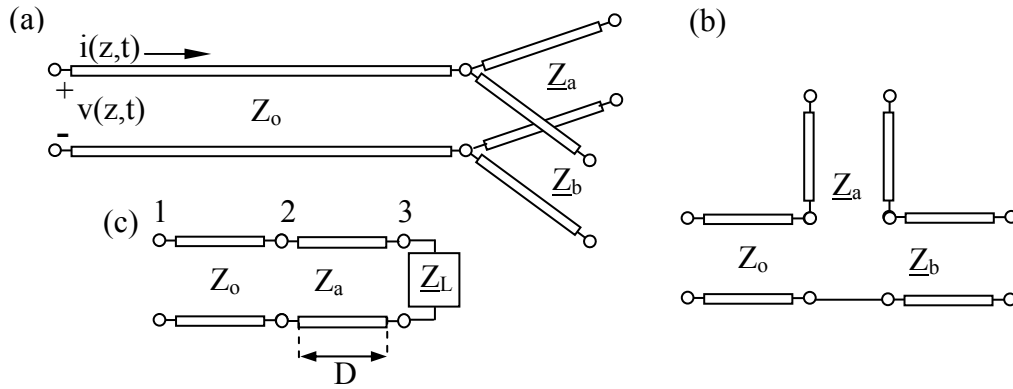

Figure 7.2.4   Multiple connected TEM lines.

Figure 7.2.4(c) illustrates how TEM lines can be concatenated. In this case the impedance $\underline{Z}_1$ seen at the left-hand terminals could be determined by transforming the impedance $\underline{Z}_L$ at terminals (3) to the impedance $\underline{Z}_2$ that would be seen at terminals (2). The impedance seen at (2) could then be transformed a second time to yield the impedance seen at the left-hand end. The algorithm for this might be:

$$\underline{Z}_L \to \underline{Z}_{Ln} \to \underline{\Gamma}_3 \to \underline{\Gamma}_2 \to \underline{Z}_{n2} \to \underline{Z}_2 \to \underline{Z}_{n2'} \to \underline{\Gamma}_{2'} \to \underline{\Gamma}_1 \to \underline{Z}_{n1} \to \underline{Z}_1 \qquad (7.2.40)$$

Note that Z_{n2} is normalized with respect to Z_a and Z_{n2}' is normalized with respect to Z_o; both are defined at junction (2). Also, $\underline{\Gamma}_2$ is the reflection coefficient at junction (2) within the line Z_a, and $\underline{\Gamma}_2$' is the reflection coefficient at junction (2) within the line Z_o.

Example 7.2B
A 100-ohm air-filled TEM line is terminated at $z = 0$ with a capacitor $C = 10^{-11}$ farads. What is $\underline{\Gamma}(z)$? At what positions $z < 0$ are voltage minima located on the line when $f = 1/2\pi$ GHz? What is the VSWR? At $z = -\lambda/4$, what is the equivalent impedance?

Solution: The normalized load impedance $Z_L/Z_o \equiv Z_{Ln} = 1/j\omega C Z_o = -j/(10^9 \times 10^{-11} \times 100) = -j$, and (7.2.22) gives $\underline{\Gamma}_L = (\underline{Z}_{Ln} - 1)/(\underline{Z}_{Ln} + 1) = -(1+j)/(1-j) = -j$. $\underline{\Gamma}(z) = \underline{\Gamma}_L e^{2jkz} = -je^{2jkz}$. (7.2.34) gives $|\underline{V}(z)| \propto |1 + \underline{\Gamma}(z)| = |1 - je^{2jkz}| = 0$ when $e^{2jkz} = -j = e^{-j(\pi/2 + n2\pi)}$, where n $= 0,1,2\ldots$ Therefore $2jkz = -j(\pi/2 + n2\pi)$, so $z(nulls) = -(\pi/2 + n2\pi)\lambda/4\pi = -(\lambda/8)(1 + 4n)$. But $f = 10^9/2\pi$, and so $\lambda = c/f = 2\pi c \times 10^{-9} = 0.6\pi$ [m]. (7.2.34) gives VSWR $= (1 + |\underline{\Gamma}|)/(1 - |\underline{\Gamma}|) = \infty$. At $z = -\lambda/4$, $\underline{\Gamma} \to -\underline{\Gamma}_L = +j$ via (7.2.20), so by (7.2.38) $\underline{Z} = Z_o[1 + \underline{\Gamma}]/[1 - \underline{\Gamma}] = 100[1 + j]/[1 - j] = j100 = j\omega L_o \Rightarrow L_o = 100/\omega = 100/10^9 = 10^{-7}$ [H].

Example 7.2C

The VSWR observed on a 100-ohm air-filled TEM transmission line is 2. The voltage minimum is 15 cm from the load and the distance between minima is 30 cm. What is the frequency of the radiation? What is the impedance $\underline{Z}_L$ of the load?

Solution: The distance between minima is $\lambda/2$, so $\lambda = 60$ cm and $f = c/\lambda = 3\times10^8/0.6 = 500$ MHz. The load impedance is $\underline{Z}_L = Z_o[1 + \underline{\Gamma}_L]/[1 - \underline{\Gamma}_L]$ (7.2.38) where $|\underline{\Gamma}_L| = $ (VSWR - 1)/(VSWR + 1) = 1/3 from (5.2.83). $\underline{\Gamma}_L$ is rotated on the Smith chart 180 degrees counter-clockwise (toward the load) from the voltage minimum, corresponding to a quarter wavelength. The voltage minimum must lie on the negative real $\underline{\Gamma}$ axis, and therefore $\underline{\Gamma}_L$ lies on the positive real $\underline{\Gamma}$ axis. Therefore $\underline{\Gamma}_L = 1/3$ and $\underline{Z}_L = 100(1 + 1/3)/(1 - 1/3) = 200$ ohms.

7.3   *Methods for matching transmission lines*

7.3.1   Frequency-dependent behavior

This section focuses on the frequency-dependent behavior introduced by obstacles and impedance transitions in transmission lines, including TEM lines, waveguides, and optical systems. Frequency-dependent transmission line behavior can also be introduced by loss, as discussed in Section 8.3.1, and by the frequency-dependent propagation velocity of waveguides and optical fibers, as discussed in Sections 9.3 and 12.2.

The basic issue is illustrated in Figure 7.3.1(a), where an obstacle reflects some fraction of the incident power. If we wish to eliminate losses due to reflections we need to cancel the reflected wave by adding another that has the same magnitude but is 180° out of phase. This can easily be done by adding another obstacle in front of or behind the first with the necessary properties, as suggested in (b). However, the reflections from the further obstacle can bounce between the two obstacles multiple times, and the final result must consider these additional rays too. If the reflections are small the multiple reflections become negligible. This strategy works for any type of transmission line, including TEM lines, waveguides and optical systems.
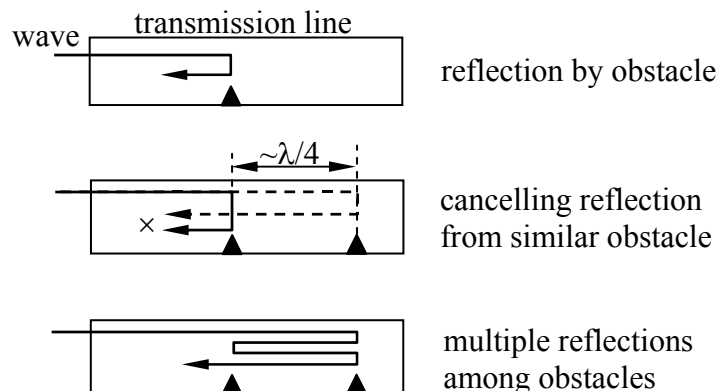

Figure 7.3.1   Cancellation of reflections on transmission lines.

The most important consequence of any such tuning strategy to eliminate reflections is that the two reflective sources are often offset spatially, so the relative phase between them is wavelength dependent. If multiple reflections are important, this frequency dependence can increase substantially. Rather than consider all these reflections in a tedious way, we can more directly solve the equations by extending the analysis of Section 7.2.3, which is summarized below in the context of TEM lines having characteristic admittance Y_o and a termination of complex impedance $\underline{Z}_L$:

$$\underline{V}(z) = \underline{V}_+ e^{-jkz} + \underline{V}_- e^{jkz} \ [V] \tag{7.3.1}$$

$$\underline{I}(z) = Y_o \left(\underline{V}_+ e^{-jkz} - \underline{V}_- e^{jkz} \right) [A] \tag{7.3.2}$$

$$\underline{\Gamma}(z) \equiv \left(\underline{V}_- e^{jkz} \right) / \left(\underline{V}_+ e^{-jkz} \right) = \left(\underline{V}_- / \underline{V}_+ \right) e^{2jkz} = \left(\underline{Z}_n - 1 \right) / \left(\underline{Z}_n + 1 \right) \tag{7.3.3}$$

The normalized impedance $\underline{Z}_n$ is defined as:

$$\underline{Z}_n \equiv \underline{Z}/Z_o = \left[1 + \underline{\Gamma}(z) \right] / \left[1 - \underline{\Gamma}(z) \right] \tag{7.3.4}$$

$\underline{Z}_n$ can be related to $\underline{\Gamma}(z)$ by dividing (7.3.1) by (7.3.2) to find $\underline{Z}(z)$, and the inverse relation (7.3.3) follows. Using (7.3.3) and (7.3.4) in the following sequences, the impedance $\underline{Z}(z_2)$ at any point on an unobstructed line can be related to the impedance at any other point z_1:

$$\underline{Z}(z_1) \Leftrightarrow \underline{Z}_n(z_1) \Leftrightarrow \underline{\Gamma}(z_1) \Leftrightarrow \underline{\Gamma}(z_2) \Leftrightarrow \underline{Z}_n(z_2) \Leftrightarrow \underline{Z}(z_2) \tag{7.3.5}$$

The five arrows in (7.3.5) correspond to application of equations (7.3.3) and (7.3.4) in the following left-to-right sequence: (4), (3), (3), (4), (4), respectively.

One standard problem involves determining $\underline{Z}(z)$ (for $z < 0$) resulting from a load impedance $\underline{Z}_L$ at $z = 0$. One approach is to replace the operations in (7.3.5) by a single equation, derived in (7.2.24):

$$\underline{Z}(z) = Z_o \left(\underline{Z}_L - jZ_o \tan kz \right) / \left(Z_o - j\underline{Z}_L \tan kz \right) \qquad (\textit{impedance transformation}) \tag{7.3.6}$$

For example, if $\underline{Z}_L = 0$, then $\underline{Z}(z) = -jZ_o \tan kz$, which means that $\underline{Z}(z)$ can range between $-j\infty$ and $+j\infty$, depending on z, mimicking any reactance at a single frequency. The impedance repeats at distances of $\Delta z = \lambda$, where $k(\Delta z) = (2\pi/\lambda)\Delta z = 2\pi$. If $\underline{Z}_L = Z_o$, then $\underline{Z}(z) = Z_o$ everywhere.

Example 7.3A
What is the impedance at 100 MHz of a 100-ohm TEM line $\lambda/4$ long and connected to a: 1) short circuit? 2) open circuit? 3) 50-ohm resistor? 4) capacitor $C = 10^{-10}$ F?

Solution: In all four cases the relation between $\underline{\Gamma}(z = 0) = \underline{\Gamma}_L$ at the load, and $\underline{\Gamma}(z = -\lambda/4)$ is the same [see (7.3.3)]: $\underline{\Gamma}(z = -\lambda/4) = \underline{\Gamma}_L e^{2jkz} = \underline{\Gamma}_L e^{2j(2\pi/\lambda)(-\lambda/4)} = -\underline{\Gamma}_L$. Therefore in all four cases we see from (7.3.4) that $\underline{Z}_n(z = -\lambda/4) = (1 - \underline{\Gamma}_L)/(1 + \underline{\Gamma}_L) = 1/\underline{Z}_n(0)$. $\underline{Z}_n(z = 0)$ for these four cases is: 0, ∞, 0.5, and $1/j\omega CZ_o = 1/(j2\pi 10^8 10^{-10} 100) = 1/j2\pi$, respectively. Therefore $\underline{Z}(z = -\lambda/4) = 100\underline{Z}_n^{-1}$ ohms, which for these four cases equals: ∞, 0, 200, and $j200\pi$ ohms, respectively. Since the impedance of an inductor is $\underline{Z} = j\omega L$, it follows that $j200\pi$ is equivalent at 100 MHz to $L = 200\pi/\omega = 200\pi/200\pi 10^8 = 10^{-8}$ [Hy].

7.3.2   Smith chart, stub tuning, and quarter-wave transformers

A common problem is how to cancel reflections losslessly, thus forcing all incident power into a load. This requires addition of one or more reactive impedances in series or in parallel with the line so as to convert the impedance at that point to Z_o, where it must remain for all points closer to the source or next obstacle. Software tools to facilitate this have been developed, but a simple graphical tool, the *Smith chart*, provides useful insight into what can easily be matched and what cannot. Prior to computers it was widely used to design and characterize microwave systems.

The key operations in (7.3.5) are rotation on the *gamma plane* $[\underline{\Gamma}(z_1) \Leftrightarrow \underline{\Gamma}(z_2)]$ and the conversions $\underline{Z}_n \Leftrightarrow \underline{\Gamma}$, given by (7.3.3–4). Both of these operations can be accommodated on a single graph that maps the one-to-one relationship $\underline{Z}_n \Leftrightarrow \underline{\Gamma}$ on the complex gamma plane, as suggested in Figure 7.3.2(a). Conversions $\underline{\Gamma}(z_1) \Leftrightarrow \underline{\Gamma}(z_2)$ are simply rotations on the gamma plane. The gamma plane was introduced in Figure 7.2.3. The Smith chart simply overlays the equivalent normalized impedance values $\underline{Z}_n$ on the gamma plane; only a few key values are indicated in the simplified version shown in (a). For example, the loci for which the real R_n and imaginary parts X_n of $\underline{Z}_n$ remain constant lie on segments of circles $(\underline{Z}_n \equiv R_n + jX_n)$.
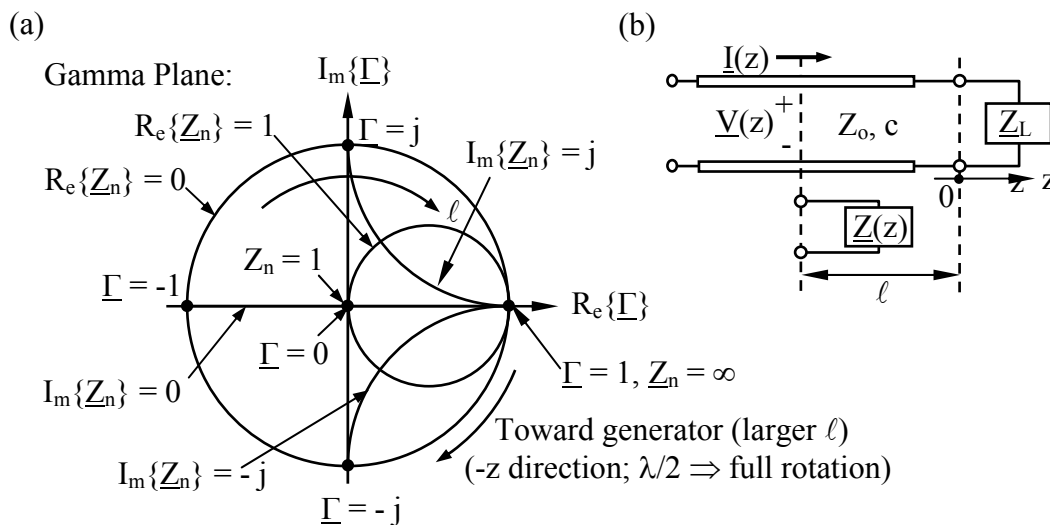

Figure 7.3.2   Relation between the gamma plane and the Smith chart.

Rotation on the gamma plane relates the values of $\underline{Z}_n$ and $\underline{\Gamma}$ at one z to their values at another, as suggested in Figure 7.3.2(b). Since $\underline{\Gamma}(z) = (\underline{V}_-/\underline{V}_+)e^{2jkz} = \underline{\Gamma}_L e^{2jkz} = \underline{\Gamma}_L e^{-2jk\ell}$, and since $e^{j\phi}$ corresponds to counter-clockwise rotation as ϕ increases, movement toward the generator (-z direction) corresponds to clockwise rotation in the gamma plane. The exponent of $e^{-2jk\ell}$ is $-j4\pi\ell/\lambda$, so a full rotation on the gamma plane corresponds to movement ℓ down the line of only $\lambda/2$.

A simple example illustrates the use of the Smith chart. Consider an inductor having $j\omega L = j100$ on a 100-ohm line. Then $\underline{Z}_n = j$, which corresponds to a point at the top of the Smith chart where $\underline{\Gamma} = +j$ (normally $\underline{Z}_n \neq \underline{\Gamma}$). If we move toward the generator $\lambda/4$, corresponding to rotation of $\underline{\Gamma}(z)$ half way round the Smith chart, then we arrive at the bottom where $\underline{Z}_n = -j$ and $\underline{Z} = Z_o\underline{Z}_n = -j100 = 1/j\omega C$. So the equivalent capacitance C at the new location is $1/100\omega$ farads.

The Smith chart has several other interesting properties. For example, rotation half way round the chart (changing $\underline{\Gamma}$ to $-\underline{\Gamma}$) converts any normalized impedance into the corresponding normalized admittance. This is easily proved: since $\underline{\Gamma} = (\underline{Z}_n - 1)/(\underline{Z}_n + 1)$, conversion of $\underline{Z}_n \rightarrow \underline{Z}_n^{-1}$ yields $\underline{\Gamma}' = (\underline{Z}_n^{-1} - 1)/(\underline{Z}_n^{-1} + 1) = (1 - Z_n)/(Z_n + 1) = -\underline{\Gamma}$ [Q.E.D.][35] Pairs of points with this property include $\underline{Z}_n = \pm j$ and $\underline{Z}_n = (0,\infty)$.

Another useful property of the Smith chart is that the voltage-standing-wave ratio (VSWR) equals the maximum positive real value $R_{n\,max}$ of $\underline{Z}_n$ lying on the circular locus occupied by $\underline{\Gamma}(z)$. This is easily shown from the definition of VSWR:

$$\text{VSWR} \equiv |\underline{V}_{max}|/|\underline{V}_{min}| = \left(\left|\underline{V}_+ e^{-jkz}\right| + \left|\underline{V}_- e^{+jkz}\right|\right)\Big/\left(\left|\underline{V}_+ e^{-jkz}\right| - \left|\underline{V}_- e^{+jkz}\right|\right)$$
$$\equiv \left(1 + |\underline{\Gamma}|\right)\big/\left(1 - |\underline{\Gamma}|\right) = R_{n\,max} \qquad (7.3.7)$$

A more important use of the Smith chart is illustrated in Figure 7.3.3, where the load $60 + j80$ is to be matched to a 100-ohm TEM line so all the power is dissipated in the 60-ohm resistor. In particular the length ℓ of the transmission line in Figure 7.3.3(a) is to be chosen so as to transform $\underline{Z}_L = 60 + 80j$ so that its real part becomes Z_o. The new imaginary part can be cancelled by a reactive load (L or C) that will be placed either in position M or N. The first step is to locate $\underline{Z}_n$ on the Smith chart at the intersection of the $R_n = 0.6$ and $X_n = 0.8$ circles, which happen to fall at $\underline{\Gamma} = 0.5j$. Next we locate the gamma circle $\underline{\Gamma}(z)$ along which we can move by varying ℓ. This intersects the $R_n = 1$ circle at point "a" after rotating toward the generator "distance A". Next we can add a negative reactance to cancel the reactance $jX_n = +1.18j$ at point "a" to yield $\underline{Z}_n(a) = 1$ and $\underline{Z} = Z_o$. A negative reactance is a capacitor C in series at location M in the circuit. Therefore $1/j\omega C = -1.18jZ_o$ and $C = (1.18\omega Z_o)^{-1}$. The required line length ℓ corresponds to $\sim 0.05\lambda$, a scale for which is printed on the perimeter of official charts as illustrated in Figure 7.3.4.

---

[35] Q.E.D. is an abbreviation for the Latin phrase "quod erat demonstratum", or "that which was to be demonstrated".
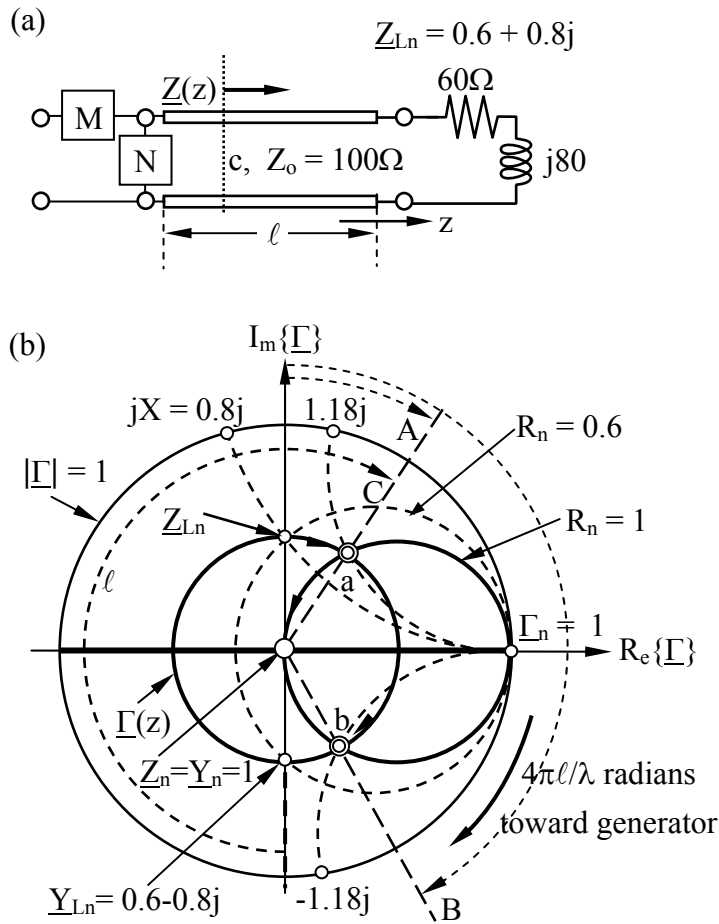
Figure 7.3.3   Matching a reactive load using the Smith chart.

More than three other matching schemes can be used here. For example, we could lengthen ℓ to distance "B" and point "b", where a positive reactance of $X_n = 1.18$ could be added in series at position M to provide a match. This requires an inductor $L = 1.18 Z_o/\omega$.

Alternatively, we could note that $\underline{Z}_{Ln}$ corresponds to $\underline{Y}_{Ln} = 0.6 - 0.8j$ on the opposite side of the chart ($\underline{\Gamma} \rightarrow -\underline{\Gamma}$), where the fact that both $\underline{Z}_{Ln}$ and $\underline{Y}_{Ln}$ have the same real parts is a coincidence limited to cases where $\underline{\Gamma}_L$ is pure imaginary. Rotating toward the generator distance C again puts us on the $G_n = 1$ circle ($\underline{Y}_n \equiv G_n + jB_n$), so we can add a negative admittance B_n of $-1.18j$ to yield Y_o. Adding a negative admittance in parallel at $z = -\ell$ corresponds to adding an inductor L in position N, where $-jZ_oX_n = 1/j\omega L$, so $L = (1.18\omega Z_o)^{-1}$. By rotating further to point "b" a capacitor could be added in parallel instead of the inductor. Generally one uses the shortest line length possible and the smallest, lowest-cost, lowest-loss reactive element.

IMPEDANCE OR ADMITTANCE COORDINATES
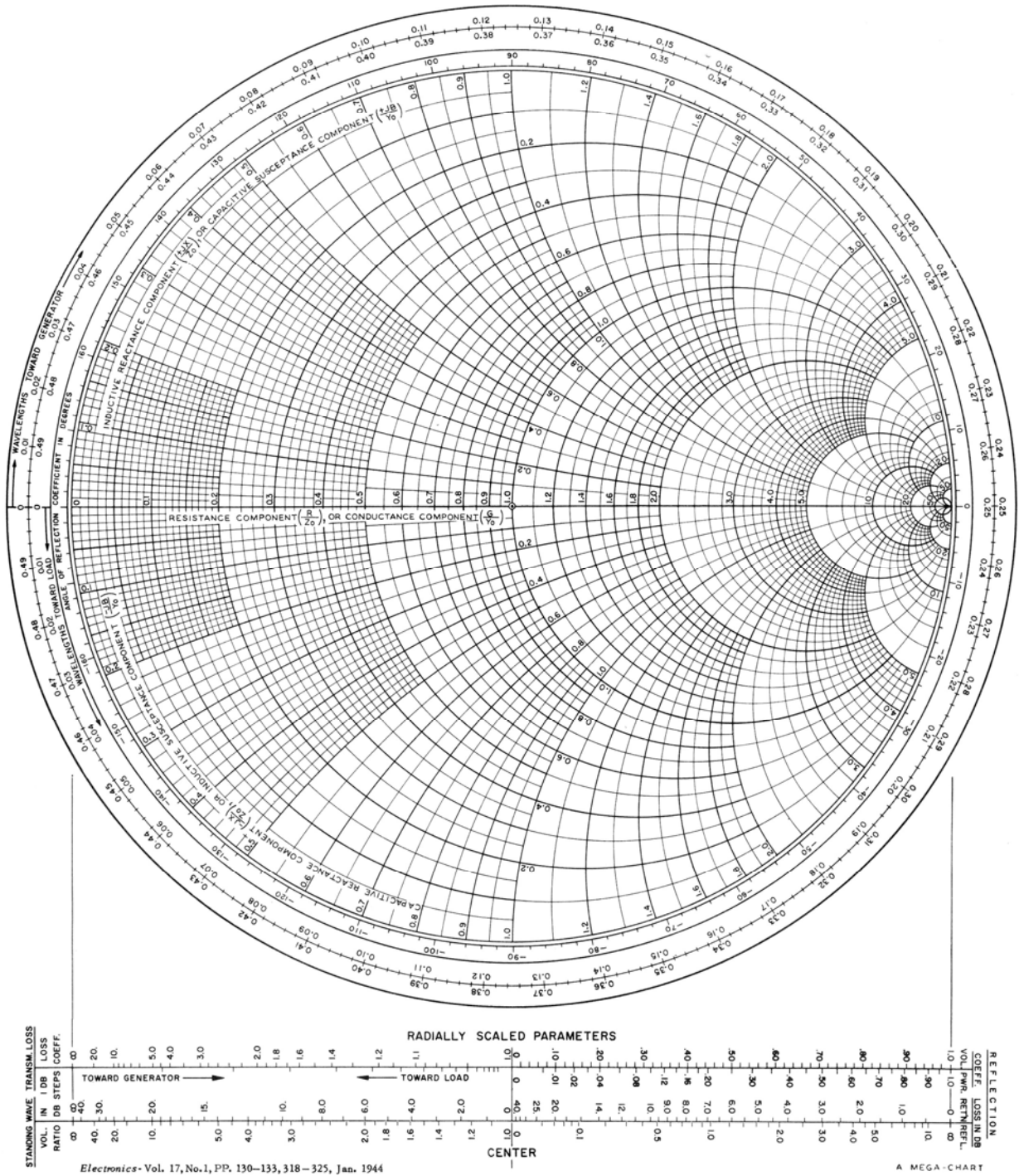
Figure 7.3.4   Smith chart.

Often printed circuits do not add capacitors or inductors to tune devices, but simply print an extra TEM line on the circuit board that is open- or short-circuit at its far end and is cut to a length that yields the desired equivalent L or C at the given frequency ω.

One useful approach to matching resistive loads is to insert a quarter-wavelength section of TEM line of impedance Z_A between the load Z_L and the feed line impedance Z_o. Then $Z_{Ln} = Z_L/Z_A$ and one quarter-wave-length down the TEM line where $\underline{\Gamma}$ becomes $-\underline{\Gamma}$, the normalized impedance becomes the reciprocal, $Z'_n = Z_A/Z_L$ and the total impedance there is $Z' = Z_A^2/Z_L$. If this matches the output transmission line impedance Z_o so that $Z_o = Z_A^2/Z_L$ then there are no reflections. The quarter-wavelength section is called a *quarter-wave transformer* and has the impedance $Z_A = (Z_L Z_o)^{0.5}$. A similar technique can be used if the load is partly reactive without the need for L's or C's, but the length and impedance of the transformer must be adjusted. For example, any line impedance Z_A will yield a normalized load impedance that can be rotated on a Smith chart to become a real impedance; if Z_A and the transformer length are chosen correctly, this real impedance will match Z_o. Matching usually requires iteration with a Smith chart or a numerical technique.

7.4   TEM resonances

7.4.1   Introduction

Resonators are widely used for manipulating signals and power, although unwanted resonances can sometimes limit system performance. For example, resonators can be used either as band-pass filters that remove all frequencies from a signal except those near the desired resonant frequency ω_n, or as band-stop filters that remove unwanted frequencies near ω_n and let all frequencies pass. They can also be used effectively as step-up transformers to increase voltages or currents to levels sufficient to couple all available energy into desired loads without reflections. That is, the matching circuits discussed in Section 7.3.2 can become sufficiently reactive for badly mismatched loads that they act like band-pass resonators that match the load only for a narrow band of frequencies. Although simple RLC resonators have but one natural resonance and complex RLC circuits have many, distributed electromagnetic systems can have an infinite number.

A *resonator* is any structure that can trap oscillatory electromagnetic energy so that it escapes slowly or not at all. Section 7.4.2 discusses energy trapped in TEM lines terminated so that outbound waves are reflected back into the resonator, and Section 9.4 treats cavity resonators formed by terminating rectangular waveguides with short circuits that similarly reflect and trap otherwise escaping waves. In each of these cases boundary conditions restricted the allowed wave structure inside to patterns having integral numbers of half- or quarter-wavelengths along any axis of propagation, and thus only certain discrete resonant frequencies ω_n can be present.

All resonators dissipate energy due to resistive losses, leakages, and radiation, as discussed in Section 7.4.3. The rate at which this occurs depends on where the peak currents or voltages in the resonator are located with respect to the resistive or radiating elements. For example, if the resistive element is in series at a current null or in parallel at a voltage null, there is no dissipation. Since dissipation is proportional to resonator energy content and to the squares of current or voltage, the decay of field strength and stored energy is generally exponential in time. Each resonant frequency f_n has its own rate of energy decay, characterized by the dimensionless

quality factor Q_n, which is generally the number of radians $\omega_n t$ required for the total energy w_{Tn} stored in mode n to decay by a factor of $1/e$. More importantly, $Q \cong f_o/\Delta f$, where f_n is the resonant frequency and Δf_n is the half-power full-width of resonance n.

Section 7.4.4 then discusses how resonators can be coupled to circuits for use as filters or transformers, and Section 7.4.5 discusses how arbitrary waveforms in resonators are simply a superposition of orthogonal modes, each decaying at its own rate.

7.4.2   TEM resonator frequencies

A *resonator* is any structure that traps electromagnetic radiation so it escapes slowly or not at all. Typical *TEM resonators* are terminated at their ends with lossless elements such as short- or open-circuits, inductors, or capacitors. Complex notation is used because resonators are strongly frequency-dependent. We begin with the expressions (7.1.55) and (7.1.58) for voltage and current on TEM lines:

$$\underline{V}(z) = \underline{V}_+ e^{-jkz} + \underline{V}_- e^{+jkz} \quad [V] \tag{7.4.1}$$

$$\underline{I}(z) = Y_o \left[\underline{V}_+ e^{-jkz} - \underline{V}_- e^{+jkz} \right] \quad [A] \tag{7.4.2}$$

For example, if both ends of a TEM line of length D are open-circuited, then $\underline{I}(z) = 0$ at $z = 0$ and $z = D$. Evaluating (7.4.1) at $z = 0$ yields $\underline{V}_- = \underline{V}_+$. At the other boundary:[36]

$$\underline{I}(D) = 0 = Y_o \underline{V}_+ \left(e^{-jkD} - e^{+jkD} \right) = -2jY_o \underline{V}_+ \sin(kD) = -2jY_o \underline{V}_+ \sin(2\pi D/\lambda) \tag{7.4.3}$$

To satisfy (7.4.3), $\sin(2\pi D/\lambda) = 0$, and so λ is restricted to specific resonances:

$$\lambda_n = 2D/n = c/f_n \text{ for } n = 0,1,2,3,... \tag{7.4.4}$$

That is, at resonance the length of this open-circuited line is $D = n\lambda_n/2$, as suggested in Figure 7.4.1(a) for $n = 1$. The corresponding resonant frequencies are:

$$f_n = c/\lambda_n = nc/2D \text{ [Hz]} \tag{7.4.5}$$

By our definition, static storage of electric or magnetic energy corresponds to a resonance at zero frequency. For example, in this case the line can hold a static charge and store electric energy at zero frequency ($n = 0$) because it is open-circuited at both ends. Because the different modes of a resonator are spatially orthogonal, the total energy stored in a resonator is the sum of the energies stored in each of the resonances separately, as shown later in (7.4.20).

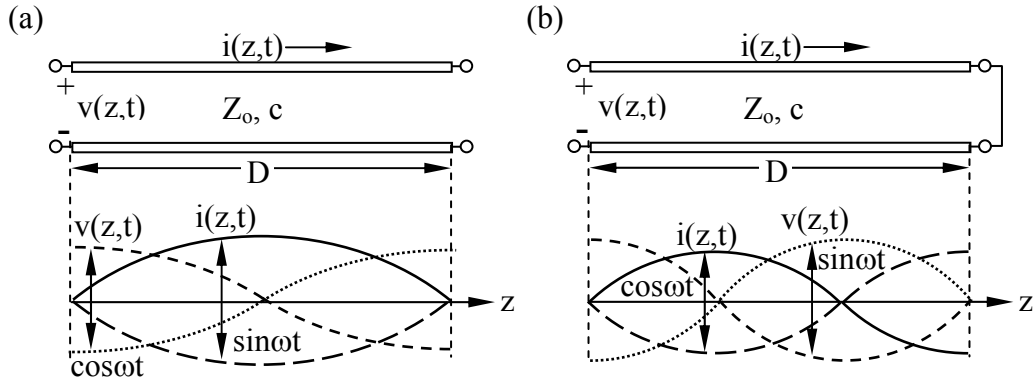[36] We use the identity $\sin\phi = (e^{j\phi} - e^{-j\phi})/2j$.

Figure 7.4.1 Voltage and current on TEM resonators.

The time behavior corresponding to (7.4.2) when $2Y_o\underline{V}_+ = I_o$ is:

$$i(t,z) = R_e\{\underline{I}e^{j\omega t}\} = I_o\sin\omega t\,\sin(2\pi z/\lambda) \tag{7.4.6}$$

where $\omega = 2\pi c/\lambda$. The corresponding voltage $v(t,z)$ follows from (7.4.1), $\underline{V}_- = \underline{V}_+$, and our choice that $2Y_o\underline{V}_+ = I_o$:

$$\underline{V}(z) = \underline{V}_+\left[e^{-jkz} + e^{+jkz}\right] = 2\underline{V}_+\cos(2\pi z/\lambda) \tag{7.4.7}$$

$$v(t,z) = Re\{\underline{V}e^{j\omega t}\} = Z_o I_o\cos\omega t\,\cos(2\pi z/\lambda) \tag{7.4.8}$$

Both $v(z,t)$ and $i(z,t)$ are sketched in Figure 7.4.1(a) for $n = 1$. The behavior of $i(t,z)$ resembles the motion of a piano string at resonance and is 90° out of phase with $v(z, t)$ in both space and time.

Figure 7.4.1(b) illustrates one possible distribution of voltage and current on a TEM resonator short-circuited at one end and open-circuited at the other. Since $i(t) = 0$ at the open circuit and $v(t) = 0$ at the short circuit, boundary conditions are satisfied by the illustrated $i(t,z)$ and $v(t,z)$. In this case:

$$D = (\lambda_n/4)(2n+1)\text{ for }n = 0,1,2,... \tag{7.4.9}$$

$$f_n = c/\lambda_n = c(2n+1)/4D\text{ [Hz]} \tag{7.4.10}$$

For $n = 0$ the zero-frequency solution for Figure 7.4.1(a) corresponds to the line being charged to a DC voltage V_o with zero current. The electric energy stored on the line is then $DCV_o^2/2$ [J], where the electric energy density on a TEM line (7.1.32) is:

$$W_e = C\left\langle v^2(t,z)\right\rangle/2 = C|\underline{V}|^2/4 \left[\text{J m}^{-1}\right]$$ (7.4.11)

The extra factor of 1/2 in the right-hand term of (7.4.11) results because $\left\langle\cos^2\omega t\right\rangle = 0.5$ for non-zero frequencies. A transmission line short-circuited at both ends also has a zero-frequency resonance corresponding to a steady current flowing around the line through the two short circuits at the ends, and the voltage across the line is zero everywhere.[37] The circuit of Figure 7.4.1(b) cannot store energy at zero frequency, however, and therefore has no zero-frequency resonance.

There is also a simple relation between the electric and magnetic energy storage in resonators because $Z_o = (L/C)^{0.5}$ (7.1.31). Using (7.4.11), (7.4.6), and (7.4.8) for $n > 0$:

$$\left\langle W_e\right\rangle = C\left\langle v^2(t,z)\right\rangle/2 = C\left\langle\left[Z_o I_o \cos\omega_n t \cos(2\pi z/\lambda_n)\right]^2\right\rangle/2$$ (7.4.12)

$$= \left(Z_o^2 I_o^2 C/4\right)\cos^2(2\pi z/\lambda_n) = \left(LI_o^2/4\right)\left[\cos^2(2\pi z/\lambda_n)\right]\left[\text{J m}^{-1}\right]$$ (7.4.13)

$$\left\langle W_m\right\rangle = L\left\langle i^2(t,z)\right\rangle/2 = L\left\langle\left[I_o \sin\omega_n t \sin(2\pi z/\lambda_n)\right]^2\right\rangle/2$$ (7.4.14)

$$= \left(LI_o^2/4\right)\sin^2(2\pi z/\lambda_n)\left[\text{J m}^{-1}\right]$$ (7.4.15)

Integrating these two time-average energy densities $\left\langle W_e\right\rangle$ and $\left\langle W_m\right\rangle$ over the length of a TEM resonator yields the important result that at any resonance the total time-average stored electric and magnetic energies w_e and w_m are equal; the fact that the lengths of all open- and/or short-circuited TEM resonators are integral multiples of a quarter wavelength λ_n is essential to this result. Energy conservation also requires this because periodically the current or voltage is everywhere zero together with the corresponding energy; the energy thus oscillates between magnetic and electric forms at twice f_n.

All resonators, not just TEM, exhibit equality between their time-average stored electric and magnetic energies. This can be proven by integrating Poynting's theorem (2.7.24) over the volume of any resonator for the case where the surface integral of $\overline{S}\bullet\hat{n}$ and the power dissipated P_d are zero:[38]

$$0.5\oiint_A \overline{S}\bullet\hat{n}\,da + \oiiint_V \left[\left\langle P_d(t)\right\rangle + 2j\omega(W_m - W_e)\right]dv = 0$$ (7.4.16)

$$\therefore w_m \equiv \oiiint_V W_m\,dv = \oiiint_V W_e\,dv = w_e \qquad \text{(energy balance at resonance)}$$ (7.4.17)

---

[37] Some workers prefer not to consider the zero-frequency case as a resonance; by our definition it is.
[38] We assume here that μ and ε are real quantities so W is real too.

This proof also applies, for example, to TEM resonators terminated by capacitors or inductors, in which case the reactive energy in the termination must be balanced by the line, which then is not an integral number of quarter wavelengths long.

Any system with spatially distributed energy storage exhibits multiple resonances. These resonance modes are generally orthogonal so the total stored energy is the sum of the separate energies for each mode, as shown below for TEM lines.

Consider first the open-ended TEM resonator of Figure 7.4.1(a), for which the voltage of the nth mode, following (7.4.7), might be:

$$\underline{V}_n(z) = \underline{V}_{no} \cos(n\pi z/D) \tag{7.4.18}$$

The total voltage is the sum of the voltages associated with each mode:

$$\underline{V}(z) = \sum_{n=0}^{\infty} \underline{V}(n) \tag{7.4.19}$$

The total electric energy on the TEM line is:

$$
\begin{aligned}
w_{eT} &= \int_0^D \left(C|\underline{V}(z)|^2 \Big/ 4 \right) dz = (C/4) \int_0^D \sum_m \sum_n \left(\underline{V}_m(z) \underline{V}_n^*(z) \right) dz \\
&= (C/4) \int_0^D \sum_m \sum_n \left[\underline{V}_{mo} \cos(m\pi z/D) \underline{V}_{no}^* \cos(n\pi z/D) \right] dz \\
&= (C/4) \sum_n \int_0^D |\underline{V}_{no}|^2 \cos^2(n\pi z/D) dz = (CD/8) \sum_n |\underline{V}_{no}|^2 \\
&= \sum_n w_{eTn}
\end{aligned}
\tag{7.4.20}
$$

where the total electric energy stored in the nth mode is:

$$w_{eTn} = CD|\underline{V}_{no}|^2 \Big/ 8 \text{ [J]} \tag{7.4.21}$$

Since the time average electric and magnetic energies in any resonant mode are equal, the total energy is twice the value given in (7.4.21). Thus the total energy w_T stored on this TEM line is the sum of the energies stored in each resonant mode separately because all m≠n cross terms in (7.4.20) integrate to zero. Superposition of energy applies here because all TEM$_m$ resonant modes are spatially orthogonal. The same is true for any TEM resonator terminated with short or open circuits. Although spatial orthogonality may not apply to the resonator of Figure 7.4.2(a), which is terminated with a lumped reactance, the modes are still orthogonal because they have different frequencies, and integrating $v_m(t)v_n(t)$ over time also yields zero if m≠n.
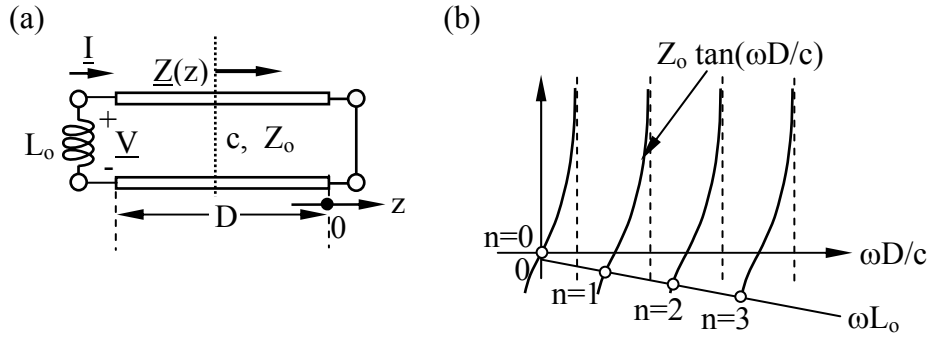
Figure 7.4.2   Inductively loaded TEM transmission line resonator.

Other types of resonator also generally have orthogonal resonant modes, so that in general:

$$w_T = \sum_n w_{Tn} \tag{7.4.22}$$

If a TEM resonator is terminated with a reactive impedance such as $j\omega L$ or $1/j\omega C$, then energy is still trapped but the resonant frequencies are non-uniformly distributed. Figure 7.4.2(a) illustrates a lossless short-circuited TEM line of length D that is terminated with an inductor L_o. Boundary conditions immediately yield an expression for the resonant frequencies ω_n. The impedance of the inductor is $j\omega L_o$ and that of the TEM line follows from (7.3.6) for $Z_L = 0$:

$$\underline{Z}(z) = Z_o(\underline{Z}_L - jZ_o\tan kz)/(\underline{Z}_o - jZ_L\tan kz) = -jZ_o\tan kz \tag{7.4.23}$$

Since the current $\underline{I}$ and voltage $\underline{V}$ at the inductor junction are the same for both the transmission line and the inductor, their ratios must also be the same except that we define $\underline{I}$ to be flowing out of the inductor into the TEM line, which changes the sign of $+j\omega L_o$; so:

$$\underline{V}/\underline{I} = -j\omega L_o = jZ_o\tan kD \tag{7.4.24}$$

$$\omega_n = -\frac{Z_o}{L_o}\tan kD = -\frac{Z_o}{L_o}\tan(\omega_n D/c) > 0 \tag{7.4.25}$$

The values of ω_n that satisfy (7.4.25) are represented graphically in Figure 7.4.2(b), and are spaced non-uniformly in frequency. The resonant frequency $\omega_o = 0$ corresponds to direct current and pure magnetic energy storage. Figure 7.4.2(b) yields ω_n for a line shorted at both ends when $L_o = 0$, and shows that for small values of L_o (perturbations) that the shift in resonances $\Delta\omega_n$ are linear in L_o.

We generally can tune resonances to nearby frequencies by changing the resonator slightly. Section 9.4.2 derives the following expression (7.4.26) for the fractional change $\Delta f/f$ in any resonance f as a function of the incremental increases in average electric ($\Delta\omega_e$) and magnetic

- 218 -

$(\Delta\omega_m)$ energy storage and, equivalently, in terms of the incremental volume that was added to or subtracted from the structure, where W_e and W_m are the electric and magnetic energy densities in that added $(+\Delta v_{vol})$ or removed $(-\Delta v_{vol})$ volume, and w_T is the total energy associated with f. The energy densities can be computed using the unperturbed values of field strength to obtain approximate answers.

$$\Delta f/f = \left(\Delta w_e - \Delta w_m\right)/w_T = \Delta v_{ol}\left(W_e - W_m\right)/w_T \qquad \text{(frequency perturbation)} \qquad (7.4.26)$$

A simple example illustrates its use. Consider the TEM resonator of Figure 7.4.2(a), which is approximately short-circuited at the left end except for a small tuning inductance L_o having an impedance $|j\omega L| \ll Z_o$. How does L_o affect the resonant frequency f_1? One approach is to use (7.4.25) or Figure 7.4.2(b) to find w_n. Alternatively, we may use (7.4.26) to find $\Delta f = - f_1 \times \Delta w_m/w_T$, where $f_1 \cong c/\lambda \cong c/2D$ and $\Delta w_m = L_o|\underline{I}'|^2/4 = |\underline{V}'|^2/4\omega^2 L$, where $\underline{I}'$ and $\underline{V}'$ are exact. But the unperturbed voltage at the short-circuited end of the resonator is zero, so we must use $\underline{I}'$ because perturbation techniques require that only small fractional changes exist in parameters to be computed, and a transition from zero to any other value is not a perturbation. Therefore $\Delta w_m = L_o|\underline{I}_o|^2/4$. To cancel $|\underline{I}_o|^2$ in the expression for Δf, we compute w_T in terms of voltage: $w_T = 2w_m = 2\int_0^D \left(L|\underline{I}(z)|^2/4\right)dz = DLI_o^2/4$. Thus:

$$\Delta f = \Delta f_n = -f_n\left(\frac{\Delta w_m}{w_T}\right) = -f_n\frac{|\underline{I}_o|^2/4}{DL|\underline{I}_o|^2/4} = -f_n\frac{L_o}{LD} \qquad (7.4.27)$$

7.4.3 Resonator losses and Q

All resonators dissipate energy due to resistive losses, leakage, and radiation. Since dissipation is proportional to resonator energy content and to the squares of current or voltage, the decay of field strength and stored energy is generally exponential in time. Each resonant frequency f_n has its own rate of energy decay, characterized by the dimensionless *quality factor Q_n*, which is generally the number of radians $\omega_n t$ required for the total energy w_{Tn} stored in mode n to decay by a factor of $1/e$:

$$w_{Tn}(t) = w_{Tn0}e^{-\omega_n t/Q_n} \text{J} \quad [\quad] \qquad (7.4.28)$$

Q_n is easily related to P_n, the power dissipated by mode n:

$$P_n \cong -dw_{Tn}/dt = \omega_n w_{Tn}/Q_n \text{W} [\quad] \qquad (7.4.29)$$

$$Q_n \cong \omega_n w_{Tn}/P_n \qquad \text{(quality factor Q)} \qquad (7.4.30)$$

The rate of decay for each mode depends on the location of the resistive or radiating elements relative to the peak currents or voltages for that mode. For example, if a resistive element

experiences a voltage or current null, there is no dissipation. These relations apply to all resonators, for example, RLC resonators: (3.5.20–23).

Whether a resonator is used as a band-pass or band-stop filter, it has a bandwidth $\Delta\omega$ within which more than half the peak power is passed or stopped, respectively. This half-power bandwidth $\Delta\omega$ is simply related to Q by (3.5.36):

$$Q_n \cong \omega_n/\Delta\omega_n \qquad (7.4.31)$$

The concept and utility of Q and the use of resonators in circuits are developed further in Section 7.4.4.

Loss in TEM lines arises because the wires are resistive or because the medium between the wires conducts slightly. In addition, lumped resistances may be present, as suggested in Figure 7.4.3(b) and (d). If these resistances do not significantly perturb the lossless voltage and current distributions, then the power dissipated and Q of each resonance ω_n can be easily estimated using *perturbation techniques*. The perturbation method simply involves computing power dissipation using the voltages or currents appropriate for the lossless case under the assumption that the fractional change induced by the perturbing element is small (perturbations of zero-valued parameters are not allowed). The examples below illustrate that perturbing resistances can be either very large or very small.
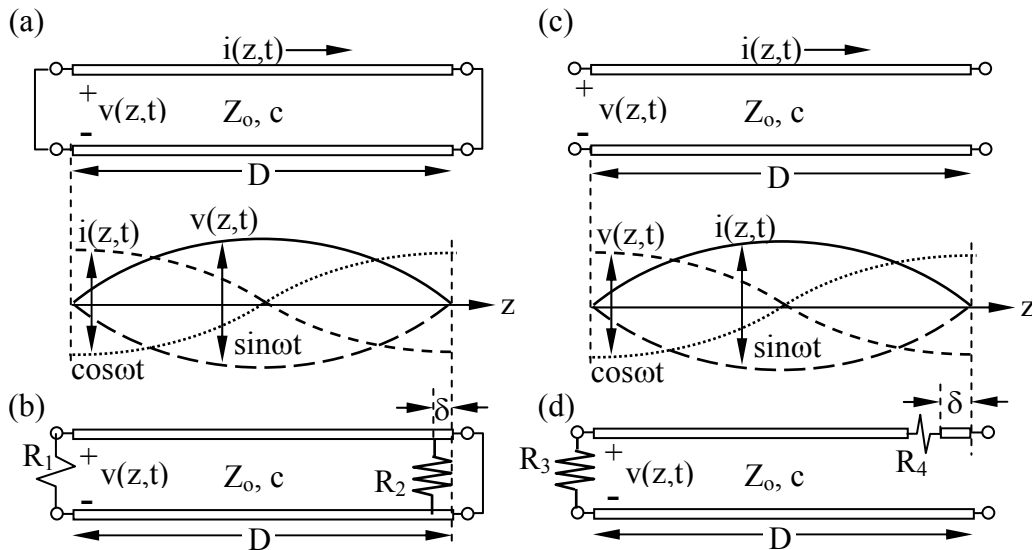

Figure 7.4.3   TEM resonators perturbed by loss.

Consider first the illustrated ω_1 resonance of Figure 7.4.3(a) as perturbed by the small resistor $R_1 \ll Z_o$; assume R_2 is absent. The nominal current on the TEM line is:

$$i(t,z) = R_e\{\underline{I}e^{j\omega t}\} = I_o \sin \omega t \cos(\pi z/D) \qquad (7.4.32)$$

The power P_1 dissipated in R_1 at $z = 0$ using the unperturbed current is:

$$P_1 = \left\langle i^2(t, z = 0) \right\rangle R_1 = I_o^2 R_1/2 \qquad (7.4.33)$$

The corresponding total energy w_{T1} stored in this unperturbed resonance is twice the magnetic energy:

$$w_{T1} = 2\int_0^D \left(L\left\langle i^2(t) \right\rangle/2 \right) dz = DLI_o^2/4 \text{ [J]} \qquad (7.4.34)$$

Using (7.4.30) for Q and (7.4.5) for ω we find:

$$Q_1 \cong \omega_1 w_{T1}/P_1 = (\pi c/D)\left(DLI_o^2/4 \right)/\left(I_o^2 R_1/2 \right) = \pi c L/2R_1 = (Z_o/R_1)\pi/2 \qquad (7.4.35)$$

Thus $Q_1 \cong Z_o/R_1$ and is high when $R_1 \ll Z_o$; in this case R_1 is truly a perturbation, so our solution is valid.

A more interesting case involves the loss introduced by R_2 in Figure 7.4.3(b) when R_1 is zero. Since the unperturbed shunting current at that position on the line is zero, we must use instead the unperturbed voltage $v(z,t)$ to estimate P_1 for mode 1, where that nominal line voltage is:

$$v(z,t) = V_o \sin \omega t \sin(\pi z/D) \qquad (7.4.36)$$

The associated power P_1 dissipated at position δ, and total energy w_{T1} stored are:

$$P_1 \cong \left\langle v(\delta,t)^2 \right\rangle/R_2 = V_o^2 \sin^2(\pi\delta/D)/2R_2 \cong (V_o\pi\delta/D)^2/2R_2 \text{ [W]} \qquad (7.4.37)$$

$$w_{T1} \cong 2\int_0^D \left(C\left\langle v^2(z,t) \right\rangle/2 \right) dz = DCV_o^2/4 \text{ [J]} \qquad (7.4.38)$$

Note that averaging $v^2(z,t)$ over space and time introduces two factors of 0.5. Using (7.4.35) for Q and (7.4.5) for ω we find:

$$Q_1 \cong \omega_1 w_{T1}/P_1 = (\pi c/D)\left(DCV_o^2/4 \right)/\left[(V_o\pi\delta/D)^2/2R_2 \right] = (D/\delta)^2 (R_2/2\pi Z_o) \qquad (7.4.39)$$

Thus Q_1 is high and R_2 is a small perturbation if $D \gg \delta$, even if $R_2 < Z_o$. This is because a leakage path in parallel with a nearby short circuit can be a perturbation even if its conductance is fairly high.

In the same fashion Q can be found for the loss perturbations of Figure 7.4.3(d). For example, if $R_4 = 0$, then [following (7.4.39)] the effect of R3 is:

$$Q_1 \cong \omega_1 w_{T1}/P_1 = (\pi c/D)\left(DCV_o{}^2\big/4\right)\Big/\left(V_o{}^2\big/2R_3\right) = (\pi/2)(R_3/Z_o) \qquad (7.4.40)$$

In this case R_3 is a perturbation if $R_3 \gg Z_o$. Most R_4 values are also perturbations provided $\delta \ll$ D, similar to the situation for R_2, because any resistance in series with a nearby open circuit will dissipate little power because the currents there are so small.

Example 7.4A

What is the Q of a TEM resonator of length D characterized by ω_o, C, and G?

Solution: Equation (7.4.40) says $Q = \omega_o w_T/P_d$, where the power dissipated is given by (7.1.61): $P_d = \int_0^D \left(G|\underline{V}(z)|^2\big/2\right)dz$. The total energy stored w_T is twice the average stored electric energy $w_T = 2w_e = 2\int_0^D \left(C|\underline{V}(z)|^2\big/4\right)dz$ [see 7.1.32)]. The voltage distribution $|\underline{V}(z)|$ in the two integrals cancels in the expression for Q, leaving Q = $2\omega_o C/G$.

7.4.4   Coupling to resonators

Depending on how resonators are coupled to circuits, they can either pass or stop a band of frequencies of width $\sim\Delta\omega_n$ centered on a resonant frequency ω_n. This effect can be total or partial; that is, there might be total rejection of signals either near resonance or far away, or only a partial enhancement or attenuation. This behavior resembles that of the series and parallel RLC resonators discussed in Section 3.5.2.

   Figure 7.4.4 shows how both series and parallel RLC resonators can block all the available power to the load resistor R_L near resonance, and similar behavior can be achieved with TEM resonators as suggested below; these are called *band-stop filters*. Alternatively, both series and parallel RLC resonators can pass to the load resistor the band near resonance, as suggested in Figure 3.5.3; these are called *band-pass filters*. In Figure 7.4.4(a) the series LC resonator shorts out the load R near resonance, while in (b) the parallel LC resonator open-circuits the load conductance G; the resonant band is stopped in both cases.
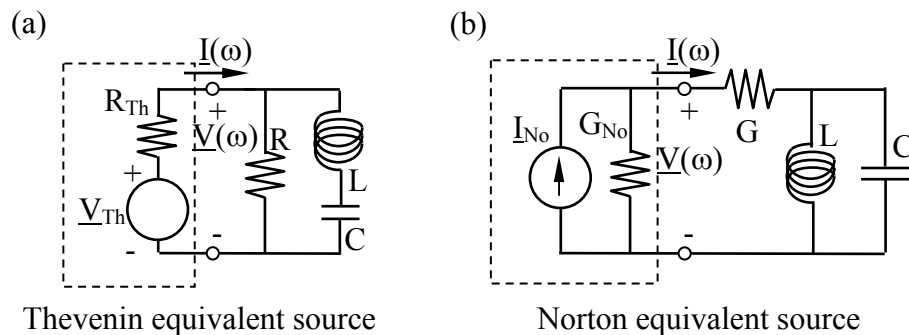

Figure 7.4.4   Band-stop RLC resonators.

The half-power width $\Delta\omega$ of each resonance is inversely proportional to the *loaded Q*, where Q_L was defined in (3.5.40), P_{DE} is the power dissipated externally (in the source resistance R_{Th}), and P_{DI} is the power dissipated internally (in the load R):

$$Q_L \equiv \omega w_T / (P_{DI} + P_{DE}) \qquad\qquad \text{(loaded Q)} \quad (7.4.41)$$

$$\Delta\omega_n = \omega_n / Q_n \left[\text{radians s}^{-1} \right] \qquad\qquad \text{(half-power bandwidth)} \quad (7.4.42)$$

When $\omega = (LC)^{-0.5}$ the LC resonators are either open- or short-circuit, leaving only the source and load resistors, R_{Th} and R_L. At the frequency f of maximum power transfer the fraction of the available power that can be passed to the load is determined by the ratio $Z_n' = R_L / R_{Th}$. For example, if the power source were a TEM transmission line of impedance $Z_o \equiv R_{Th}$, then the minimum fraction of incident power reflected from the load (7.2.22) would be:

$$|\underline{\Gamma}|^2 = |(\underline{Z}_n' - 1)/(\underline{Z}_n' + 1)|^2 \qquad\qquad (7.4.43)$$

The fraction reflected is zero only when the normalized load resistance $Z_n' = 1$, i.e., when $R_L = R_{Th}$. Whether the maximum transfer of power to the load occurs at resonance ω_n (band-pass filter) or only at frequencies removed more than $\sim\Delta\omega$ from ω_n (band-stop filter) depends on whether the current is blocked or passed at ω_n by the LC portion of the resonator. For example, Figures 3.5.3 and 7.4.4 illustrate two forms of band-pass and band-stop filter circuits, respectively.

Resonators can be constructed using TEM lines simply by terminating them at both ends with impedances that reflect most or all incident power so that energy remains largely trapped inside, as illustrated in Figure 7.4.5(a). Because the load resistance R_L is positioned close to a short circuit ($\delta \ll \lambda/4$), the voltage across R_L is very small and little power escapes, even if $R_L \cong Z_o$. The Q for the ω_1 resonance is easily calculated by using (7.4.40) and the expression for line voltage (7.4.36):

$$v(z,t) = V_o \sin \omega t \sin (\pi z/D) \qquad\qquad (7.4.44)$$

$$Q \equiv \omega_1 w_T / P_D = (\pi c/D)(DCV_o{}^2/4)/\left[V_o{}^2 \sin^2 (\pi\delta/D)/2R_L \right]$$
$$= (\pi R_L/2Z_o)/\sin^2 (\pi\delta/D) \cong (D/\delta)^2 R_L/2\pi Z_o \quad (\text{for } \delta \ll D) \qquad (7.4.45)$$

Adjustment of δ enables achievement of any desired Q for any given R_L in an otherwise lossless system. If we regard R_L as internal to the resonator then the Q calculated above is the *internal Q*, Q_I.
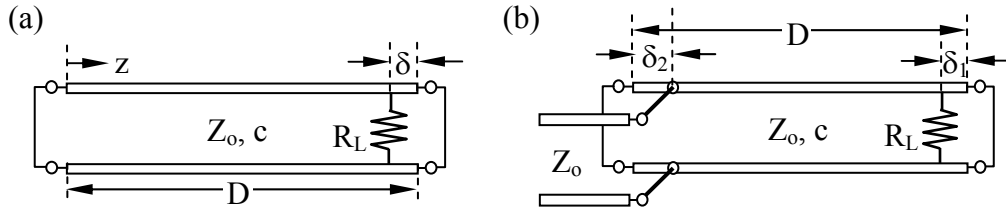
Figure 7.4.5   Coupled TEM resonator.

We may connect this resonator externally by adding a feed line at a short distance δ_2 from its left end, as illustrated in Figure 7.4.5(b). If the feed line is matched at its left end then the *external Q*, Q_E, associated with power dissipated there is given by (7.4.45) for $\delta = \delta_2$ and $R_L = Z_o$. By adjusting δ_2 any Q_E can be obtained. Figures 3.5.3 and 7.4.4 suggest how the equivalent circuits for either band-pass or band-stop filters can match all the available power to the load if $R_{Th} = R_L$ and therefore $Q_E = Q_I$. Thus all the available power can be delivered to R_L in Figure 7.4.5(b) for any small δ_1 by selecting δ_2 properly; if δ_2 yields a perfect match at resonance, we have a *critically coupled resonator*. If δ_2 is larger than the critically coupled value, then the input transmission line is too strongly coupled, $Q_E < Q_I$, and we have an *over-coupled resonator*; conversely, smaller values of δ_2 yield $Q_E > Q_I$ and undercoupling. The bandwidth of this band-pass filter $\Delta\omega$ is related to the *loaded Q*, Q_L, as defined in (7.4.41) where:

$$Q_L^{-1} = Q_I^{-1} + Q_E^{-1} = \Delta\omega/\omega \qquad\qquad (7.4.46)$$

If the band-pass filter of Figure 7.4.5(b) is matched at resonance so $Q_E = Q_I$, it therefore has a bandwidth $\Delta\omega = 2\omega/Q_I$, where Q_I is given by (7.4.39) and is determined by our choice of δ_1. Smaller values of δ_1 yield higher values for Q_L and narrower bandwidths $\Delta\omega$. In the special case where R_L corresponds to another matched transmission line with impedance Z_o, then a perfect match at resonance results here when $\delta_1 = \delta_2$.

Many variations of the coupling scheme in Figure 7.4.5 exist. For example, the feed line and resonator can be isolated by a shunt consisting of a large capacitor or a small inductor, both approximating short circuits relative to Z_o, or by a high-impedance block consisting of a small capacitor or large inductor in series. Alternatively, an external feed line can be connected in place of R_4 in Figure 7.4.3(d). In each weakly coupled case perturbation methods quickly yield Q_I and Q_E, and therefore Q_L, $\Delta\omega$, and the impedance at resonance.

The impedance at resonance can be found once Q_E, Q_I, and Z_o for the feedline are known, and once it is known whether the resonance is a series or parallel resonance. Referring to Figures 3.5.3 and 7.4.4 for equivalent circuits for band-pass and band-stop filters, respectively, it is clear that if $Q_E = Q_I$, then band-pass resonators are matched at resonance while band-stop series-resonance resonators are short circuits and parallel-resonance resonators are open circuits. Away from resonance band-pass resonators become open circuits for series resonances and short circuits for parallel resonances, while both types of band-stop resonator become matched loads if $Q_E = Q_I$. At resonance all four types of resonator have purely real impedances and reflection coefficients Γ that can readily be found by examining the four equivalent circuits cited above.

Sometimes unintended resonances can disrupt systems. For example, consider a waveguide that can propagate two modes, only one of which is desired. If a little bit of the unwanted mode is excited at one end of the waveguide, but cannot escape through the lines connected at each end, then the second mode is largely trapped and behaves as a weakly coupled resonator with its own losses. At each of its resonances it will dissipate energy extracted from the main waveguide. If the internal losses happen to cause $Q_E = Q_I$ for these parasitic resonances, no matter how weakly coupled they are, they can appear as a matched load positioned across the main line; dissipation by parasitic resonances declines as their internal and external Q's increasingly differ.

The ability of a weakly coupled resonance to have a powerful external effect arises because the field strengths inside a low-loss resonator can rise to values far exceeding those in the external circuit. For example, the critically coupled resonator of Figure 7.4.5(b) for $R_L = Z_o$ and $\delta_1 = \delta_2$, has internal voltages $v(z,t) = V_o \sin\omega t \sin(\pi z/D)$ given by (7.4.44), where the maximum terminal voltage is only $V_o \sin(\pi\delta/D) \cong V_o \pi\delta/D << V_o$. Thus a parasitic resonance can slowly absorb energy from its surroundings at its resonant frequency until its internal fields build to the point that even with weak coupling it has a powerful effect on the external fields and thus reaches an equilibrium value. It is these potentially extremely strong resonant fields that enables critically coupled resonators to couple energy into poorly matched loads--the fields in the resonator build until the power dissipated in the load equals the available power provided. In some cases the fields can build to the point where the resonator arcs internally, as can happen with an empty microwave oven without an extra internal load to prevent it.

This analysis of the resonant behavior of TEM lines is approximate because the resonator length measured in wavelengths is a function of frequency within $\Delta\omega$, so exact answers require use the TEM analysis methods of Sections 7.2–3, particularly when $\Delta\omega$ becomes a non-trivial fraction of the frequency difference between adjacent resonances.

Example 7.4B

Consider a variation of the coupled resonator of Figure 7.4.5(b) where the resonator is open-circuited at both ends and the weakly coupled external connections at δ_1 and δ_2 from the ends are in series with the 100-ohm TEM resonator line rather than in parallel. Find δ_1 and δ_2 for: $Q_L = 100$, $Z_o = 100$ ohms for both the feed line and resonator, $R_L = 50$ ohms, and the resonator length is $D \cong \lambda/2$, where λ is the wavelength within the resonator.

Solution: For critical coupling, $Q_E = Q_I$, so the resonator power lost to the input line, $|\underline{I}_2|^2 Z_o/2$, must equal that lost to the load, $|\underline{I}_1|^2 R_L/2$, and therefore $|\underline{I}_1|/|\underline{I}_2| = (Z_o/R_L)^{0.5} = 2^{0.5}$. Since the $\lambda/2$ resonance of an open-circuited TEM line has $I(z) \cong I_o \sin(\pi z/D) \cong \pi z/D$ for $\delta << D/\pi$ (high Q), therefore $|\underline{I}_1|/|\underline{I}_2| = [\sin(\pi\delta_1/D)]/[\sin(\pi\delta_2/D)] \cong \delta_1/\delta_2 \cong 2^{0.5}$. Also, $Q_L = 100 = 0.5 \times Q_I = 0.5\omega_o w_T/P_{DI}$, where: $\omega_o = 2\pi f_o = 2\pi c/\lambda = \pi c/D$; $w_T = 2w_m = 2\int_0^D (L|I|^2/4)dz \cong LI_o^2 D/4$; and $P_{DI} = |I(\delta_1)|^2 R_L/2 = I_o^2 \sin^2(\pi\delta_1/D)R_L/2 \cong (I_o\pi\delta_1/D)^2 R_L/2$. Therefore $Q_I = \omega_o w_T/P_{DI} = 200 = (\pi c/D)(LI_o^2 D/4)/[(I_o\pi\delta_1/D)^2 R_L/2] = (D/\delta_1)^2 (Z_o/R_L)/\pi$, where $cL = Z_o = 100$. Thus $\delta_1 = \pi^{-0.5}D/10$ and $\delta_2 = \delta_1 2^{-0.5} = (2\pi)^{-0.5}D/10$.

7.4.5 Transients in TEM resonators

TEM and cavity resonators have many resonant modes, all of which can be energized simultaneously, depending on initial conditions. Because Maxwell's equations are linear, the total fields can be characterized as the linear superposition of fields associated with each excited mode. This section illustrates how the relative excitation of each TEM resonator mode can be determined from any given set of initial conditions, e.g. from $v(z, t = 0)$ and $i(z, t = 0)$, and how the voltage and current subsequently evolve. The same general method applies to modal excitation of cavity resonators. By using a similar orthogonality method to match boundary conditions in space rather than in time, the modal excitation of waveguides and optical fibers can be found, as discussed in Section 9.3.3.

   The central concept developed below is that any initial condition in a TEM resonator at time zero can be replicated by superimposing some weighted set of voltage and current modes. Once the phase and magnitudes of those modes are known, the voltage and current are then known for all time. The key solution step uses the fact that the mathematical functions characterizing any two different modes a and b, e.g. the voltage distributions $\underline{V}_a(z)$ and $\underline{V}_b(z)$, are spatially othogonal: $\int \underline{V}_a(z)\underline{V}_b{}^*(z)\,dz = 0$.

   Consider the open-circuited TEM resonator of Figure 7.4.3(c), for which $\underline{V}_{n-} = \underline{V}_{n+}$ for any mode n because the reflection coefficient at the open circuit at $z = 0$ is +1. The resulting voltage and current on the resonator for mode n are:[39]

$$\underline{V}_n(z) = \underline{V}_{n+}e^{-jk_nz} + \underline{V}_{n-}e^{+jk_nz} = 2\underline{V}_{n+}\cos k_nz \qquad (7.4.47)$$

$$\underline{I}_n(z) = Y_o\left(\underline{V}_{n+}e^{-jk_nz} - \underline{V}_{n-}e^{+jk_nz}\right) = -2jY_o\underline{V}_{n+}\sin k_nz \qquad (7.4.48)$$

where $k_n = \omega_n/c$ and (7.4.5) yields $\omega_n = n\pi c/D$. We can restrict the general expressions for voltage and current to the moment $t = 0$ when the given voltage and current distributions are $v_o(z)$ and $i_o(z)$:

$$v(z,t = 0) = v_o(z) = \sum_{n=0}^{\infty} R_e\left\{\underline{V}_n(z)e^{j\omega_nt}\right\}_{t=0} = \sum_{n=0}^{\infty} R_e\left\{2\underline{V}_{n+}\cos k_nz\right\} \qquad (7.4.49)$$

$$i(z,t = 0) = i_o(z) = \sum_{n=0}^{\infty} R_e\left\{\underline{I}_n(z)e^{j\omega_nt}\right\}_{t=0} = Y_o\sum_{n=0}^{\infty} I_m\left\{2\underline{V}_{n+}\sin k_nz\right\} \qquad (7.4.50)$$

---

[39] Where we recall $\cos\phi = (e^{j\phi} + e^{-j\phi})/2$ and $\sin\phi = (e^{j\phi} - e^{-j\phi})/2j$.

We note that these two equations permit us to solve for both the real and imaginary parts of $\underline{V}_{n+}$, and therefore for $v(z,t)$ and $i(z,t)$. Using spatial orthogonality of modes, we multiply both sides of (7.4.49) by $\cos(m\pi z/D)$ and integrate over the TEM line length D, where $k_n = n\pi z/D$:

$$\int_0^D v_o(z)\cos\,(m\pi z/D)\,dz = \int_0^D \sum_{n=0}^{\infty} R_e\{2\underline{V}_{n+}\cos\,k_n z\}\cos(m\pi z/D)\,dz$$

$$= \sum_{n=0}^{\infty} R_e\{2\underline{V}_{n+}\}\int_0^D \cos(n\pi z/D)\cos(m\pi z/D)\,dz = 2R_e\{\underline{V}_{n+}\}(D/2)\delta_{mn} \qquad (7.4.51)$$

where $\delta_{mn} \equiv 0$ if $m \neq n$, and $\delta_{mn} \equiv 1$ if $m = n$. Orthogonality of modes thus enables this integral to single out the amplitude of each mode separately, yielding:

$$R_e\{\underline{V}_{n+}\} = D^{-1}\int_0^D v_o(z)\cos(n\pi z/D)\,dz \qquad (7.4.52)$$

Similarly, we can multiply (7.4.50) by $\sin(m\pi z/D)$ and integrate over the length D to yield:

$$Im\{\underline{V}_{n+}\} = Z_o D^{-1}\int_0^D i_o(z)\sin(n\pi z/D)\,dz \qquad (7.4.53)$$

Once $\underline{V}_n$ is known for all n, the full expressions for voltage and current on the TEM line follow, where $\omega_n = \pi n c/D$:

$$v(z,t) = \sum_{n=0}^{\infty} R_e\{\underline{V}_n(z)e^{j\omega_n t}\} \qquad (7.4.54)$$

$$i(z,t) = \sum_{n=0}^{\infty} R_e\{\underline{I}_n(z)e^{j\omega_n t}\} = Y_o \sum_{n=0}^{\infty} I_m\{2_-V_{n+}\sin k_n z\} \qquad (7.4.55)$$

In general, each resonator mode decays exponentially at its own natural rate, until only the longest-lived mode remains.

As discussed in Section 9.3.3, the relative excitation of waveguide modes by currents can be determined in a similar fashion by expressing the fields in a waveguide as the sum of modes, and then matching the boundary conditions imposed by the given excitation currents at the spatial origin (not time origin). The real and imaginary parts of the amplitudes characterizing each waveguide propagation mode can then be determined by multiplying both sides of this boundary equation by spatial sines or cosines corresponding to the various modes, and integrating over the surface defining the boundary at $z = 0$. Arbitrary spatial excitation currents generally excite both propagating and evanescent modes in some combination. Far from the excitation point only the

- 227 -

propagating modes are evident, while the evanescent modes are evident principally as a reactance seen by the current source.

Chapter 8: Fast Electronics and Transient Behavior on TEM Lines

8.1 *Propagation and reflection of transient signals on TEM transmission lines*

8.1.1 Lossless transmission lines

The speed of computation and signal processing is limited by the time required for charges to move within and between devices, and by the time required for signals to propagate between elements. If the devices partially reflect incoming signals there can be additional delays while the resulting reverberations fade. Finally, signals may distort as they propagate, smearing pulse shapes and arrival times. These three sources of delay, i.e., propagation plus reverberation, device response times, and signal distortion are discussed in Sections 8.1, 8.2, and 8.3, respectively. These same issues apply to any system combining transmission lines and circuits, such as integrated analog or digital circuits, printed circuit boards, interconnections between circuits or antennas, and electrical power lines.

Transmission lines are usually paired parallel conductors that convey signals between devices. They are fundamental to every electronic system, from integrated circuits to large systems. Section 7.1.2 derived from Maxwell's equations the behavior of transverse electromagnetic (TEM) waves propagating between parallel plate conductors, and Section 7.1.3 showed that the same equations also govern any structure, even a dissipative one, for which the cross-section is constant along its length and that has at least two perfectly conducting elements between which the exciting voltage is applied. Using differential RLC circuit elements, this section below derives the same transmission-line behavior in a form that can readily be extended to transmission lines with resistive wires, as discussed later in Section 8.3.1. Since resistive wires introduce longitudinal electric fields, such lines are no longer pure TEM lines.

Equations (7.1.10) and (7.1.11) characterized the voltage v(t,z) and current i(t,z) on TEM structures with inductance L [H m^{-1}] and capacitance C [F m^{-1}] as:

$$dv/dz = -L\,di/dt \qquad\qquad\qquad (8.1.1)$$

$$di/dz = -C\,dv/dt \qquad\qquad\qquad (8.1.2)$$

These expressions were combined to yield the *wave equation* (7.1.14) for lossless TEM lines:

$$\left(d^2/dz^2 - LC\,d^2/dt^2\right)v(z,t) = 0 \qquad \text{(TEM wave equation)} \qquad (8.1.3)$$

One general solution to this wave equation is (7.1.16):

$$v(z,t) = v_+(z-ct) + v_-(z+ct) \qquad \text{(TEM voltage)} \qquad (8.1.4)$$

which corresponds to the superposition of forward and backward propagating waves moving at velocity $c = (LC)^{-0.5} = (\mu\varepsilon)^{-0.5}$. The current i(t,z) corresponding to (8.1.4) follows from substitution of (8.1.4) into (8.1.1) or (8.1.2), and differentiation followed by integration:

$$i(z,t) = Y_o \left[v_+ (z - ct) - v_- (z + ct) \right] \qquad \text{(TEM current)} \qquad (8.1.5)$$

Y_o is the *characteristic admittance* of the line, and the reciprocal of the *characteristic impedance* Z_o:

$$Z_o = Y_o^{-1} = (L/C)^{0.5} \text{ [Ohms]} \qquad \text{(characteristic impedance of lossless TEM line)} \qquad (8.1.6)$$

The value of Y_o follows directly from the steps above.

A more intuitive way to derive these equations utilizes an equivalent *distributed circuit* for the transmission line composed of an infinite number of differential elements with series inductance and parallel capacitance, as illustrated in Figure 8.1.1(a). This model is easily extended to non-TEM lines with resistive wires.
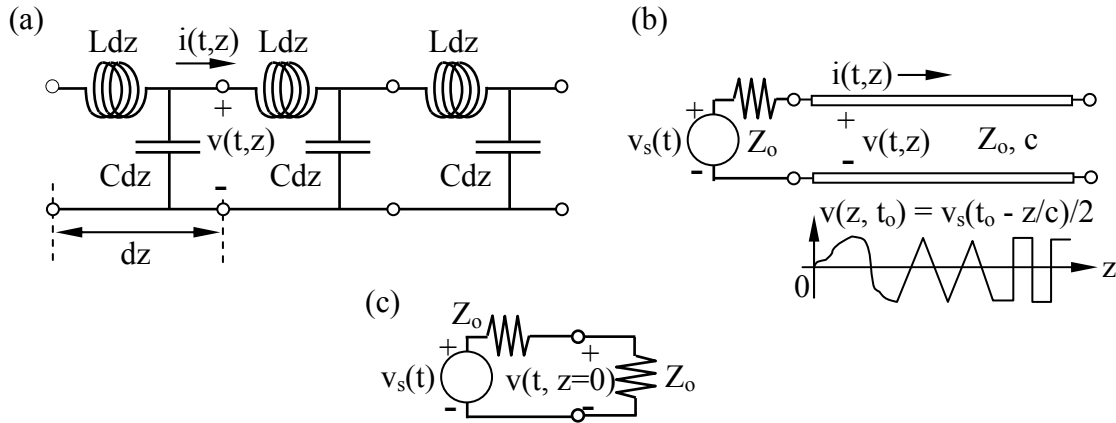
Figure 8.1.1   Distributed circuit model for lossless TEM transmission lines.

The inductance L [Henries m^{-1}] of the two conductors arises from the magnetic energy stored per meter of length, and produces a voltage drop dv across each incremental length dz of wire which is proportional to the time derivative of current through it[40]:

$$dv = -L \, dz \, (di/dt) \qquad (8.1.7)$$

Any current increase di across the distance dz, defined as di = i(t, z+dz) - i(t,z), would be supplied from charge stored in C [F m^{-1}]:

---

[40] An alternate equivalent circuit would have a second inductor in the lower branch equivalent to that in the upper branch; both would have value Ldz/2, and v(t,z) and i(t,z) would remain the same.

$$di = -C\,dz\,(dv/dt)$$
(8.1.8)

These two equations for dv and di are equivalent to (8.1.1) and (8.1.2), respectively, and lead to the same wave equation and general solutions derived in Section 7.1.2 and summarized above, where arbitrary waveforms propagate down TEM lines in both directions and superimpose to produce the total $v(z,t)$ and $i(t,z)$.

Two equivalent solutions exist for this wave equation: (8.1.4) and (8.1.9):

$$v(z,t) = f_+\left(t - z/c\right) + f_-\left(t + z/c\right)$$
(8.1.9)

The validity of (8.1.9) is easily shown by substitution into the wave equation (8.1.3), where again $c = (LC)^{-0.5}$. This alternate form is useful when relating line signals to sources or loads for which z is constant, as illustrated below. The first form (8.1.4) in terms of (z - ct) is more convenient when t is constant and z varies.

Waves can be launched on TEM lines as suggested in Figure 8.1.1(b). The line is driven by the Thevenin equivalent source $v_s(t)$ in series with the source resistance Z_o, which is matched to the transmission line in this case. Equations (8.1.4) and (8.1.5) say that if there is no negative traveling wave, then the ratio of the voltage to current for the forward wave on the line must equal $Z_o = Y_o^{-1}$. The equivalent circuit for this TEM line is therefore simply a resistor of value Z_o, as suggested in Figure 8.1.1(c). If the source resistance is also Z_o, then only half the source voltage $v_s(t)$ appears across the TEM line terminals at z = 0. Therefore the voltages at the left terminals (z = 0) and on the line $v(t,z)$ are:

$$v(t,z = 0) = v_s(t)/2 = v_+(t,z = 0)$$
(8.1.10)

$$v(t,z) = v_+\left(t - z/c\right) = v_s\left(t - z/c\right)/2 \qquad \text{(transmitted signal)}$$
(8.1.11)

where we have used the solution form of (8.1.9). The propagating wave in Figure 8.1.1(b) has half the amplitude of the Thevenin source $v_s(t)$ because the source was matched to the line so as to maximize the power transmitted from the given voltage $v_s(t)$. Note that (8.1.11) is the same as (8.1.10) except that z/c was subtracted from each. Equality is preserved if all arguments in an equation are shifted the same amount.

If the Thevenin source resistance were R, then the voltage-divider equation would yield the terminal and propagating voltage $v(t,z)$:

$$v(t,z) = v_s\left(t - z/c\right)\left[Z_o/(R + Z_o)\right]$$
(8.1.12)

This more general expression reduces to (8.1.10) when $R = Z_o$ and z = 0.

Example 8.1A

A certain integrated circuit with $\mu = \mu_o$ propagates signals at velocity c/2, and its TEM wires exhibit $Z_o = 100$ ohms. What are ε, L, and C for these TEM lines?

Solution: $c = (\mu_o\varepsilon_0)^{-0.5}$, and $v = c/2 = (\mu_o\varepsilon)^{-0.5}$; so $\varepsilon = 4\varepsilon_o$. Since $v = (LC)^{-0.5}$ and $Z_o = (L/C)^{0.5}$, $L = Z_o/v = 200/c = 6.67 \times 10^{-7}$ [Hy], and $C = 1/vZ_o = 1/200c = 1.67 \times 10^{-11}$ [F].

8.1.2   Reflections at transmission line junctions

If a transmission line connecting a source to a load is sufficiently short, then the effects of the line on reflections can be modeled by simply replacing it with a small lumped capacitor across the source terminals representing the capacitance between the wires, and a resistor in series with an inductor and the load, representing the resistance and inductance of the wires. If, however, the line length D is such that the propagation time $\tau_{line} = D/c$ is a non-trivial fraction of the shortest time constant of the load $\tau_{load,}$ then we should use transmission line models governed by the wave equation (e.g., 8.1.3). That is, the TEM wave equation should be used unless the line length D is:

$$D \ll c\tau_{load} \tag{8.1.13}$$

For larger values of D the propagation delays become important and a transmission line model must be used, as explained in Section 8.1.1. Section 8.1.1 also explained how signals are launched and propagate on TEM lines, and how the Thevenin equivalent circuit (8.1.6) for a passive transmission line as seen by the source is simply a resistor $Z_o = (L/C)^{0.5}$. This *characteristic impedance* Z_o of the transmission line is the ratio of the forward voltage $v_+(t,z)$ to the associated current $i_+(z,t)$. TEM signals are partially transmitted and partially reflected at each junction they encounter, where these junctions may be the intended load or simply places where the impedance Z_o of the transmission line changes. Sometimes multiple transmission lines meet at such junctions.

Section 7.2.2 (7.2.7) derived the *reflection coefficient* Γ for an arbitrary TEM wave $v_+(t,z)$ reflected by a load resistance R at z, where the normalized impedance of the load is $R_n = R/Z_o$:

$$v_-(t,z) = \Gamma v_+(t,z) \tag{8.1.14}$$

$$\Gamma = (R_n - 1)/(R_n + 1) \tag{8.1.15}$$

$$R_n \equiv R/Z_o \tag{8.1.16}$$

It is important to distinguish the difference between Γ for purely resistive loads, which is real, and $\underline{\Gamma}(\omega)$, which is complex and applies to any complex load impedance $\underline{Z}_L$. Here R and Γ are real.

Consider the example illustrated in Figure 8.1.2(a), where a TEM line is characterized by impedance Z_o and phase velocity c. The line is D meters long, open circuit at the right-hand end, and driven by a unit-step[41] voltage u(t). The equivalent circuit at the source end of the line is illustrated in (b), which is simply a voltage divider that places $v_s(t)/2$ volts across the line. But the voltage across the line equals the sum of the forward and backward moving waves, where a passive line at rest has no backward wave. Therefore the forward wave here at z = 0+ is simply u(t)/2, and the result is a voltage of 0.5 volts that moves down the line at velocity c, as illustrated in Figure 8.1.2(c) for t = t_o. The associated current i(z, t) is plotted in (d) for t = t_o, and is proportional to the voltage.
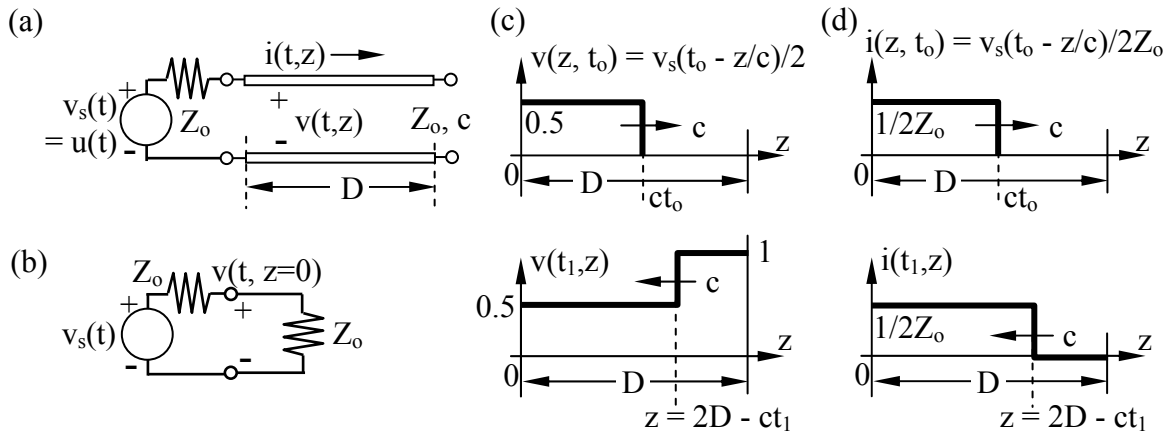
Figure 8.1.2   Step-function transients on a lossless transmission line.

Once the transient reaches the right-hand end, boundary conditions must again be satisfied, so there is a reflected voltage wave having $v_-(t, z=D) = \Gamma v_+(t, z=D)$, where $\Gamma = +1$, as given by (8.1.15) for $R_n \rightarrow \infty$. The total voltage on the line (8.1.9) is the sum of the forward and backward waves, each of value 0.5 volts, as illustrated in Figure 8.1.2(c) for $t = t_1 > D/c$. At t_1 the reflected voltage step is propagating leftward toward the source. The current at $t = t_1$ is plotted in Figure 8.1.2(d).

Although these voltage and current transients are most easily represented and understood graphically, they can also be derived and represented algebraically. For example, v(t, z=0) = u(t)/2 here, and therefore for t < D/c we have $v(t,z) = v_+(t - z/c) = u(t - z/c)/2$. Note that if we translate an argument on one side of an equation, we must impose the same translation on the other; thus $v(t,0) \rightarrow v(t - z/c)$ forces $u(t,0) \rightarrow u(t - z/c)$. Once the wave reflects from the open circuit we have $v(z,t) = v_+(t - z/c) + v_-(t + z/c)$. At z = D for t < 3D/c the boundary condition at the open circuit requires $\Gamma = +1$, so $v_-(t + D/c) = v_+(t - D/c) = u(t - D/c)/2$. From $v_-(t + D/c)$ we can find the more general expression $v_-(t + z/c)$ simply by operating on their arguments: $v_-(t + z/c) = v_-(t + D/c - D/c + z/c) = u(t - 2D/c + z/c)$. The total voltage for t < 2D/c is the sum of these forward and backward waves: $v(t,z) = [u(t - z/c) + u(t - 2D/c + z/c)]/2$. The same approach can represent line currents and also more complex examples.

---

[41] We use the notation u(t) to represent a *unit-step function* that is zero for t < 0, and unity for t ≥ 0. A *unit impulse* is represented by δ(t), which is zero for all |t| > ε in the limit where ε → 0, and the integral of δ(t) ≡ 0. ∫δ(t)dt = u(t).

When the reflected wave arrives back at the source, $\Gamma = 0$ because this source is matched to the transmission line. In this special case there are no further reflections. Steady state is therefore one volt on the line everywhere, with $v_+ = v_- = 0.5$ in perpetuity. The total line current is the difference between the forward and backward wave (8.1.5), as plotted in Figure 8.1.2(d) for t_1. The steady-state current is therefore zero. These steady state values correspond to $\omega \to 0$ and $\lambda \to \infty$, so the line is then much shorter than any wavelength of interest and can be considered static. We can easily see that an open-circuit line connected to a voltage source via any impedance at all will eventually assume the same voltage as the source, and the current will be zero, as it is here.

If the line were short-circuited at the right-hand end, then $\Gamma = -1$ and the voltage $v(z)$ at t_1 would resemble that of the current in Figure 8.1.2(d), with the values 0.5 and 0 volts, while the current $i(z)$ at t_1 would resemble that of the voltage in (c), with the values $0.5/Z_o$ and $1/Z_o$. The steady state values for voltage and current in this short-circuit case are zero and $1/Z_o$, respectively.

If the first transmission line were connected to a second passive infinite line of impedance Z_b, as illustrated in Figure 8.1.3(a), then the same computations would yield $v(t,z)$ and $i(t,z)$ on the first transmission line, where $R_n = Z_b/Z_o$. The solution on the second line follows from the boundary conditions: $v(t)$ and $i(t)$ are both continuous across the boundary. The resulting waveforms $v(t_1,z)$ and $i(t_1,z)$ at time $D/c < t_1 < 2D/c$ are plotted in Figure 8.1.3(b) for the case $R_n = 0.5$, so $\Gamma = -1/3$. In this case the current is increased by the reflection while the voltage is diminished. Independent of the incident waveform, the fraction of the incident power that is reflected is $(v_-/v_+)^2 = \Gamma^2$, where the reflection coefficient Γ is given by (8.1.15); the transmitted fraction is $1 - \Gamma^2$.
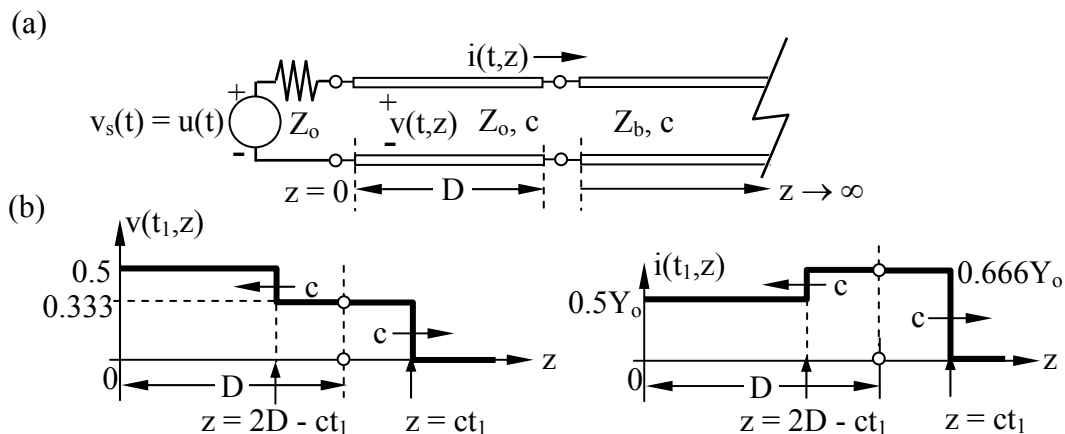

Figure 8.1.3   Step function incident upon a mismatched TEM line.

The principal consequence of this reflection phenomenon is that the voltage across a device may not be what was intended if there is an impedance mismatch between the TEM line and the device. This is an issue only when the line is sufficiently long that line delays are non-negligible

compared to circuit time constants (8.1.13). The analysis above is for linear resistive loads, but most loads are non-linear or reactive, and their treatment is discussed in Section 8.1.4.

8.1.3   Multiple reflections and reverberations

The reflected waves illustrated in Figures 8.1.2 and 8.1.3 eventually impact the source and may be reflected yet again. Since superposition applies if the sources and loads are linear, the contributions from each reflection can be separately determined and then added to yield the total voltage and current. That is, the reflected $v_-(t,z)$ will yield its own reflection at the source, and the fate of this reflection can be followed independently of the original forward wave. As usual when analyzing linear circuits, all sources are set to zero when determining the contribution of an independent source such as $v_-(t,z)$.

This paradigm is illustrated in Figure 8.1.4, which involves a unit-step current source driving an open-circuited TEM line that is characterized by Z_o, c, and length D. Figures 8.1.4(a), (b), and (c) illustrate the circuit, the voltage at t_1, and the current at t_1, respectively, where t_1 = D/2c. The reflection coefficient $\Gamma = 1$ (8.1.15) for the open circuit at z = D, so the incident Z_o-volt step is reflected positively, and the total voltage where they superimpose is $2Z_o$ volts, as illustrated in (d) for t_2 = 1.5D/c. The current at this moment is $Y_o[v_+(t - z/c) - v_-(t + z/c)]$, which is zero where the forward and reverse waves overlap, as illustrated in (e). When $v_-(t + z/c)$ is reflected from the left-hand end it sees $\Gamma = +1$ because, when using superposition, we consider the current source to be zero, corresponding to an open circuit. Thus an additional Z_o volts, associated with v_{+2}, adds to v_{+1} and v_{-1} to yield a total of $3Z_o$ volts, as illustrated in (f) at t_3; the notation v_{+i} refers to the ith forward wave v_+. This process continues indefinitely, with the voltage continuing to increase by Z_o volts every D/c seconds until something breaks down. Voltage breakdowns are expected when current sources feed open circuits; the finite rate of voltage increase is related to the total capacitance of the TEM line.
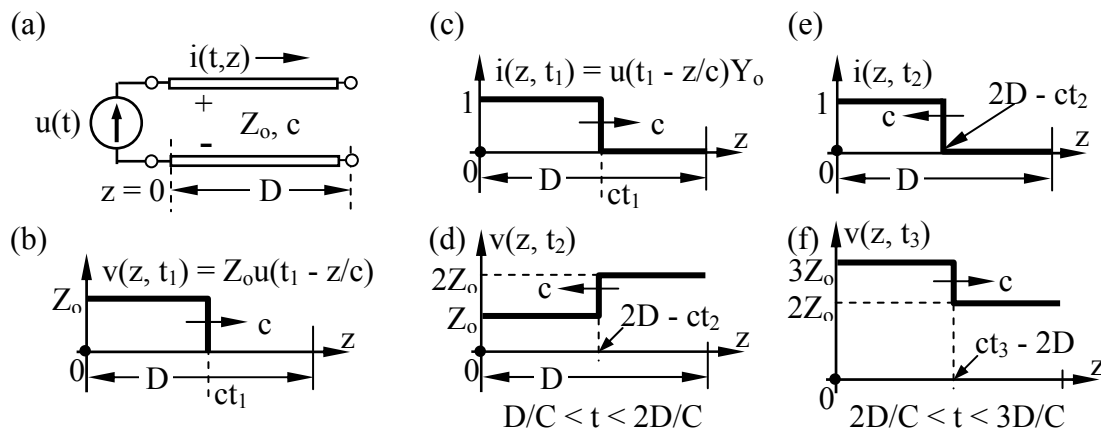

Figure 8.1.4   Transients for a current source driving an open-circuited TEM line.

The behavior of the current i(t,z) is interesting too. Figure 8.1.4(c) illustrates how the one-ampere current from the current source propagates down the line at velocity c, and (e) shows how the "message" that the line is open-circuited is returned: the current is returning to zero.

When this left-moving wave is reflected at the left-hand end the current is again forced to be one ampere by the current source. Thus the current distribution (c) also applies to (f) at t_3. This oscillation between one and zero amperes continues indefinitely, much like an unresolved argument between two people, each end of the line forcing the current to satisfy its own boundary conditions while that message propagates back and forth at velocity c.

Example 8.1B

A unit step voltage source u(t) with no source resistance drives a short-circuited air-filled TEM line of length D and characteristic impedance Z_o = 1 ohm. What current i(t) flows through the short circuit at the end of this TEM line?

Solution: The unit step will propagate down the line, be reflected at the short circuit at z = D where the reflection coefficient Γ_D = -1, and travel back to the voltage source at z = 0, which this transient sees as a short circuit (Δv = 0), also having Γ_S = -1. So, after a round-trip delay of 2D/c, the voltage everywhere on the line is zero, after which a new step voltage travels down the line and superimposes on the first step voltage, thus adding a second step to the current i(t, z=D). This process continues indefinitely as i(t) steps in 1-ampere increments every 2D/C seconds monotonically toward infinity, which is the expected current when a voltage source is short-circuited. The effect of the line is simply to slow this result as the current and stored magnetic energy on the line build up. More precisely, v(t, z=0) $\equiv$ u(t) = v_{+1}(t, z=0). v_{+1}(t, z=D) = u(t - D/c), so the TEM line presents an equivalent circuit at z = D having Thevenin voltage v_{Th} = $2v_{+1}$(t, D) = 2u(t - D/c), and Thevenin impedance Z_o; this yields i(t) = 2u(t - D/c)/Z_o for t < 3D/c. Therefore v_{-1}(t, D) = $\Gamma_D v_{+1}$(t, D) = -u(t - D/c), so v_{+2}(t, 0) = $\Gamma_S v_{-1}$(t, 0) = u(t - 2D/c). At z = D this second step increases the Thevenin voltage by 2u(t - 3D/c) and increases the current by 2u(t - 3D/c)/Z_o, where Z_o = 1 ohm. Therefore i(t) = $\Sigma_{n=0}^{\infty}$ 2u(t - [2n+1]D/c).

8.1.4   Reflections by mnemonic or non-linear loads

Most junctions involve mnemonic[42] or non-linear loads, where *mnemonic loads* are capacitors, inductors, or other energy storage devices that have characteristics depending on the past. *Non-linear loads* include diodes, transistors, and voltage- or current-dependent capacitors and inductors. In either case the response to arbitrary waveforms cannot be determined by the simple methods described in the previous section. However by simply replacing the transmission line by its equivalent circuit, the voltage and current can generally be easily found, first at the junction and then on the transmission line.

The equivalent circuit for an unexcited transmission line is simply a resistor of value Z_o because the ratio $\Delta v/\Delta i$ for any excitation is always Z_o. Determining the voltage across this Z_o is generally straightforward even if the source driving the line contains capacitors, inductors,

---

[42] Mnemonic means "involving memory".

diodes, or similar devices. The forward-propagating wave voltage is simply the terminal voltage, as demonstrated in Figures 8.1.2–4.

The Thevenin equivalent circuit for an energized TEM line has a Thevenin voltage source V_{Th} in series with the Thevenin impedance of the line: $Z_{Th} = Z_o$. Note that the equivalent impedance for a TEM line is exactly Z_o, regardless of any loads on the line. The influence of the load at the far end of the line is manifest only in reflected waves that may propagate from it toward the observer, as discussed in the previous section.

The Thevenin equivalent voltage of any linear system is simply its open-circuit voltage. The open-circuit voltage of a transmission line is twice the amplitude of any incident voltage waveform because the reflection coefficient Γ for an open circuit is +1, which doubles the incidence voltage at the junction position z_J:

$$V_{Th}(t,z_J) = v_+(t,z_J) + v_-(t,z_J) = 2v_+(t,z_J) \qquad (8.1.17)$$

The procedure for analyzing a TEM line terminated by any load at $z = z_J$ is then to: 1) solve for the wave $v_+(t - z/c)$ traveling toward the load of interest, 2) set $V_{Th} = 2v_+(t - z_J/c)$ and $Z_{Th} = Z_o$, 3) solve for the terminal voltage $v(t, z_J)$, 4) solve for $v_-(t,z_J)$, and 5) find $v_-(t + z/c)$, where we define z as increasing toward the load:

$$v_-(t,z_J) = v(t,z_J) - v_+(t,z_J) \equiv v_-\big(t + [z_J/c]\big) \qquad (8.1.18)$$

$$v_-(t + z/c) = v_-\big(t + [(z - z_J)/c], z_J\big) \qquad \text{(wave reflected by load)} \quad (8.1.19)$$

Equation (8.1.19) says $v_-(t + z/c)$ is simply the $v_-(t, z_J)$ given by (8.1.18), but delayed by $(z_J - z)/c$.

This procedure is best demonstrated by a simple example. Figure 8.1.5(a) illustrates a TEM line driven by a matched unit step voltage source and terminated with a capacitor C. This voltage step, reduced by a factor of two by the voltage divider, propagates toward the capacitor at velocity c, as illustrated in (b). The capacitor sees the Thevenin equivalent circuit illustrated in (c); it consists of Z_o in series with a Thevenin voltage source that is twice v_+, where $v_+(t,D)$ is a 0.5-volt step delayed by the propagation time D/c. Therefore $V_{Th} = u(t - D/c)$, as illustrated in Figure 8.1.5(c) and (d). The solution to the circuit problem of (c) is the junction voltage $v_J(t)$ plotted in (e); it rises exponentially toward its 1-volt asymptote with a time constant $\tau = Z_oC$ seconds.

To solve for $v_-(t,z_J)$ we subtract $v_+(t,z_J)$ from $v_J(t)$, as shown in (8.1.18) and illustrated in (f); this then yields $v_-(t + z/c)$ using (8.1.19). The total voltage $v(t_1,z)$ on the line at time $D/c < t_1 < 2D/c$ is plotted in (g) and is the sum of $v_+(t - z/c)$, which is 0.5 volts, and $v_-(t + z/c)$. The corresponding current $i(t_1,z)$ is plotted in (h) and equals Y_o times the difference between the forward and reverse voltage waves, as given by (8.1.5). When $v_-(t,z)$ arrives at the source, it can be treated just as such waves were treated in Section 8.1.3. In this case the source is matched, so there are no further reflections.
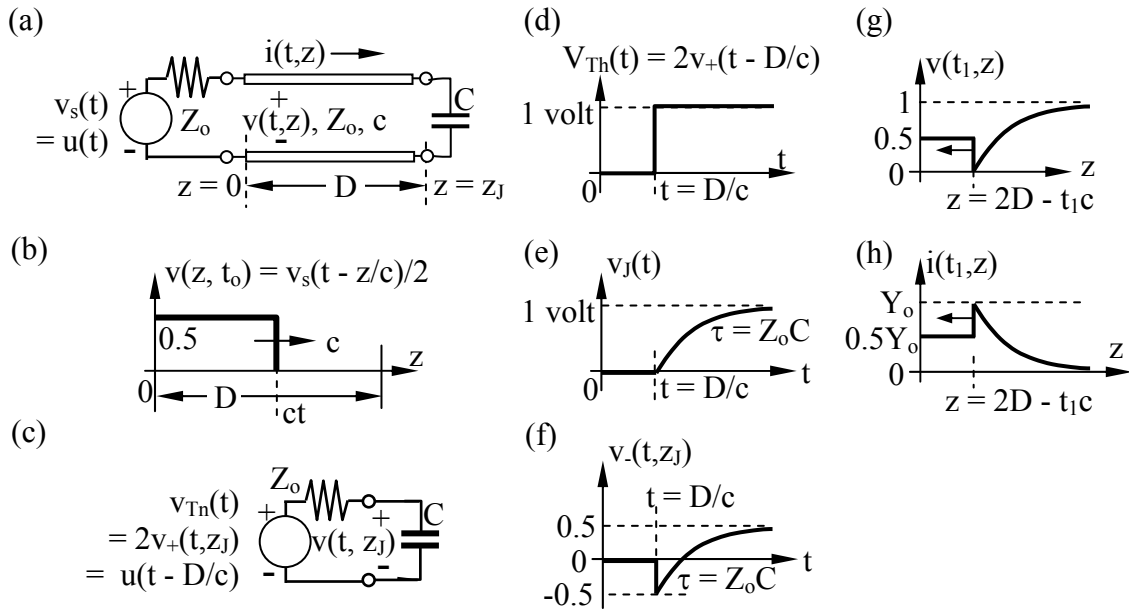
(a)

(b)

$v(z, t_o) = v_s(t - z/c)/2$

(c)

$v_{Tn}(t)$
$= 2v_+(t,z_J)$
$= u(t - D/c)$

(d)

$V_{Th}(t) = 2v_+(t - D/c)$

1 volt

$t = D/c$

(e)

$v_J(t)$

1 volt

$\tau = Z_oC$

$t = D/c$

(f)

$v_-(t,z_J)$

$t = D/c$

0.5
0
-0.5

$\tau = Z_oC$

(g)

$v(t_1,z)$

1
0.5
0

$z = 2D - t_1c$

(h)

$i(t_1,z)$

Y_o
$0.5Y_o$
0

$z = 2D - t_1c$

Figure 8.1.5 Transient voltages and currents on a capacitor-terminated TEM line.

Most digital circuits are non-linear, so this same technique is often used to determine the waveforms on longer TEM lines. Consider the circuit and ramp-pulse voltage source illustrated in Figure 8.1.6(a) and (b).

(a)

(b)

$v_s(t)$

1 volt

10^{-9} sec

(c)

$v(t, z_J) = v_J(t)$

$v_{Tn}(t)$
$= 2v_s(t-D/c)$

1 volt

(d)

$V_{Th}(t) =$
$2v_+(t - D/c)$

2

$D/c \quad 10^{-9}$ s

(e)

$v_J(t)$

1

$D/c \quad 10^{-9}$ s

(f)

$v_+(t - D/c)$

1

$D/c \quad 10^{-9}$ s

(g)

$v_-(t,z_J) = v_J - v_+$

0.5

$D/c \quad 10^{-9}$ s

(h)

$v(t_1,z)$

0.5
c

10^{-9} c

D

$2D - ct_1$

(i)

$i(t_1,z)$

0

$D \quad z$

c

-0.5Y_o

10^{-9} c
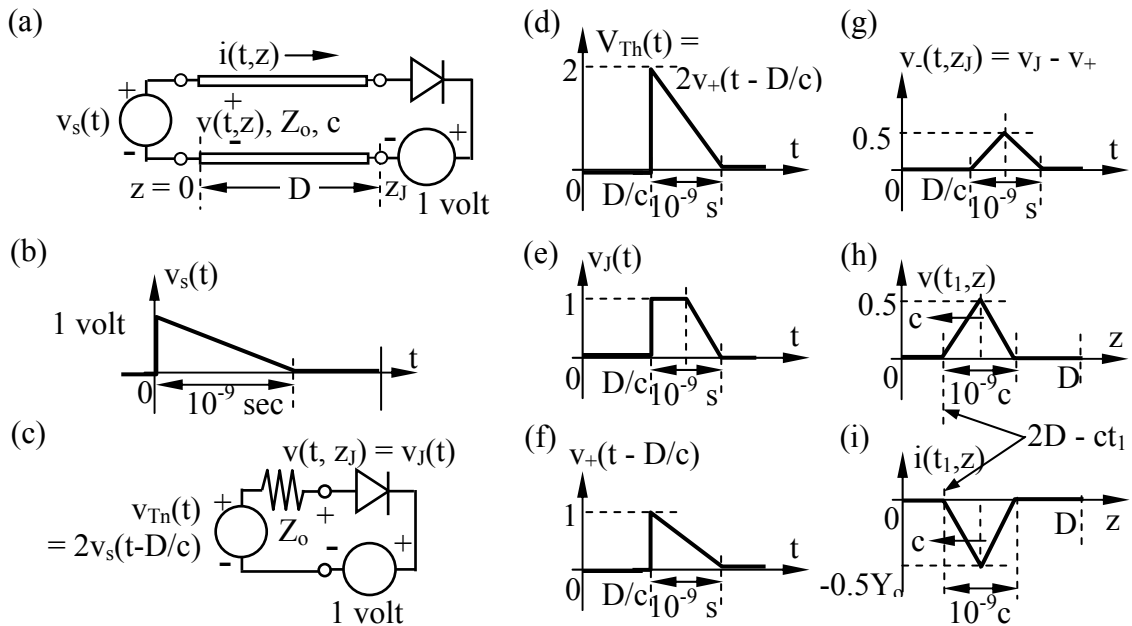
Figure 8.1.6 Transient TEM waveforms produced by reflection from a non-linear load.

In this case there is no source resistance (an arbitrary choice), so the full value of the source voltage appears across the TEM line. Part (c) shows the equivalent circuit of the transmission

line driving the load, which consists of a back-biased diode. The Thevenin voltage $V_{Th}(t,z_J) = 2v_+(t,z_J)$ is plotted in (d), the resulting junction voltage $v_J(t)$ is plotted in (e), $v_+(t,z_J)$ is plotted in (f), and $v_-(t,z_J) = v_J(t,z_J) - v_+(t,z_J)$ is plotted in (g). The line voltage and currents at $D/c < t_1 < 2D/c$ are plotted in (h) and (i), respectively. Note that these reflected waveforms do not resemble the incident waveform.

Example 8.1C

If the circuit illustrated in Figure 8.1.5(a) were terminated by L instead of C, what would be $v(t, D)$, $v_-(t, D)$, and $v(t,0)$?

Solution: Figures 8.1.5(a–d) still apply, except that L replaces C. The one-volt Thevenin step voltage at $z = D$ in series with the Thevenin line impedance Z_o yields a voltage $v(t, D)$ across the inductor of $u(t - D/c)e^{-tL/R}$. Therefore $v_-(t, D) = v(t, D) - v_+(t, D) = u(t - D/c)e^{-(t - D/c)L/R} - 0.5u(t - D/c)$ and, since there are no further reflections at the matched load at $z = 0$, it follows that $v(t,0) = 0.5u(t) + u(t - 2D/c)e^{-(t - 2D/c)L/R} - 0.5u(t - 2D/c)$.

8.1.5   Initial conditions and transient creation

Often transmission lines have an initial voltage and current that is interrupted in some way, producing transients. For example, a charged TEM line at rest may have a switch thrown at one end that suddenly connects it to a load, or disconnects it; such a switch could be located in the middle of a line too, either in series or parallel. The solution method has two main steps: 1) determine $v_+(t,z)$ and $v_-(t,z)$ at $t = 0_-$ before the change occurs, and 2) solve for the subsequent behavior of the forward and backward moving waves for the given network configuration.
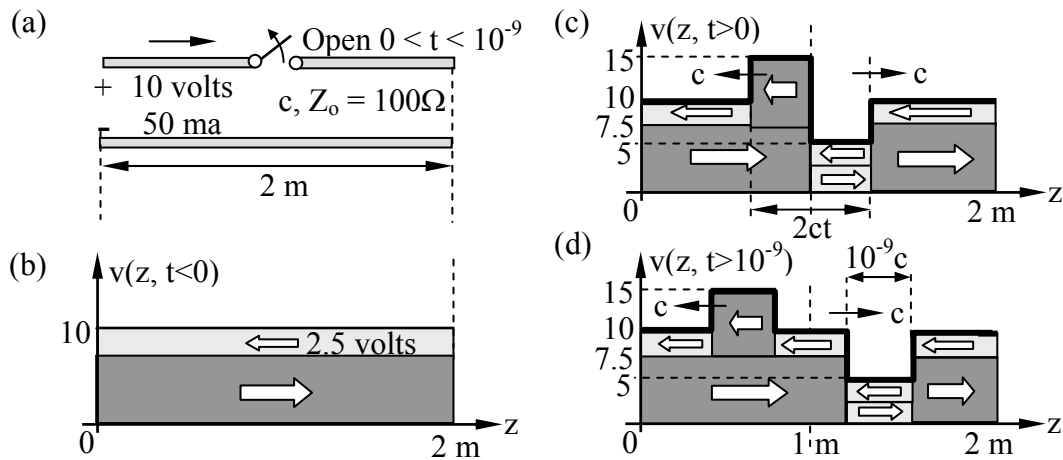

Figure 8.1.7   Transients induced by momentarily open-circuited active TEM line.

The simple example of Figure 8.1.7 illustrates the method. Assume an air-filled 2-meter long 100-ohm TEM line is feeding a 200-ohm load R with I = 50 milliamperes, when suddenly at t = 0 the line is open-circuited at z = 1 meters for 10^{-9} seconds, after which it returns to normal. What are the voltage and current on the line as a result of this temporary event?

Using the method suggested above, we first solve for the forward and backward waves prior to t = 0; the current i(t<0, z) is given as I = 50 milliamperes, and the voltage v(t<0, z) = IR is 0.05 × 200 ohms = 10 volts. Note that in steady state Z_o does not affect v(t<0, z). We know from (8.1.4) and (8.1.5) that:

$$v(z,t) = v_+(z-ct) + v_-(z+ct) \tag{8.1.20}$$

$$i(z,t) = Y_o\left[v_+(z-ct) - v_-(z+ct)\right] \tag{8.1.21}$$

Solving these two equations for v_+ and v_- yields:

$$v_+(z-ct) = \left[v(z,t) + Z_o i(z,t)\right]/2 = [10+5]/2 = 7.5 \text{ volts} \tag{8.1.22}$$

$$v_-(z+ct) = \left[v(z,t) - Z_o i(z,t)\right]/2 = [10-5]/2 = 2.5 \text{ volts} \tag{8.1.23}$$

These two voltages are shown in Figure 8.1.7(b), 7.5 volts for the forward wave and 2.5 volts for the reflected wave; this is consistent with the given 50-ma current.

When the switch opens at t = 0 for 10^{-9} seconds, it momentarily interrupts both v_+ and v_-, which see an open circuit at the switch and Γ = +1. Therefore in (c) we see 7.5 volts reflected back to the left from the switch, and 2.5 volts reflected back toward the right. At distances closer to the switch than ct [m] we therefore see 15 volts to the left and 5 volts to the right; this zone is propagating outward at velocity c. When the switch closes again, these mid-line reflections cease and the voltages and currents return to normal as the two transient pulses of 15 and 5 volts continue to propagate toward the two ends of the line, as shown in (d), where they might be reflected further.

The currents associated with Figure 8.1.7(d) can easily be surmised using (8.1.21). The effects of the switch are only felt for that brief 10^{-9}-second interval, and otherwise the current on the line is the original 50 ma. In the brief interval when the switch was open the current was forced to zero, and so zero-current pulses of duration 10^{-9} seconds propagate away from the switch in both directions.

Example 8.1D
A 100-ohm air-filled TEM line of length D is feeding 1 ampere to a 50-ohm load when it is momentarily short-circuited in its middle for a time T < D/2c. What are $v_+(z - ct)$ and $v_-(z + ct)$ prior to the short circuit, and during it?

Solution: For t < 0, $\Gamma = v_-(D - ct)/v_+(D + ct) = (Z_n - 1)/(Z_n + 1) = -0.5/1.5 = -1/3$ where Z_n = 50/100. Since the line voltage v(z,t) equals the current i times the load resistance (v = 50 volts), it follows that $v_+ + v_- = 2v_+/3 = 50$, and therefore $v_+ = v_+(z - ct) = $ 75 volts, and $v_-(z + ct) = -25$. During the short circuit the voltage within a distance d = ct of the short is altered. On the source side the short circuit reflects $v_- = -v_+ = -75$,

- 240 -

so the total voltage $(v_+ + v_-)$ within ct meters of the short circuit is zero, and on the load side $v_+ = -v_- = 25$ is reflected, so the total voltage is again zero. The currents left and right of the short are different, however, because the original $v_+ \neq v_-$, and $i_+ = v_+/Z_o$. Therefore, on the source side near the short circuit, $i = (v_+ - \Gamma v_+)/Z_o = 2v_+/Z_o = 2 \times 75/50 = 3$ [A]. On the load side near the short circuit, $I = -2 \times 25/50 = -1$ [A].

8.2   Limits posed by devices and wires

8.2.1   Introduction to device models

Most devices combine conducting elements with semiconductors, insulators, and air in a complex structure that stores, switches, or transforms energy at rates limited by characteristic time constants governed by Maxwell's equations and kinematics. For example, in *vacuum tubes* electrons are boiled into vacuum by a hot negatively charged *cathode* with small fractions of an electron volt of energy[43]. Such tubes switch state only so fast as the free electrons can cross the vacuum toward the positively charged *anode*, and only as fast as permitted by the RL or RC circuit time constants that control the voltages accelerating or retarding the electrons. The same physical limits also apply to most semiconductor devices, as suggested in Section 8.2.4, although sometimes quantum effects introduce non-classical behavior, as illustrated in Section 12.3.1 for laser devices.

Design of vacuum tubes for use above ~100 MHz was difficult because high voltages and very small dimensions were required to shorten the electron transit time to fractions of a radio frequency (RF) cycle. The wires connecting the cathode, anode, and any grids to external circuits also contributed inductance that limited speed. Trade-offs were required. For example, as the cathode-anode gap was diminished to shorten electron transit times, the capacitance C between the cathode and anode increased together with delays associated with their RC time constant τ_{RC}. Exactly the same physical issues of gap length, capacitance, and τ_{RC} arise in most semiconductor devices. The kinematics of electrons in vacuum was discussed in Sections 5.1.2–3, and the behavior of simple RL and RC circuits was discussed in Section 3.5.1.

8.2.2   Semiconductor device models

One simple example illustrates typical sources of lag in semiconductor devices. Both pnp and npn *transistors* are composed of p-n junctions that contribute device-related delay.   Field-effect transistors exhibit related lags.   Figures 8.2.1(a) and (b) present a DC p-n junction i-v characteristic and a circuit model that exhibits approximately correct delay characteristics for the case where low-loss metal wires are used for interconnecting devices.

---

[43] A thermal energy E_o of one electron volt ($\cong 1.6 \times 10^{-19}$ [J]) corresponds to a temperature T of ~11,600K, where $E_o = kT$ and k is Boltzmann's constant: $k \cong 1.38 \times 10^{-23}$ [J/^{o}K]. Thus a red-hot cathode at ~1000K would boil off free electrons with thermal energies of ~0.1 e.v.
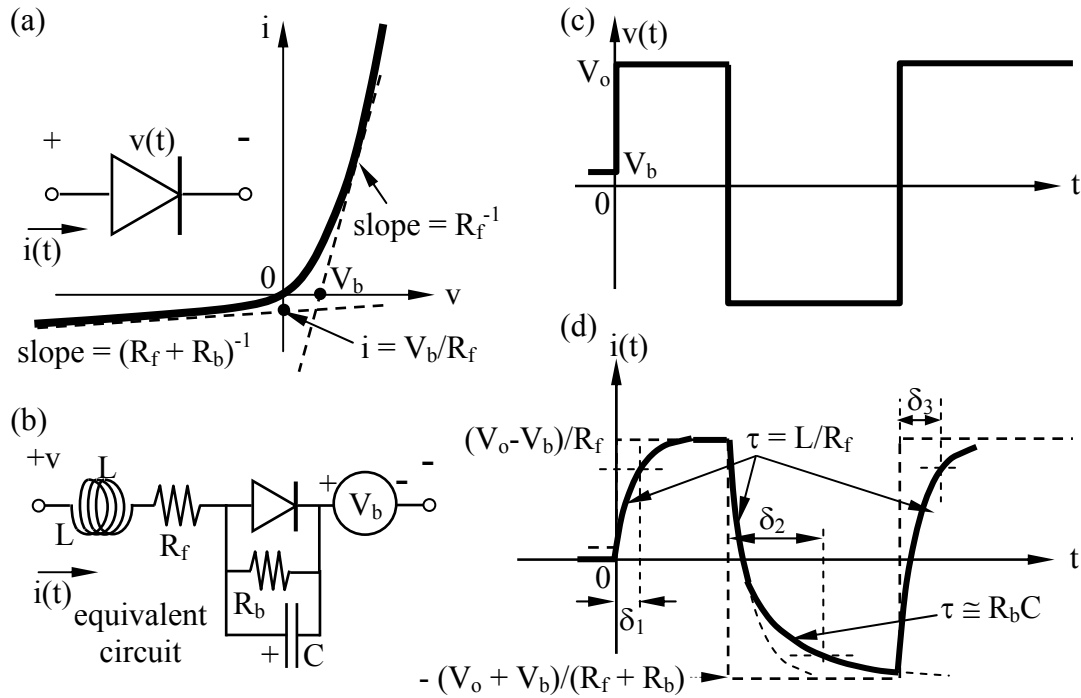
Figure 8.2.1   Circuit model for switching delays at p-n junctions.

When the ideal diode is forward biased the forward-bias resistance R_f determines the slope of the i-v characteristic. When it is back-biased beyond $\sim V_b$ the ideal diode becomes an open circuit so the junction capacitance C becomes important and the back-bias resistance $R_f + R_b \cong R_b$ determines the slope. C arises because of the charge-free depletion region that exists in back-biased diodes, an explanation of which is given in Section 8.2.4. C decreases as the back-bias voltage increases because the gap width d increases and $C \cong \varepsilon A/d$ (3.1.10). The bias voltage V_b in the equivalent circuit is related to the band gap between the valence and conduction bands in the semiconductor, and is ~1 volt for silicon (see Section 8.2.4 for more discussion). The inductance L arises primarily from wires leading externally, and is discussed further in Section 8.2.3 for printed circuits, and estimated in Section 3.3.2 for isolated wires (3.3.17).

Figure 8.2.1(c) represents a typical test voltage across a p-n junction applied as suggested in Figure 8.2.1(a). It begins at that bias voltage $v(t) = V_b$ (~1 volt in silicon), for which the equivalent circuit in Figure 8.2.1(b) conducts no current and the capacitor is discharged. At $t = 0_+$, $v(t) \rightarrow V_o$ and the current i(t) increases toward $(V_o - V_b)/R_f$ with zero incremental resistance offered by the diode and voltage source, so the time constant τ is L/R_f seconds (3.5.10), as illustrated in Figure 8.2.1(d). The capacitor remains uncharged. Section 3.5.1 discusses time constants for simple circuits. As a result of diversion of energy into the inductor, the current i does not reach levels sufficient to trigger the next circuit element until $t \cong \delta_1$, which is the lag time.

This time constant δ_1 can be easily estimated. For example, a p-n junction might be attached to wires of length D = 0.001 [m] and radius $r_o = 10^{-6}$ [m], and have a forward-biased resistance R_f of ~1 ohm. In this case (3.3.17) yields the wire inductance:

$$L \cong \left(\mu_o D/16\pi\right)\ln\left(D/r_o\right) \cong 1.7\times10^{-10} \ [\text{Henries}] \tag{8.2.1}$$

Thus $\delta_1 \cong \tau = L/R = 1.7\times10^{-10}$ seconds, so the diode might handle a maximum frequency of $\sim R/2\pi L$ Hz, or ~ 1 GHz. More conservatively the diode might be used at clock frequencies below ~ 0.2 GHz. Modern computers employ shorter wires and smaller R_f in order to work faster. The circuit model in Figure 8.2.1(b) does not include the capacitance between the wire and the substrate because it is negligibly small here relative to the effects of L.

When the test voltage v(t) then goes negative, the ideal diode in Figure 8.2.1(b) continues to conduct until the current through the inductor decays to zero with the same L/R_f time constant. The current then begins charging C (as the depletion layer is cleared of charge) with a time constant $\sim R_b C$ that delays the current response for a total of $\sim \delta_2$ seconds. Note that for illustrative purposes the current scale for negative i(t) in (d) has been expanded by a very large factor (R_b/R_f) relative to the scale for positive i(t).

When the test voltage then returns to V_o from its strong negative value, it must first discharge C (re-populate the depletion layer with charge) before the ideal diode in Figure 8.2.1(c) closes, introducing a time constant of $\sim R_f C$ that we can estimate. If the capacitance C corresponds to a depletion layer of thickness $d \cong 10^{-7}$, area $A \cong 10^{-11} \ [\text{m}^2]$, and permittivity $\varepsilon \cong \sim 10\varepsilon_o$, then (3.1.10) yields $C = \varepsilon A/d \cong 10\times8.8\times10^{-12}\times10^{-11}/10^{-7} \cong 9\times10^{-15} \ [\text{F}]$. This yields $R_f C \cong 10^{-14} \ [\text{s}] << L/R_f$, so L/R_f would dominate the entire transition, resulting in a total lag of $\delta_3 \cong \delta_1$ seconds. In reality i(t) in this RLC circuit would ring at $\omega \cong (LC)^{-0.5}$ radians per second as i(t) and the ringing decay toward the asymptote $i \cong (V_o - V_b)/R_f$.

In most bi-polar transistor circuits using metal wires it is L/R_f that controls the maximum clock speed for the system, which is limited by the slowest junction and the most inductive wires in the entire integrated circuit. In MOS integrated circuits, however, the resistivity of the polysilicon or diffusion layers used for conductors is sufficiently high that the wire inductance is often no longer controlling, as discussed in Section 8.3.1. Wire inductance is most easily reduced by using shorter wider wires, which also reduces wire resistance. Longer paths can be accommodated by using matched TEM lines, as discussed in Section 8.1.

8.2.3   Quasistatic wire models

The lag time for the p-n junction of Figure 8.2.1 was dominated by L/R_f, where L originated in the wires connected to the junction. The effects of depletion layer capacitance C were negligible in comparison for the assumed device parameters. In this section we examine the effects of wire capacitance and cross-section in limiting clock or signal frequencies.

In most integrated circuits the wires are planar and deposited on top of an insulating layer located over a conducting ground plane, as suggested in Figure 8.2.2.
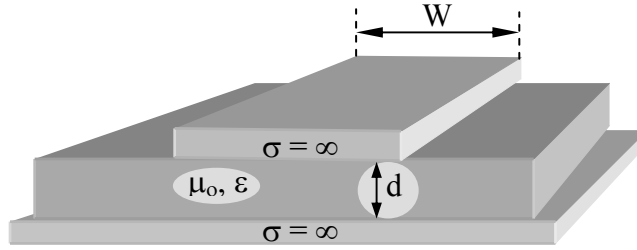
Figure 8.2.2   Idealized model for printed or integrated circuit wire.

The capacitance and inductance per unit length are C' and L', respectively, which follow from (3.1.10) and (3.2.6) under the assumption that fringing fields are negligible:

$$C' = \varepsilon W/d \ \left[F\,m^{-1}\right] \tag{8.2.2}$$

$$L' = \mu d/W \ \left[H\,m^{-1}\right] \tag{8.2.3}$$

$$L'C' = \mu\varepsilon \tag{8.2.4}$$

Printed circuit wires with width $W \cong 1$ mm and length $D \cong 3$ cm printed over dielectrics with thickness $d \cong 1$-mm and permittivity $\varepsilon \cong 4\varepsilon_o$ would add capacitance C and inductance L to the connected device, where:

$$C = \varepsilon WD/d = 9 \times 8.8 \times 10^{-12} \times 10^{-3} \times 0.03/10^{-3} = 2.4 \times 10^{-12} \ [F] \tag{8.2.5}$$

$$L = \mu dD/W = 1.2 \times 10^{-6} \times 10^{-3} \times 0.03/10^{-3} = 3.6 \times 10^{-8} \ [H] \tag{8.2.6}$$

These values combine with nominal one-ohm forward-bias resistances of p-n junctions to yield the time constants L/R = 3.6×10^{-8} seconds, and RC = 2.4×10^{-12} seconds. Again the limit is posed by inductance. Such printed circuit boards would be limited to frequencies $f \leq \sim 1/2\pi\tau \cong 4$ MHz.

The numbers cited here are not nearly so important as the notion that interconnections can strongly limit frequencies of operation and circuit utility. The quasistatic analysis above is valid because the physical dimensions here are much smaller than the shortest wavelength (at f = 4 MHz): $\lambda = c/f \cong 700$ meters in air or ~230 meters in a dielectric with $\varepsilon = 9\varepsilon_o$.

We have previously ignored wire resistance in comparison to the nominal one-ohm resistance of forward-biased p-n junctions. If the printed wires above are d =10 microns thick and have the conductivity σ of copper or aluminum, then their resistance is:

$$R = D/dW\sigma = 0.03/\left(10^{-5} \times 10^{-3} \times 5 \times 10^{7}\right) = 0.06 \ [ohms] \tag{8.2.7}$$

Since some semiconductor devices have forward resistances much less than this, wires are sometimes made thicker or wider to compensate. Wider printed wires also have lower inductance [see (8.2.6)]. Width and thickness are particularly important for power supply wires, which often carry large currents.

If (8.2.7) is modified to represent wires on integrated circuits where the dimensions in microns are length D = 100, thickness d = 0.1, and width W = 1, then R = 20 ohms and far exceeds typical forward p-n junction resistances. Limiting D to 20 microns while increasing d to 0.4 and W to 2 microns would lower R to 0.5 ohms. The resistivities of polysilicon and diffusion layers often used for conductors can be 1000 times larger, posing even greater challenges. Clearly wire resistance is another major constraint for IC circuit design as higher operating frequencies are sought.

Example 8.2A

A certain integrated circuit device having forward resistance R_f = 0.1 ohms is fed by a polysilicon conductor that is 0.2 microns wide and thick, 2 microns long, and supported 0.1 micron above the ground plane by a dielectric having $\varepsilon = 10\varepsilon_o$. The conductivity of the polysilicon wire is ~5×10^4 S m^{-1}. What limits the switching time constant τ of this device?

Solution: R, L, and C for the conductor can be found from (8.2.7), (8.2.6), and (8.2.5), respectively. $R = D/dW\sigma = 2\times10^{-6}/[(0.2\times10^{-7})^2 5\times10^4] = 1000$, $L \cong \mu dD/W = 1.26\times10^{-6}\times10^{-7}\times2\times10^{-6}/(0.2\times10^{-6}) = 1.26\times10^{-12}$ [Hy]. $C = \varepsilon WD/d = 8.85\times10^{-11}\times0.2\times10^{-6}\times2\times10^{-6}/10^{-7} = 3.54\times10^{-16}$ [F]. $RC \cong 3.54\times10^{-13}$ [s], $L/R = 1.26\times10^{-15}$ [s], and $(LC)^{0.5} = 2.11\times10^{-14}$ [seconds/radian], so RC limits the switching time. If metal substituted for polysilicon, then LC would pose the limit here.

8.2.4   Semiconductors and idealized p-n junctions

Among the most commonly used *semiconductors* are silicon (Si), germanium (Ge), gallium arsenide (GAs), and indium phosphide (InP). Semiconductors at low temperatures are insulators since all electrons are trapped in the immediate vicinity of their host atoms. The periodic atomic spacing of crystalline semiconductors permits electrons of sufficient energy to propagate freely without scattering, however. Diodes and transistors therefore exhibit conductivities that depend on the applied voltages and resulting electron energy distributions. The response times of these devices are determined by electron kinematics and the response times of the circuits and structures determining voltages and field strengths within the device.

The quantum mechanical explanation of such electron movement invokes the wave nature of electrons, which is governed by the Schroedinger wave equation (not explained here, although it is similar to the electromagnetic wave equation). The consequence is that semiconductors can be characterized by an *energy diagram* that shows possible electron energy states as a function of position in the z direction, as illustrated in Figure 8.2.3(a). At low temperatures all electrons occupy energy states in the lower *valence band*, corresponding to bound orbits around atoms. However a second *conduction band* of possible energy states occupied by freely moving electrons exists at higher energies separated from the valence band by an *energy gap* E_g that

varies with material, but is ~1 e.v. for silicon. For example, a photon with energy $hf = E_p > E_g$ can excite a bound electron in the valence band to a higher energy state in the conduction band where that electron can move freely and conduct electricity. In fact this photo-excitation mechanism is often used in semiconductor photo detectors to measure the intensity of light.
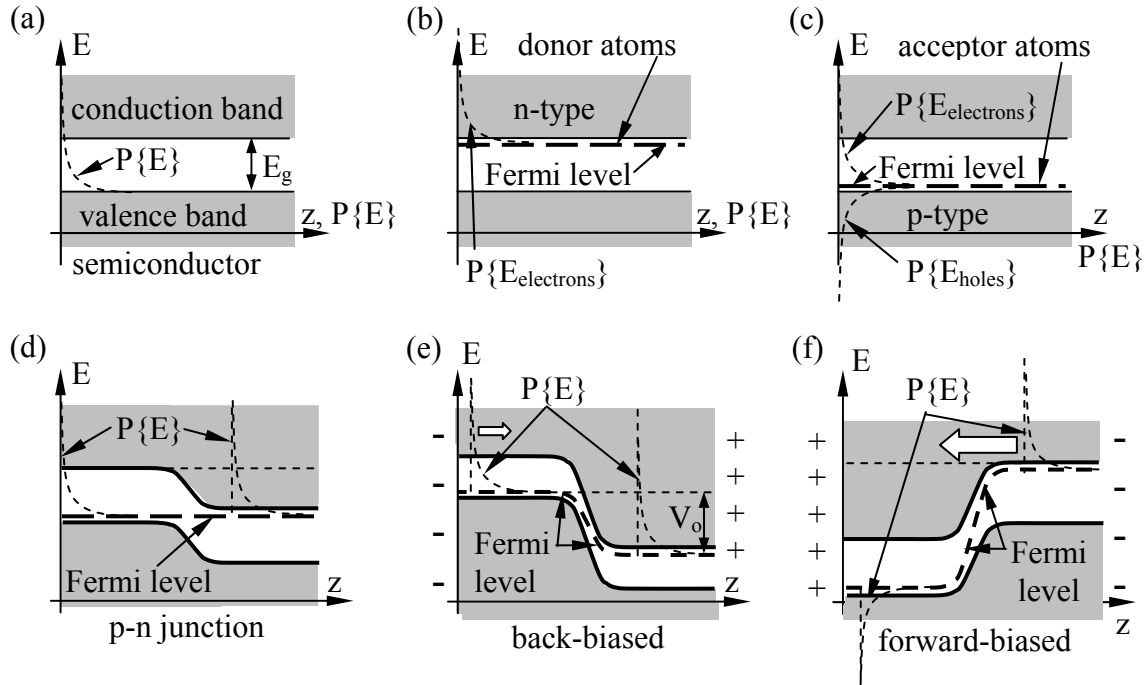

Figure 8.2.3   Energy diagrams for p and n semiconductors and p-n junctions.

The probability that an unbound electron in thermal equilibrium at temperature T has energy E is governed by the *Boltzmann distribution* $P\{E\} = e^{-E/kT}/kT$, where the *Boltzmann constant* k = 1.38×10^{-23} and one electron volt is 1.6×10^{-19} [J]. Therefore *thermal excitation* will randomly place a few free electrons in the conduction band since $P\{E > E_g\} > 0$. However this provides only extremely limited conductivity because a gap of ~1 e.v. corresponds to a temperature of E/k $\cong 11,600K$, much greater than room temperature.[44]

To boost the conductivity of semiconductors a small fraction of doping atoms are added that either easily release one electron (called *donor atoms*), or that easily capture an extra electron (*acceptor atoms*). These atoms assume energy levels that are just below the conduction band edge (donor atoms) or just above the valence band edge (acceptor atoms), as illustrated in Figure 8.2.3(b) and (c), respectively. These energy gaps are quite small so a significant fraction of the donor and acceptor atoms are typically ionized at room temperature.

The probability that an electron has sufficient energy to leap a gap E_a is the integral of the Boltzmann probability distribution from E_a to E→∞, and E_a is sufficiently small that this integral can approach unity for some dopant atoms. The base energy for the Boltzmann distribution is

---

[44] The Fermi level of an undoped semiconductor is midway between the valence and conduction bands, so ~one-half electron volt is actually sufficient to produce excitation, although far more electrons exist in the valence band itself.

the *Fermi level*; electrons fill the available energy levels starting with the lowest and ending with the highest being (on average) at the Fermi level. The Fermi level usually lies very close to the energy level associated with the donor or acceptor atoms, as illustrated in Figure 8.2.3(b–f). Since holes are positively charged, their exponential Boltzmann distribution appears inverted on the energy diagram, as illustrated in (c).

For each ionized donor atom there is an electron in the conduction band contributing to conductivity. For each negatively ionized acceptor atom there is a vacated positively charged "hole" left behind. An adjacent electron can easily jump to this hole, effectively moving the hole location to the space vacated by the jumping electron; in this fashion holes can migrate rapidly and provide nearly the same conductivity as electrons in the conduction band. Semiconductors doped with donor atoms so that free electrons dominate the conductivity are *n-type semiconductors* (negative carriers dominate), while holes dominate the conductivity of *p-type semiconductors* doped with acceptors (positive carriers dominate). Some semiconductors are doped to produce both types of carriers. The conductivity of homogeneous semiconductors is proportional to the number of charge carriers, which is controlled primarily by doping density and temperature.

If a p-n junction is short-circuited, the Fermi level is the same throughout as shown in Figure 8.2.3(d). Therefore the tails of the Boltzmann distributions on both sides of the junction are based at the same energy, so there is no net flow of current through the circuit.

When the junction is back-biased by V_o volts as illustrated in (e), the Fermi level is depressed correspondingly. The dominant current comes from the tail of the Boltzmann distribution of electrons in the conduction band of the p-type semiconductor; these few electrons will be pulled to the positive terminal and are indicated by the small white arrow. Some holes thermally created in the n-type valence band may also contribute slightly. Because the carriers come from the high tail of the Boltzmann thermal distribution, the reverse current in a p-n junction is strongly dependent upon temperature and can be used as a thermometer; it is not very dependent upon voltage once the voltage is sufficiently negative. When a p-n junction is back-biased, the electric field pulls back most low-energy free electrons into the n-type semiconductor, and pulls the holes into the p-type semiconductor, leaving a carrier-free layer, called a charge-depletion region, that acts like a capacitor. Larger values of V_o yield larger gaps and smaller capacitance.

When the junction is forward-biased by V_o volts, as illustrated in (f), the charge-free layer disappears and the current flow is dominated by the much greater fraction of electrons excited into the conduction band in the n-type semiconductor because almost all of them will be pulled by the applied electric field across the junction before recombining with a positive ion. This flow of electrons (opposite to current flow) is indicated by the larger white arrow, and is proportional to that fraction of P{E} that lies beyond the small energy gap separating the Fermi level and the lower edge of the conduction band. Holes in the valence band can also contribute significantly to this current. The integral over energy of the exponential probability distribution P{E} above threshold E_g - v is another exponential for $0 < v < E_g$, which is proportional to the population of conducting electrons, and which approximates the i(v) relation for a p-n junction illustrated in Figure 8.2.1(a) for v > 0.

Transistors are semiconductor devices configured so that the number of carriers (electrons plus holes) available in a junction to support conductivity is controlled 1) by the number injected into the junction by a p-n interface biased so as to inject the desired number (e.g., as is done in p-n-p or n-p-n transistors), or 2) by the carriers present that have not been pulled to one side or trapped by electric fields (e.g., field-effect transistors). In general, small bias currents and voltages can thereby control the current flowing across much larger voltage gaps with power amplification factors of 100 or more. Although the range of device designs is very large, most can be understood semi-classically as suggested above, without the full quantum mechanical descriptions needed for precise characterization.

The response time of p-n junctions and transistors is usually determined by either the RC, RL, or LC time constants that limit the rise and fall times of voltages and currents applied to the device terminals, or by the field relaxation time ε/σ (4.3.3) of the semiconductor material within the device itself. In extremely fast devices the response time sometimes is $\tau \cong D/v$, where D is the junction dimension [m] of interest, and v is the velocity of light ($v = [\mu\varepsilon]^{-0.5}$) or of the transiting electrons ($v = \int a \, dt$, $f = ma = eE$).

Although these physical models for semiconductor junctions are relatively primitive, they do approximately explain most phenomena.

Example 8.2B
What are the approximate temperature dependences of the currents flowing in forward- and reverse-biased p-n junctions?

Solution: If the bias voltage exceeds the gap voltage, and kT is large compared to the energy gap between the donor level and the conduction band, then essentially all donors are ionized and further temperature changes have little effect on forward biased p-n junctions; see Figure 8.2.3(f). The carrier concentration in reverse-biased diodes is proportional to $\int_{E_0+E_g}^{\infty} e^{-E/kT} \, dE$, and therefore to T [Figure 8.2.3(e)].

8.3   *Distortions due to loss and dispersion*

8.3.1   Lossy transmission lines

In most electronic systems transmission line loss is a concern because business strategy generally dictates reducing wire diameters and costs until such issues arise. For example, the polysilicon often used for conductors in integrated silicon devices has noticeable resistance.

The *TEM circuit model* of Figure 8.3.1 incorporates two types of loss. The series resistance R per meter arises from the finite conductivity of the wires, while the parallel conductance G per meter arises from leakage currents flowing between the wires through the medium separating them.
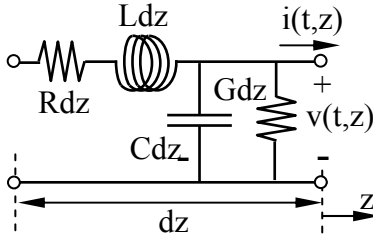
Figure 8.3.1   Distributed circuit model for lossy TEM transmission lines.

When these lossy elements are included, we obtain the *telegraphers' equations*:

$$dv/dz = -Ri - L\,di/dt \qquad \text{(telegraphers' equation)} \qquad (8.3.1)$$

$$di/dz = -Gv - C\,dv/dt \qquad \text{(telegraphers' equation)} \qquad (8.3.2)$$

If the wires are resistive, then current flowing through them introduces longitudinal electric fields E_z, violating the TEM assumption: $E_z = H_z = 0$. Since rigorous solution of Maxwell's equations for the non-TEM case is challenging, the telegraphers' equations are often used instead if the loss is modest. The same problem does not arise with G because it does not violate the TEM assumption, as shown in Section 7.1.3. Since propagation in such *lossy TEM lines* is frequency dependent, the telegraphers' equations (8.3.1–2) and their solutions are generally expressed using complex notation[45]:

$$d\underline{V}(z)/dz = -(R + j\omega L)\underline{I}(z) \qquad \text{(telegraphers' equation)} \qquad (8.3.3)$$

$$d\underline{I}(z)/dz = -(G + j\omega C)\underline{V}(z) \qquad \text{(telegraphers' equation)} \qquad (8.3.4)$$

Differentiating (8.3.3) with respect to z, and substituting (8.3.4) for $d\underline{I}(z)/dz$ yields the wave equation for lossy TEM lines, where the sign of k^2 is chosen so that k is real, consistent with the lossless solutions discussed earlier in Section 7.1.2:

$$d^2\underline{V}(z)/dz^2 = (R + j\omega L)(G + j\omega C)\underline{V}(z) = -\underline{k}^2\underline{V}(z) \qquad \text{(wave equation)} \qquad (8.3.5)$$

$$\underline{k} = \left[-(R + j\omega L)(G + j\omega C)\right]^{0.5} = k' - jk'' \qquad \text{(TEM propagation constant)} \qquad (8.3.6)$$

Since the second derivative of $\underline{V}(z)$ equals a constant times itself, it must be expressible as the sum of exponentials that have this property:

$$\underline{V}(z) = \underline{V}_+ e^{-j\underline{k}z} + \underline{V}_- e^{+j\underline{k}z} \qquad \text{(TEM voltage solution)} \qquad (8.3.7)$$

---

[45] Complex notation is discussed in Section 2.3.2 and Appendix B. In general, $v(t) = R_e\{\underline{V}e^{j\omega t}\}$, where $R_e\{\bullet\ e^{j\omega t}\}$ is omitted from equations.

Differentiating (8.3.7) with respect to z and substituting the result in (8.3.3) yields both $\underline{I}(z)$ and $\underline{Y}_o$:

$$\underline{I}(z) = \underline{Y}_o\left(\underline{V}_+e^{-j\underline{k}z} - \underline{V}_-e^{+j\underline{k}z}\right) \qquad \text{(TEM current solution)} \qquad (8.3.8)$$

$$\underline{Z}_0 = \frac{1}{\underline{Y}_o} = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \qquad \text{(characteristic impedance)} \qquad (8.3.9)$$

When R = G = 0, (8.3.9) reduces to the well known result $Z_o = (L/C)^{0.5}$.

Thus two new properties emerge when TEM lines are dissipative: 1) because $\underline{k}$ is complex and a non-linear function of frequency, waves are attenuated and dispersed as they propagate in a frequency-dependent manner, and 2) $\underline{Z}_o$ is complex and frequency dependent. Both k' and k" (8.3.6) are functions of frequency, so signals propagating on lossy lines change shape, partly because different frequency components propagate and decay differently. The resulting attenuation and dispersion are discussed in Sections 8.3.1 and 8.3.2, respectively. Reflections are affected at junctions by losses, and also are attenuated with distance so the impedance of a lossy line $\underline{Z}(z) \rightarrow \underline{Z}_o$ regardless of load as $\underline{V}_-(z)$ becomes negligible. Reflections by junctions involving lossy lines are simply analyzed by replacing Z_o by a complex impedance $\underline{Z}_o$ in the expressions developed in Section 7.2 for lossless lines.

Waves propagating only in the +z direction obey (8.3.7), which becomes:

$$\underline{V}(z) = \underline{V}_+e^{-j\underline{k}z} = \underline{V}_+e^{-jk'z}e^{-k"z} \qquad \text{(decaying propagating wave)} \qquad (8.3.10)$$

One combination of R, L, C, and G is particularly interesting because it results in zero dispersion and a frequency-independent decay that does not distort waveforms. We may discover this combination by evaluating $\underline{k}$ using (8.3.6):

$$\underline{k} = \left[-(R + j\omega L)(G + j\omega C)\right]^{0.5} = \omega\left\{LC\left[1 - j(R/\omega L)\right]\left[1 - j(G/\omega C)\right]\right\}^{0.5} \qquad (8.3.11)$$

It follows from (8.3.11) that if R/L = G/C, then the phase velocity ($v_p = \omega/k' = [LC]^{-0.5}$) and the decay rate (k" = $R[C/L]^{0.5}$) are both frequency independent:

$$\underline{k} = (LC)^{0.5}(\omega - jR/L) = k' - jk" \qquad \text{(distortionless line)} \qquad (8.3.12)$$

The ability to avoid signal distortion due to frequency-dependent absorption was first exploited by telephone companies who added small inductors periodically in series with their longer phone lines in order to reduce R/L so that it balanced G/C; the result was called a

distortionless line, and the coils are called *Pupin coils* after their inventor[46]. The consequences of dispersion are explored in Section 8.3.2.

Another limit is sometimes of interest when the effects of R dominate those of ωL. This occurs, for example, in resistive polysilicon or diffusion lines in integrated circuits, which may be approximately modeled by eliminating L and G from Figure 8.3.1. Then $\underline{k}$ (8.3.11) becomes:

$$\underline{k} \cong (-j\omega RC)^{0.5} = (\omega RC/2)^{0.5} - j(\omega RC/2)^{0.5} = k' - jk'' \tag{8.3.13}$$

The square root of -j was chosen to correspond to a decaying wave rather than to exponential growth. The phase and group velocities for this line are the same:

$$v_p = \omega/k' = (2\omega/RC)^{0.5} \ [\text{m s}^{-1}] \tag{8.3.14}$$

$$v_g = (\partial k'/\partial \omega)^{-1} = 2(\omega/RC)^{0.5} \ [\text{m s}^{-1}] \tag{8.3.15}$$

Although it is not easy to relate these frequency-dependent velocities to delays in digital circuits, they demonstrate that such delays exist and express their dependence on R and C. That is, larger line time constants RC lower pulse velocities and increase delays. Such lines are best used when they are short compared to the shortest wavelength of interest, $D < \lambda = v_p/f_{max} = 2\pi(2/rc\omega_{max})^{0.5}$. In polysilicon lines $\lambda_{min} \cong 1$ mm for $\omega_{max} = 10^{10}$. The response to arbitrary waveform excitation can be computed by: 1) Fourier transforming the signal, 2) propagating each frequency component as dictated by (8.3.13), and then 3) reconstructing the signal at the new location with an inverse Fourier transform. Typical values for R and C in metal, *polysilicon*, and diffusion lines are presented in Table 8.3.1, and correspond to velocities much less than c. The costs of these three options for forming conductors are unequal and must also be considered when designing fast integrated circuits.

Table 8.3.1 Resistance and capacitance per meter for typical integrated circuit lines.

Parameter	Metal	Polysilicon	Diffusion
R [$M\Omega$ m^{-1}]	0.06	50	50
C [nF m^{-1}]	0.1	0.2	1

Provided R is not so large compared to ωL that the TEM approximation is invalid because of strong longitudinal electric fields, then the power dissipated is:

$$P_d = (R|\underline{I}|^2 + G|\underline{V}|^2)/2 \ [\text{W m}^{-1}] \tag{8.3.16}$$

---

[46] Pupin coils had to be inserted at least every $\lambda/10$ meters in order to avoid additional distortions, but the shortest λ for telephone voice signals is ~$c/f = 3\times10^8/3000 = 100$ km.

8.3.2   Dispersive transmission lines

Different frequency components propagate at different velocities on *dispersive transmission lines*. The nature and consequences of dispersion are discussed further in Section 9.5.2. Consider first a square-wave computer clock pulse at F Hz propagating along a dispersive TEM line. The Fourier transform of this signal has its fundamental at F Hz, with odd harmonics at 3F, 5F, etc., each of which has its own phase velocity, as suggested in Figure 8.3.2.
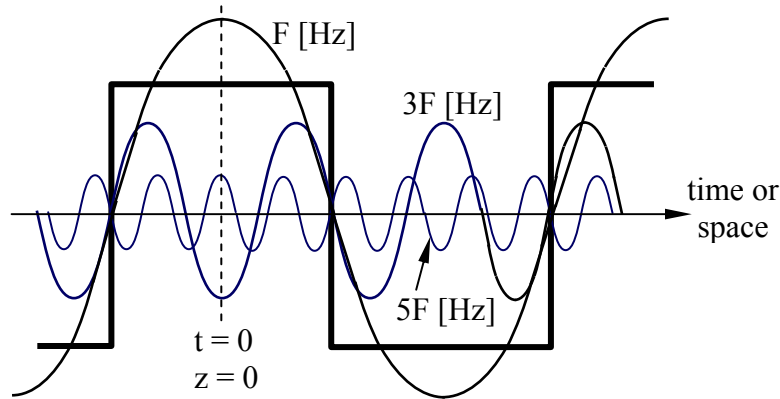

Figure 8.3.2   Square wave and its constituent sinusoids.

Significant pulse distortion occurs if a strong harmonic is shifted as much as ~90° relative to the fundamental. To determine the relative phase shift between fundamental and harmonic we can first multiply the difference in phase velocity at F and 3F, e.g., v_{pF} - v_{p3F}, by the propagation time T of interest. This yields the spatial offset between these two harmonics, which we might limit to $\lambda/4$ for 3F. That is, we might expect significant distortion over a transmission line of length D = v_pT meters if:

$$\left(v_{pF} - v_{p3F}\right)T > \sim \lambda_{3F}/4 = v_{p3F}/(4 \times 3F) \tag{8.3.17}$$

There is a similar limit to the propagation distance of narrowband pulse signals before *waveform distortion* becomes unacceptable. Digital communications systems commonly use narrowband pulses s(t) for both wireless and cable signaling. For example, the square wave in Figure 8.3.2 could also represent the amplitude envelope $A(t) = \Sigma_i\, a_i\, \cos\omega_i t$ of an underlying sinusoid $\cos\omega_o t$, where $\omega_o \gg \omega_{i>0}$ and together they occupy a narrow bandwidth. That is:

$$s(t) = \left(\cos\omega_o t\right)\sum_{i=1} a_i \cos\omega_i t = 0.5\sum_i a_i \left\{\cos\left[\left(\omega_o + \omega_i\right)t + \cos\left[\left(\omega_o - \omega_i\right)t\right]\right]\right\} \tag{8.3.18}$$

Since each frequency $\omega_o \pm \omega_i$ propagates at a slightly different phase velocity, a narrowband pulse will also distort when a strong harmonic is $\sim\lambda/4$ out of phase relative to the original wave envelope, which is much larger than $\lambda = 2\pi c/\omega_o$ for narrowband signals. Some applications are more sensitive to dispersive distortion than others; for example, distorted digital signals can be

generally be regenerated distortion free, while analog signals require inverse distortion, which is often uneconomic.

Distortion of narrowband signals is usually computed in terms of the *group velocity* v_g, which is the velocity of propagation for the waveform envelope and equals the velocity of energy or information, which can never exceed c, the velocity of light in vacuum. The sine wave that characterizes the average frequency of a narrowband pulse propagates at the *phase velocity* v_p, which can be greater or less than c. Narrowband pulse signals (e.g., digitally modulated sinusoids) distort when the accumulated difference Δ in the envelope propagation distances between the high- and low-frequency end of the signal spectrum differs by more than a small fraction of the minimum pulse width W[m] (e.g., the length of a zero or one). Since the difference in group velocity across the bandwidth B[Hz] is $(\partial v_g/\partial f)B$ [m/s], and the pulse travel time is D/v_g, where D is propagation distance, it follows that the difference in envelope propagation distance across the band is:

$$\Delta = \frac{\partial v_g}{\partial f}\frac{BD}{v_g} \ [m] \tag{8.3.19}$$

Since the minimum pulse width W is $\sim v_g/B$ [m], the requirement that D << W implies that the maximum distortion-free propagation distance D is:

$$D \ll \left(\frac{v_g}{B}\right)^2 \left(\frac{\partial v_g}{\partial f}\right)^{-1} \tag{8.3.20}$$

Group and phase velocity are discussed further in Section 9.5.2 and their effect on distortion is explored in Section 12.2.2.

Example 8.3A

Typical 50-ohm coaxial cables for home distribution of television and internet signals have series resistance $R \cong 0.02(f_{MHz}^{0.5})$ ohms m^{-1}. Assume $\varepsilon = 4\varepsilon_o$, $\mu = \mu_o$. How far can signals propagate before attenuating 60 dB?

Solution: Since conductivity $G \cong 0$, (8.3.11) says $\underline{k} = \omega[LC(1 - jR/\omega L)]^{0.5} = k' + jk''$. The imaginary part of $\underline{k}$ corresponds to exponential decay. For $R << \omega L$, $\underline{k} \cong \omega(LC)^{0.5}(1 - jR/\omega L)^{0.5}$, so $k'' = -(\omega RC)^{0.5}$. To find C we note the phase velocity $v = (\mu_o 4\varepsilon_o)^{-0.5} = (LC)^{-0.5} = c/2 \cong 1.5\times10^8$ [m s^{-1}], and $Z_o = 50 = (L/C)^{0.5}$. Therefore $C = 2/cZ_o = 2/(3\times10^8\times50) \cong 1.33\times10^{-10}$ [F]. Thus at 100 MHz, $k'' = -(\omega RC)^{0.5} = -(2\pi10^8\times0.2\times1.33\times10^{-10})^{0.5} \cong -0.017$. Since power decays as $e^{-2k''z}$, 60 dB corresponds to $e^{-2k''z} = 10^{-6}$, so $z = -\ln(10^{-6})/2k'' = 406$ meters. At 100 MHz the approximation $R << \omega L$ is quite valid. As coaxial cable systems boost data rates and their maximum frequency above 100-200 MHz, the increased attenuation requires amplifiers at intervals so short as to motivate switching to optical fibers that can propagate signals hundred of kilometers without amplification.

Chapter 9: Electromagnetic Waves

9.1 Waves at planar boundaries at normal incidence

9.1.1 Introduction

Chapter 9 treats the propagation of plane waves in vacuum and simple media, at planar boundaries, and in combinations confined between sets of planar boundaries, as in waveguides or cavity resonators. Chapter 10 then discusses how such waves can be generated and received by antennas and antenna arrays.

More specifically, Section 9.1 explains how plane waves are reflected from planar boundaries at normal incidence, and Section 9.2 treats reflection and refraction when the waves are incident at arbitrary angles. Section 9.3 then explains how linear combinations of such waves can satisfy all boundary conditions when they are confined within parallel plates or rectangular cylinders acting as waveguides. By adding planar boundaries at the ends of such waveguides, waves can be trapped at the resonant frequencies of the resulting cavity, as explained in Section 9.4. Sections 9.5 then treat waves in anisotropic, dispersive, and ionized media, respectively.

9.1.2 Introduction to boundary value problems

Section 2.2 showed how uniform planes waves could propagate in any direction with any polarization, and could be superimposed in any combination to yield a total electromagnetic field. The general electromagnetic *boundary value problem* treated in Sections 9.1–4 involves determining exactly which, if any, combination of waves matches any given set of *boundary conditions*, which are the relations between the electric and magnetic fields adjacent to both sides of each boundary. These boundaries can generally be both active and passive, the active boundaries usually being sources. Boundary conditions generally constrain $\overline{E}$ and/or $\overline{H}$ for all time on the boundary of the two- or three-dimensional region of interest.

The uniqueness theorem presented in Section 2.8 states that only one solution satisfies all Maxwell's equations if the boundary conditions are sufficient. Therefore we may solve boundary value problems simply by hypothesizing the correct combination of waves and testing it against Maxwell's equations. That is, we leave undetermined the numerical constants that characterize the chosen combination of waves, and then determine which values of those constraints satisfy Maxwell's equations. This strategy eases the challenge of hypothesizing the final answer directly. Moreover, symmetry and other considerations often suggest the nature of the wave combination required by the problem, thus reducing the numbers of unknown constants that must be determined.

The four basic steps for solving boundary value problems are:

1) Determine the natural behavior of each homogeneous section of the system in isolation (absent its boundaries).

2) Express this natural behavior as the superposition of waves characterized by unknown constants; symmetry and other considerations can minimize the number of waves required. Here our basic building blocks are usually uniform plane waves, but other more compact expansions are typically used if the symmetry of the problem permits, as illustrated in Section 4.5.2 for cylindrical and spherical geometries, Section 7.2.2 for TEM transmission lines, and Section 9.3.1 for waveguide modes.

3) Write equations for the boundary conditions that must be satisfied by these sets of superimposed waves, and then solve for the unknown constants.

4) Test the resulting solution against any of Maxwell's equations that have not already been imposed.

Variations of this four-step procedure can be used to solve almost any problem by replacing Maxwell's equations with their approximate equivalent for the given problem domain.

9.1.3   Reflection from perfect conductors

One of the simplest examples of a boundary value problem is that of a uniform plane wave in vacuum normally incident upon a planar perfect conductor at $z \geq 0$, as illustrated in Figure 9.1.1(a). Step 1 of the general boundary-problem solution method of Section 9.1.2 is simply to note that electromagnetic fields in the medium can be represented by superimposed uniform plane waves.
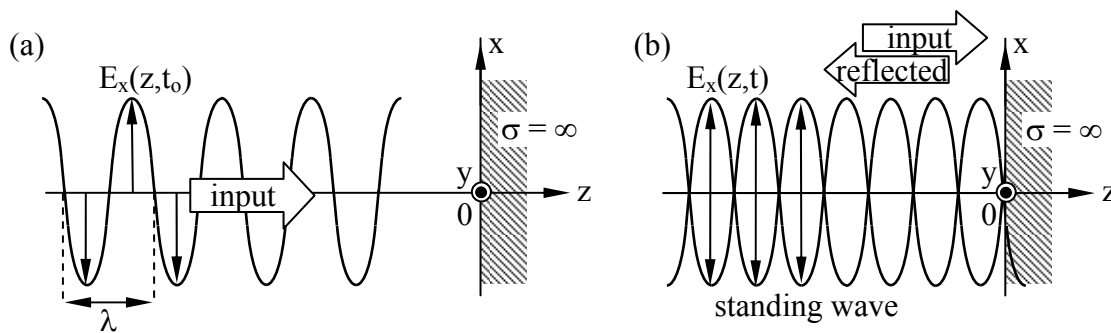
Figure 9.1.1   Plane wave at normal incidence reflecting from a perfect conductor.

For this incompletely defined example, the initial part of Step 2 of the method involves refinement of the problem definition by describing more explicitly the incident wave, for example:

$$\overline{E}(z,t) = \hat{x}E_o \cos(\omega t - kz) \ \left[Vm^{-1}\right] \tag{9.1.1}$$

where the wave number $k = 2\pi/\lambda = \omega/c = \omega(\mu_o\varepsilon_o)^{0.5}$, (2.3.24). The associated magnetic field (2.3.25) is:

$$\overline{H}(z,t) = \hat{y}(E_o/\eta_o)\cos(\omega t - kz)\left[Am^{-1}\right]$$ (9.1.2)

This unambiguously defines the source, and the boundary is similarly unambigous: $\sigma = \infty$ and therefore $\overline{E} = 0$ for $z \geq 0$. This more complete problem definition is sufficient to yield a unique solution. Often the first step in solving a problem is to ensure its definition is complete.

Since there can be no waves inside the perfect conductor, and since the source field alone does not satisfy the boundary condition $\overline{E}_{//} = 0$ at z = 0, one or more additional plane waves must be superimposed to yield a valid solution. In particular, we need to match the boundary condition $\overline{E}_{//} = 0$ at z = 0. This can be done by adding a single uniform plane wave propagating in the -z direction with an electric field that cancels the incident electric field at z = 0 for all time t. Thus we hypothesize that the total electric field is:

$$\overline{E}(z,t) = \hat{x}\left[E_o \cos(\omega t - kz) + E_1 \cos(\omega t + kz + \phi)\right]$$ (9.1.3)

where we have introduced the constants E_1 and ϕ.

In Step 3 of the method we must solve the equation (9.1.3) that characterizes the boundary value constraints:

$$\overline{E}(0,t) = \hat{x}\left[E_o \cos(\omega t - 0) + E_1 \cos(\omega t + 0 + \phi)\right] = 0$$ (9.1.4)

$$\therefore E_1 = -E_o \ , \quad \phi = 0$$ (9.1.5)

The result (9.1.5) yields the final trial solution:

$$\overline{E}(z,t) = \hat{x}E_o\left[\cos(\omega t - kz) - \cos(\omega t + kz)\right] = \hat{x}2E_o(\sin\omega t)\sin kz$$ (9.1.6)

$$\overline{H}(z,t) = \hat{y}E_o\left[\cos(\omega t - kz) + \cos(\omega t + kz)\right]/\eta_o = \hat{y}(2E_o/\eta_o)(\cos\omega t)\cos kz$$ (9.1.7)

Note that the sign of the reflected $\overline{H}$ wave is reversed from that of the reflected $\overline{E}$, consistent with the reversal of the Poynting vector for the reflected wave alone. We have used the identities:

$$\cos\alpha + \cos\beta = 2\cos\left[(\alpha + \beta)/2\right]\cos\left[(\alpha - \beta)/2\right]$$ (9.1.8)

$$\cos\alpha - \cos\beta = -2\sin\left[(\alpha + \beta)/2\right]\sin\left[(\alpha - \beta)/2\right]$$ (9.1.9)

Also note that $\overline{H}(z,t)$ is 90° out of phase with $\overline{E}(z,t)$ with respect to both time and space.

We also need a trial solution for z > 0. Inside the conductor $\overline{E} = \overline{H} = 0$, and boundary conditions (2.6.17) require a surface current:

$$\overline{J}_s = \hat{n} \times \overline{H} \left[Am^{-1} \right] \tag{9.1.10}$$

The fourth and final step of this problem-solving method is to test the full trial solution against all of Maxwell's equations. We know that our trial solution satisfies the wave equation in our source-free region because our solution is the superposition of waves that do; it therefore also satisfies Faraday's and Ampere's laws in a source-free region, as well as Gauss's laws. At the perfectly conducting boundary we require $\overline{E}_{//} = 0$ and $\overline{H}_\perp = 0$; these constraints are also satisfied by our trial solution, and therefore the problem is solved for the vacuum. Zero-value fields inside the conductor satisfy all Maxwell's equations, and the surface current $\overline{J}_s$ (9.1.10) satisfies the final boundary condition.

The nature of this solution is interesting. Note that the total electric field is zero not only at the surface of the conductor, but also at a series of null planes parallel to the conductor and spaced at intervals Δ along the z axis such that $kz_{nulls} = -n\pi$, where n = 0, 1, 2, ... That is, the null spacing $\Delta = \pi/k = \lambda/2$, where λ is the wavelength. On the other hand, the magnetic field is maximum at those planes where E is zero (the null planes of E), and has nulls where E is maximum. Since the time average power flow and the Poynting vector are clearly zero at each of these planes, there is no net power flow to the right. Except at the field nulls, however, there is reactive power, as discussed in Section 2.7.3. Because no average power is flowing via these waves and the energy and waves are approximately stationary in space, the solution is called a *standing wave*, as illustrated in Figures 7.2.3 for VSWR and 7.4.1 for resonance on perfectly reflecting TEM transmission lines.

9.1.4 Reflection from transmissive boundaries

Often more than one wave must be added to the given incident wave to satisfy all boundary conditions. For example, assume the same uniform plane wave (9.1.1–2) in vacuum is incident upon the same planar interface, where a medium having $\mu,\varepsilon \neq \mu_o,\varepsilon_o$ for $z \geq 0$ has replaced the conductor. We have no reason to suspect that the fields beyond the interface are zero, so we might try a trial solution with both a reflected wave $E_r(z,t)$ and a transmitted wave $E_t(z,t)$ having unknown amplitudes (E_r and E_t) and phases (ϕ and θ) for which we can solve:

$$\overline{E}(z,t) = \hat{x}\left[E_o \cos(\omega t - kz) + E_r \cos(\omega t + kz + \phi) \right] \qquad (z < 0) \quad (9.1.11)$$

$$\overline{E}_t(z,t) = \hat{x}E_t \cos(\omega t - k_t z + \theta) \qquad (z \geq 0) \quad (9.1.12)$$

$$\overline{H}(z,t) = \hat{y}\left[E_o \cos(\omega t - kz) - E_r \cos(\omega t + kz + \phi) \right]/\eta_o \qquad (z < 0) \quad (9.1.13)$$

$$\overline{H}(z,t) = \hat{y}E_t \cos(\omega t - k_t z + \theta)/\eta_t \qquad (z \geq 0) \quad (9.1.14)$$

where $k = \omega\sqrt{\mu_o\varepsilon_o}$, $k_t = \omega\sqrt{\mu\varepsilon}$, $\eta_o = \sqrt{\mu_o/\varepsilon_o}$, and $\eta_t = \sqrt{\mu/\varepsilon}$.

Using these four equations to match boundary conditions at $z = 0$ for $\overline{E}_{//}$ and $\overline{H}_{//}$, both of which are continuous across an insulating boundary, and dividing by E_o, yields:

$$\hat{x}\left[\cos(\omega t)+(E_r/E_o)\cos(\omega t + \phi)\right] = \hat{x}(E_t/E_o)\cos(\omega t + \theta) \qquad (9.1.15)$$

$$\hat{y}\left[\cos(\omega t)-(E_r/E_o)\cos(\omega t + \phi)\right]\big/\eta_o = \hat{y}\left[(E_t/E_o)\cos(\omega t + \theta)\right]\big/\eta_t \qquad (9.1.16)$$

First we note that for these equations to be satisfied for all time t we must have $\phi = \theta = 0$, unless we reverse the signs of E_r or E_t and let ϕ or $\theta = \pi$, respectively, which is equivalent.

Dividing these two equations by $\cos \omega t$ yields:

$$1+(E_r/E_o) = E_t/E_o \qquad (9.1.17)$$

$$\left[1-(E_r/E_o)\right]\big/\eta_o = (E_t/E_o)\big/\eta_t \qquad (9.1.18)$$

These last two equations can easily be solved to yield the *wave reflection coefficient* and the wave transmission coefficient:

$$\frac{E_r}{E_o} = \frac{(\eta_t/\eta_o)-1}{(\eta_t/\eta_o)+1} \qquad \text{(reflection coefficient)} \quad (9.1.19)$$

$$\frac{E_t}{E_o} = \frac{2\eta_t}{\eta_t + \eta_o} \qquad \text{(transmission coefficient)} \quad (9.1.20)$$

The wave transmission coefficient E_t/E_o follows from (9.1.17) and (9.1.19). When the characteristic impedance η_t of the dielectric equals that of the incident medium, η_o, there are no reflections and the transmitted wave equals the incident wave. We then have an *impedance match*. These values for E_r/E_o and E_t/E_o can be substituted into (9.1.11–14) to yield the final solution for $\overline{E}(z,t)$ and $\overline{H}(z,t)$.

The last step of the four-step method for solving boundary value problems involves checking this solution against all Maxwell's equations—they are satisfied.

Example 9.1A
A 1-Wm⁻² uniform plane wave in vacuum, $\hat{x}E_+ \cos(\omega t - kz)$, is normally incident upon a planar dielectric with $\varepsilon = 4\varepsilon_o$. What fraction of the incident power P_+ is reflected? What is $\overline{H}(t)$ at the dielectric surface $(z = 0)$?

Solution: $P_-/P_+ = |E_-/E_+|^2 = |(\eta_t - \eta_o)/(\eta_t + \eta_o)|^2$, using(5.1.19). Since $\eta_t = \sqrt{\mu_o/4\varepsilon_o} = \eta_o/2$,

therefore: $P_-/P_+ = |[(\eta_o/2) - \eta_o]/[(\eta_o/2) + \eta_o]|^2 = (-1/3)^2 = 1/9$. For the forward

wave: $\overline{E} = \hat{x}E_+ \cos(\omega t - kz)$ and $\overline{H} = \hat{y}(E_+/\eta_o)\cos(\omega t - kz)$, where $|E_+|^2/2\eta_o = 1$,

so $E_+ = (2\eta_o)^{0.5} = (2 \times 377)^{0.5} \cong 27$ [V/m]. The sum of the incident and reflected

magnetic fields at $z = 0$ is $\overline{H} = \hat{y}(E_+/\eta_o)[\cos(\omega t) - (1/3)\cos(\omega t)]$

$\cong \hat{y}(27/377)(2/3)\cos(\omega t) = 0.48\hat{y}\cos(\omega t)$ $[\text{A m}^{-1}]$

9.2   *Waves incident on planar boundaries at angles*

9.2.1   Introduction to waves propagating at angles

To determine electromagnetic fields we can generally solve a boundary value problem using the method of Section 9.1.1, the first step of which involves characterization of the basic quasistatic or dynamic fields and waves that could potentially exist within each separate region of the problem. The final solution is a linear combination of these basic fields and waves that matches all boundary conditions at the interfaces between the various regions.

So far we have considered only waves propagating along boundaries or normal to them. The general case involves waves incident upon boundaries at arbitrary angles, so we seek a compact notation characterizing such waves that simplifies the boundary value equations and their solutions. Because wave behavior at boundaries often becomes frequency dependent, it is convenient to use complex notation as introduced in Section 2.3.2 and reviewed in Appendix B, which can explicitly represent the frequency dependence of wave phenomena. For example, we might represent the electric field associated with a uniform plane wave propagating in the +z direction as $\underline{E}_o e^{-jkz}$, where:

$$\underline{E}(z) = \underline{E}_o e^{-jkz} = \underline{E}_o e^{-j2\pi z/\lambda} \tag{9.2.1}$$

$$\underline{E}_o = \hat{x}\underline{E}_{ox} + \hat{y}\underline{E}_{oy} \tag{9.2.2}$$

This notation is simpler than the time domain representation. For example, if this wave were x-polarized, then the compact complex notation $\hat{x}\underline{E}_x$ would be replaced in the time domain by:

$$\overline{E}(t) = R_e\{\hat{x}\underline{E}_x(z)e^{j\omega t}\} = \hat{x}R_e\{(R_e[\underline{E}_x(z)] + jI_m[\underline{E}_x(z)])(\cos\omega t + j\sin\omega t)\}$$
$$= \hat{x}\{R_e[\underline{E}_x(z)]\cos\omega t - I_m[\underline{E}_x(z)]\sin\omega t\} \tag{9.2.3}$$

The more general time-domain expression including both x and y components would be twice as long. Thus complex notation adequately characterizes frequency-dependent wave propagation and is more compact.

The physical significance of (9.2.1) is divided into two parts: $\overline{E}_o$ tells us the polarization, amplitude, and absolute phase of the wave at the origin, and $e^{-j2\pi z/\lambda} \equiv e^{j\phi(z)}$ tells us how the phase ϕ of this wave varies with position. In this case the phase decreases 2π radians as z increases by one wavelength λ. The physical significance of a phase shift ϕ of -2π radians for $z = \lambda$ is that observers located at $z = \lambda$ experience a delay of 2π radians; for pure sinusoids a phase shift of 2π is of course not observable.

Waves propagating in arbitrary directions are therefore easily represented by expressions similar to (9.2.1), but with a phase ϕ that is a function of x, y, and z. For example, a general plane wave would be:

$$\overline{E}(z) = \overline{E}_o e^{-jk_x x - jk_y y - jk_z z} = \overline{E}_o e^{-j\overline{k} \cdot \overline{r}} \tag{9.2.4}$$

where $\overline{r} = \hat{x}x + \hat{y}y + \hat{z}z$ and:

$$\overline{k} = \hat{x}k_x + \hat{y}k_y + \hat{z}k_z \tag{9.2.5}$$

We call $\overline{k}$ the *propagation vector* or *wave number* $\overline{k}$. The wave numbers k_x, k_y, and k_z have the dimensions of radians per meter and determine how rapidly the wave phase ϕ varies with position along the x, y, and z axes, respectively. Positions having the same value for $\overline{k} \bullet \overline{r}$ have the same phase and are located on the same *phase front*. A wave with a planar phase fronts is a *plane wave*, and if its amplitude is constant across any phase front, it is a *uniform plane wave*.
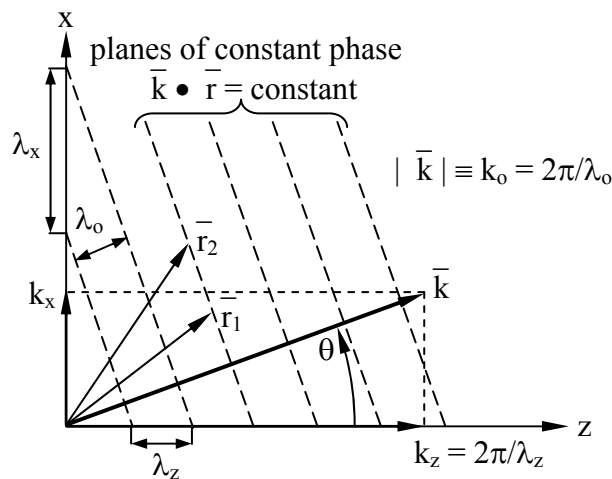
Figure 9.2.1   Uniform plane wave propagating at angle θ in the x-z plane.

The vector $\overline{k}$ points in the direction of propagation for uniform plane waves. The geometry is represented in Figure 9.2.1 for a uniform plane wave propagating in the x-z plane at an angle θ and wavelength λ_o. The planes of constant phase are perpendicular to the *wave vector* $\overline{k}$ because $\overline{k} \bullet \overline{r}$ must be constant everywhere in such a plane.

The solution (9.2.4) can be substituted into the wave equation (2.3.21):

$$\left(\nabla^2 + \omega^2 \mu\varepsilon\right)\overline{E} = 0 \tag{9.2.6}$$

This substitution yields[47]:

$$\left[-\left(k_x^2 + k_y^2 + k_z^2\right) + \omega^2 \mu\varepsilon\right]\overline{E} = 0 \tag{9.2.7}$$

$$k_x^2 + k_y^2 + k_z^2 = k_o^2 = \omega^2 \mu\varepsilon = \left|\overline{k}\right|^2 = \overline{k} \bullet \overline{k} \tag{9.2.8}$$

Therefore the figure and (9.2.7) suggest that:

$$k_x = \overline{k} \bullet \hat{x} = k_o \sin\theta, \ \ k_z = \overline{k} \bullet \hat{z} = k_o \cos\theta \tag{9.2.9}$$

The figure also includes the wave propagation vector components $\hat{x}\,k_x$ and $\hat{z}\,k_z$.

Three *projected wavelengths*, λ_x, λ_y, and λ_z, are perceived by observers moving along those three axes. The distance between successive wavefronts at 2π phase intervals is λ_o in the direction of propagation, and the distances separating these same wavefronts as measured along the x and z axes are equal or greater, as illustrated in Figure 9.2.1. For example:

$$\lambda_z = \lambda_o/\cos\theta = 2\pi/k_z \geq \lambda_o \tag{9.2.10}$$

Combining (9.2.8) and (9.2.10) yields:

$$\lambda_x^{-2} + \lambda_y^{-2} + \lambda_z^{-2} = \lambda_o^{-2} \tag{9.2.11}$$

The electric field $\overline{E}(\overline{r})$ for the wave of Figure 9.2.1 propagating in the x-z plane is orthogonal to the wave propagation vector $\overline{k}$. For simplicity we assume this wave is y-polarized:

$$\overline{E}(\overline{r}) = \hat{y}\underline{E}_o e^{-j\overline{k}\bullet\overline{r}} \tag{9.2.12}$$

---

[47] $\nabla^2 \overline{E} = \left(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2\right)\overline{E} = -\left(k_x^2 + k_y^2 + k_z^2\right)\overline{E}.$

The corresponding magnetic field is:

$$\overline{H}(\overline{r}) = -(\nabla \times \underline{\overline{E}})/j\omega\mu_o = (\hat{x}\partial\underline{E}_y/\partial z - \hat{z}\partial\underline{E}_y/\partial x)/j\omega\mu_o$$

$$= (\hat{z}\sin\theta - \hat{x}\cos\theta)(E_o/\eta_o)e^{-j\overline{k}\cdot\overline{r}}$$

(9.2.13)

One difference between this uniform y-polarized plane wave propagating at an angle and one propagating along a cartesian axis is that $\overline{H}$ no longer lies along a single axis, although it remains perpendicular to both $\overline{E}$ and $\overline{k}$. The next section treats such waves further.

Example 9.2A

If $\lambda_x = 2\lambda_z$ in Figure 9.2.1, what are θ, λ_o, and $\overline{k}$?

Solution: By geometry, $\theta = \tan^{-1}(\lambda_z/\lambda_x) = \tan^{-1}0.5 \cong 27°$.   By (9.2.11) $\lambda_o^{-2} = \lambda_x^{-2} + \lambda_z^{-2}$

$= (0.25+1)\lambda_z^{-2}$, so $\lambda_o = 1.25^{-0.5}\lambda_z = 0.89\lambda_z$. $\overline{k} = \hat{x}k_x + \hat{z}k_z = \hat{x}k_o\sin\theta + \hat{z}k_o\cos\theta$,

where $k_o = 2\pi/\lambda_o$. Alternatively, $\overline{k} = \hat{x}2\pi/\lambda_x + \hat{z}2\pi/\lambda_z$.

9.2.2   Waves at planar dielectric boundaries

Waves at planar dielectric boundaries are solved using the boundary-value-problem method of Section 9.1.4 applied to waves propagating at angles, as introduced in Section 9.2.1.

Because the behavior of waves at an interface depends upon their polarization we need a coordinate system for characterizing it. For this purpose the *plane of incidence* is defined as the plane of projection of the incident wave propagation vector $\overline{k}$ upon the interface, as illustrated in Figure 9.2.2(a). One cartesian axis is traditionally defined as being normal to this plane of incidence; in the figure it is the y axis.
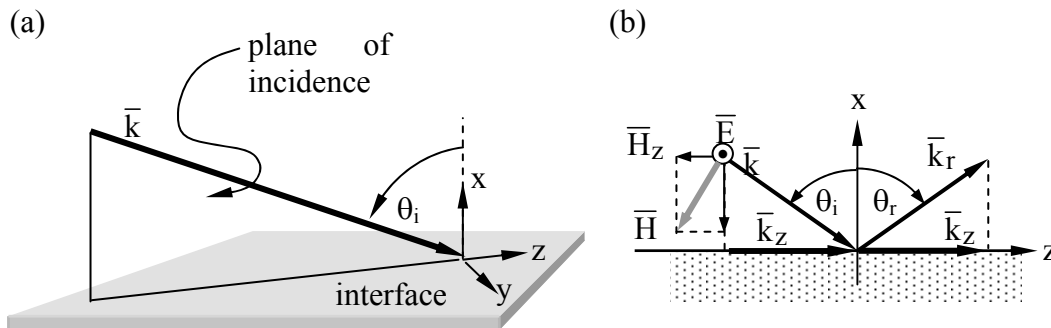
Figure 9.2.2   Uniform plane wave incident upon a planar interface.

We know from Section 2.3.4 that any pair of orthogonally polarized uniform plane waves can be superimposed to achieve any arbitrary wave polarization. For example, x- and y-polarized waves can be superimposed. It is customary to recognize two simple types of incident

electromagnetic waves that can be superimposed to yield any incident wave polarization: *transverse electric waves (TE waves)* are linearly polarized transverse to the plane of incidence (y-polarized in the figure), and *transverse magnetic waves (TM waves)* have the orthogonal linear polarization so that the magnetic field is purely transverse (again if y-polarized). TE and TM waves are typically transmitted and reflected with different amplitudes.

Consider first a TE wave incident upon the planar interface of Figure 9.2.2(b) at the incidence angle θ_i. The corresponding $\overline{H}$ lies in the x-z plane and is orthogonal to $\overline{E}$. $\overline{H}$ points downward in the figure, corresponding to power $\overline{S} = \overline{E} \times \overline{H}$ propagating toward the interface, where $\overline{S}$ is the Poynting vector for the incident wave. The wavelength of the wave above the interface is $\lambda_o = 1/(f\sqrt{\mu\varepsilon})$ in the medium characterized by permittivity ε and permeability μ. The medium into which the wave is partially transmitted is characterized by ε_t and μ_t, and there the wave has wavelength $\lambda_t = 1/(f\sqrt{\mu_t\varepsilon_t})$ and the same frequency f. This incident TE wave can be characterized by:

$$\overline{E}_i = \hat{y}\underline{E}_o e^{jk_x x - jk_z z} \left[V\ m^{-1} \right] \tag{9.2.14}$$

$$\overline{H}_i = -(\underline{E}_o/\eta)(\hat{x}\sin\theta_i + \hat{z}\cos\theta_i)e^{jk_x x - jk_z z} \left[A\ m^{-1} \right] \tag{9.2.15}$$

where the characteristic impedance of the incident medium is $\eta = \sqrt{\mu/\varepsilon}$, and $\overline{H}$ is orthogonal to $\overline{E}$.

The transmitted wave would generally be similar, but with a different η_t, θ_t, E_t, and $\overline{k}_t$. We might expect a reflected wave as well. The boundary-value-problem method of Section 9.1.2 requires expressions for all waves that might be present in both regions of this problem. In addition to the incident wave we therefore might add general expressions for reflected and transmitted waves having the same TE polarization. If still other waves were needed then no solution satisfying all Maxwell's equations would emerge until they were added too; we shall see no others are needed here. These general reflected and transmitted waves are:

$$\overline{E}_r = \hat{y}\underline{E}_r e^{-jk_{rx} x - jk_{rz} z} \left[V\ m^{-1} \right] \tag{9.2.16}$$

$$\overline{H}_r = (\underline{E}_r/\eta)(-\hat{x}\sin\theta_r + \hat{z}\cos\theta_r)e^{-jk_{rx} x - jk_{rz} z} \left[A\ m^{-1} \right] \tag{9.2.17}$$

$$\overline{E}_t = \hat{y}\underline{E}_t e^{jk_{tx} x - jk_{tz} z} \left[V\ m^{-1} \right] \tag{9.2.18}$$

$$\overline{H}_t = -(\underline{E}_t/\eta_t)(\hat{x}\sin\theta_t + \hat{z}\cos\theta_t)e^{jk_{tx} x - jk_{tz} z} \left[A\ m^{-1} \right] \tag{9.2.19}$$

Boundary conditions that must be met everywhere on the non-conducting surface at x = 0 include:

$$\overline{E}_{i\,//} + \overline{E}_{r\,//} = \overline{E}_{t\,//} \qquad\qquad\qquad (9.2.20)$$

$$\overline{H}_{i\,//} + \overline{H}_{r\,//} = \overline{H}_{t\,//} \qquad\qquad\qquad (9.2.21)$$

Substituting into (9.2.20) the values of $\overline{E}_{//}$ at the boundaries yields:

$$\underline{E}_o e^{-jk_z z} + \underline{E}_r e^{-jk_{rz} z} = \underline{E}_t e^{-jk_{tz} z} \qquad\qquad\qquad (9.2.22)$$

This equation can be satisfied for all values of z only if all exponents are equal. Therefore $e^{-jk_z z}$ can be factored out, simplifying the boundary-condition equations for both $\overline{E}_{//}$ and $\overline{H}_{//}$:

$$\underline{E}_o + \underline{E}_r = \underline{E}_t \qquad\qquad \text{(boundary condition for } E_{//}) \qquad (9.2.23)$$

$$\frac{\underline{E}_o}{\eta} \cos\theta_i - \frac{\underline{E}_r}{\eta} \cos\theta_r = \frac{\underline{E}_t}{\eta_t} \cos\theta_t \qquad\qquad \text{(boundary condition for } H_{//}) \qquad (9.2.24)$$

Because the exponential terms in (9.2.22) are all equal, it follows that the phases of all three waves must match along the full boundary, and:

$$k_{iz} = k_{rz} = k_{tz} = k_i \sin\theta_i = k_i \sin\theta_r = k_t \sin\theta_t = 2\pi/\lambda_z \qquad\qquad (9.2.25)$$

This *phase-matching condition* implies that the wavelengths of all three waves in the z direction must equal the same λ_z. It also implies that the *angle of reflection* θ_r equals the *angle of incidence* θ_i, and that the *angle of transmission* θ_t is related to θ_i by *Snell's law*:

$$\frac{\sin\theta_t}{\sin\theta_i} = \frac{k_i}{k_t} = \frac{c_t}{c_i} = \sqrt{\frac{\mu\varepsilon}{\mu_t\varepsilon_t}} \qquad\qquad \text{(Snell's law)} \qquad (9.2.26)$$

If $\mu = \mu_t$, then the angle of transmission becomes:

$$\theta_t = \sin^{-1}\left(\sin\theta_i \sqrt{\frac{\varepsilon}{\varepsilon_t}} \right) \qquad\qquad\qquad (9.2.27)$$

These phase-matching constraints, including Snell's law, apply equally to TM waves.

The magnitudes of the reflected and transmitted TE waves can be found by solving the simultaneous equations (9.2.23) and (9.2.24):

$$\underline{E}_r/\underline{E}_o = \Gamma(\theta_i) = \frac{\eta_t \cos\theta_i - \eta \cos\theta_t}{\eta_t \cos\theta_i + \eta \cos\theta_t} = \frac{\eta_n' - 1}{\eta_n' + 1} \qquad\qquad (9.2.28)$$

$$\underline{E}_t/\underline{E}_o = \underline{T}(\theta_i) = \frac{2\eta_t \cos\theta_i}{\eta_t \cos\theta_i + \eta \cos\theta_t} = \frac{2\eta_n'}{\eta_n' + 1} \qquad (9.2.29)$$

where we have defined the normalized angular impedance for TE waves as $\eta_n' \equiv$ $\eta_t\cos\theta_i/(\eta\cos\theta_t)$. The complex angular reflection and transmission coefficients $\underline{\Gamma}$ and $\underline{T}$ for TE waves approach those given by (9.1.19) and (9.1.20) for normal incidence in the limit $\theta_i \to 0$. The limit of grazing incidence is not so simple, and even the form of the transmitted wave can change markedly if it becomes evanescent, as discussed in the next section. The results for incident TM waves are postponed to Section 9.2.6. Figure 9.2.6(a) plots $|\underline{\Gamma}(\theta)|^2$ for a typical dielectric interface. It is sometimes useful to note that (9.2.28) and (9.2.29) also apply to equivalent TEM lines for which the characteristic impedances of the input and output lines are $\eta_i/\cos\theta_i$ and $\eta_t/\cos\theta_t$, respectively. When TM waves are incident, the corresponding equivalent impedances are $\eta_i\cos\theta_i$ and $\eta_t\cos\theta_t$, respectively.

Example 9.2B
What fraction of the normally incident power ($\theta_i = 0$) is reflected by a single glass camera lense having $\varepsilon = 2.25\varepsilon_o$? If $\theta_i = 30^o$, what is θ_t in the glass?

Solution: At each interface between air and glass, (9.2.28) yields for $\theta_i = 0$: $\underline{\Gamma}_L = (\eta_n' - 1)/(\eta_n' + 1)$, where $\eta_n' = (\eta_{glass}\cos\theta_i)/(\eta_{air}\cos\theta_t) = (\varepsilon_i/\varepsilon_g)^{0.5} = 1/1.5$. Thus $\underline{\Gamma}_L = (1 - 1.5)/(1 + 1.5) = -0.2$, and $|\underline{\Gamma}|^2 = 0.04$, so ~4 percent of the power is reflected from each of the two curved surfaces for each independent lense, or ~8 percent total; these reflections are incoherent so their reflected powers add. Modern lenses have many elements with different permittivities, but coatings on them reduce these reflections, as discussed in Section 7.3.2 for quarter-wave transformers. Snell's law (9.2.26) yields $\theta_t = \sin^{-1}[(\varepsilon_i/\varepsilon_t)^{0.5} \sin\theta_i] = \sin^{-1}[(1/1.5)(0.5)] = 19.5°$.

9.2.3   Evanescent waves

Figure 9.2.3 suggests why a special form of electromagnetic wave is sometimes required in order to satisfy boundary conditions. Figure 9.2.3(a) illustrates how the required equality of the z components of the incident, reflected, and transmitted wave propagation vectors $\overline{k}$ controls the angles of reflection and transmission, θ_r and θ_t. The radii of the two semi-circles correspond to the magnitudes of $\overline{k}_i$ and $\overline{k}_t$.

Figure 9.2.3(b) shows that a wave incident at a certain *critical angle* θ_c will produce a transmitted wave propagating parallel to the interface, provided $|\overline{k}_t| < |\overline{k}_i|$. Snell's law (9.2.26) can be evaluated for $\sin\theta_t = 1$ to yield:

$$\theta_c = \sin^{-1}(c_i/c_t) \text{ for } c_i < c_t \qquad \text{(critical angle)} \quad (9.2.30)$$

Figures 9.2.3(b) illustrates why phase matching is impossible with uniform plane waves when $\theta > \theta_c$; $k_z > |\overline{k}_t|$. Therefore the λ_z determined by λ and θ_i is less than λ_t, the natural wavelength of the transmission medium at frequency ω. A non-uniform plane wave is then required for phase matching, as discussed below.
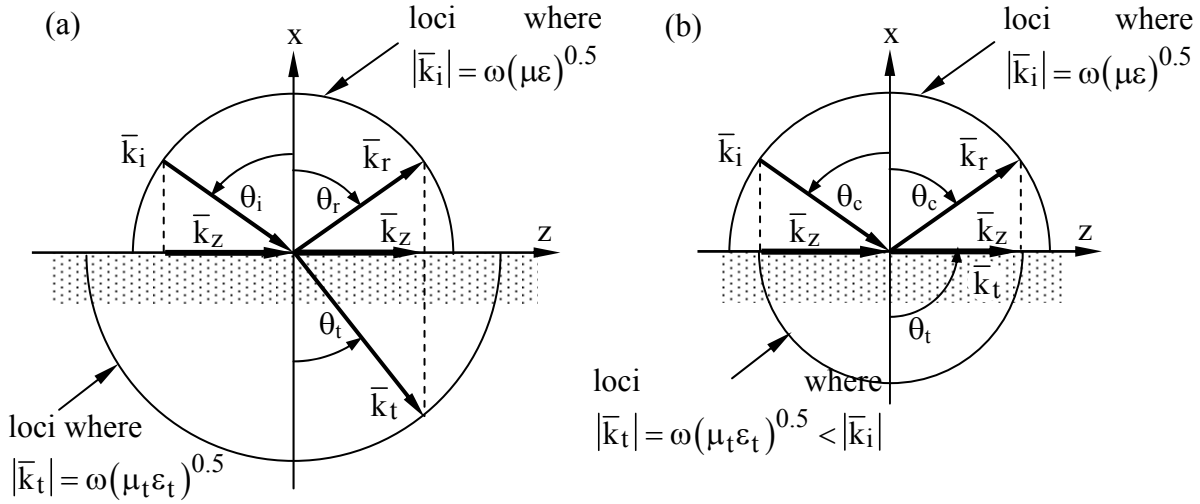

Figure 9.2.3   Angles of reflection and transmission, and the critical angle θ_c.

The wave propagation vector $\overline{k}_t$ must satisfy the wave equation $\left(\nabla^2 + k_t{}^2\right)\underline{E} = 0$. Therefore the transmitted wave must be proportional to $e^{-j\overline{k}_t \cdot \overline{r}}$, where $\overline{k}_t = \hat{k}k_t$ and $k_z = k_i \sin\theta_i$, satisfy the expression:

$$k_t{}^2 = \omega^2\mu_t\varepsilon_t = k_{tx}{}^2 + k_z{}^2 \qquad (9.2.31)$$

When $k_t{}^2 < k_z{}^2$ it follows that:

$$k_{tx} = \pm j\left(k_z{}^2 - k_t{}^2\right)^{0.5} = \pm j\alpha \qquad (9.2.32)$$

We choose the positive sign for α so that the wave amplitude decays with distance from the power source rather than growing exponentially.

The transmitted wave then becomes:

$$\underline{E}_t(x,z) = \hat{y}\underline{E}_t e^{jk_{tx}x - jk_z z} = \hat{y}\underline{E}_t e^{\alpha x - jk_z z} \qquad (9.2.33)$$

Note that x is negative in the decay region. The rate of decay $\alpha = (k_i{}^2\sin^2\theta_i - k_t{}^2)^{0.5}$ is zero when $\theta_i = \theta_c$ and increases as θ_i increases past θ_c; Waves that decay with distance from an interface and propagate power parallel to it are called *surface waves*.

The associated magnetic field $\overline{H}_t$ can be found by substituting (9.2.33) into Faraday's law:

$$\overline{H}_t = \nabla \times \overline{E}/(-j\omega\mu_t) = -\left(\underline{E}_t/\eta_t\right)\left(\hat{x}\sin\theta_t - \hat{z}\cos\theta_t\right)e^{\alpha x - jk_z z} \qquad (9.2.34)$$

This is the same expression as (9.2.19), which was obtained for normal incidence, except that the magnetic field and wave now decay with distance x from the interface rather than propagating in that direction. Also, since $\sin\theta_t > 1$ for $\theta_i > \theta_c$, $\cos\theta_t$ is now imaginary and positive, and $\overline{H}$ is not in phase with $\overline{E}$. As a result, Poynting's vector for these surface waves has a real part corresponding to real power propagating parallel to the surface, and an imaginary part corresponding to reactive power flowing perpendicular to the surface in the direction of wave decay:

$$\overline{S} = \overline{E} \times \overline{H}^* = \left(-j\alpha\hat{x} + k_z\hat{z}\right)\left(\left|\underline{E}_t\right|^2/\omega\mu_t\right)e^{2\alpha x}\left[\text{Wm}^{-2}\right] \qquad (9.2.35)$$

The reactive part flowing in the $-\hat{x}$ direction is $+j\alpha\left|\underline{E}_t\right|^2/\omega\mu_t e^{2ax}$ and is therefore inductive (+j), corresponding to an excess of magnetic stored energy relative to electric energy within this surface wave. A wave such as this one that decays in a direction for which the power flow is purely reactive is designated an *evanescent wave*.

An instantaneous view of the electric and magnetic fields of a non-uniform TE plane wave formed at such a dielectric boundary is shown in Figure 9.2.4; these correspond to the fields of (9.2.33) and (9.2.34).
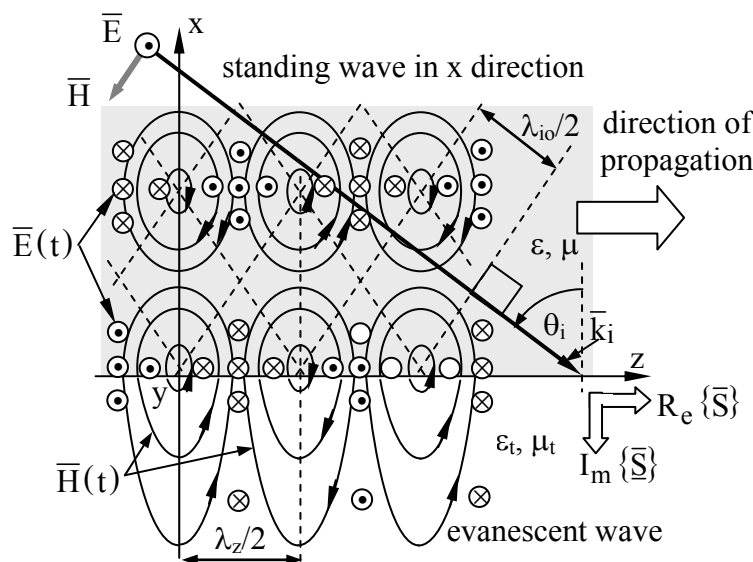
Figure 9.2.4 Evanescent wave traveling in the +z direction at a dielectric interface.

The conventional notation used here indicates field strength by the density of symbols or field lines, and the arrows indicate field direction. Small circles correspond to field lines pointing perpendicular to the page; center dots indicate field lines pointing out of the page in the +y direction and center crosses indicate the opposite, i.e., field lines pointing into the page.

The time average wave intensity in the $+\hat{z}$ direction for x negative and outside the dielectric is:

$$P_z = 0.5R_e\left\{\overline{\underline{E}}\times\overline{\underline{H}}^{*}\right\} = \left(k_z\left|\underline{E}_t\right|^2\middle/2\omega\mu_t\right)e^{\alpha x}\left[\text{W m}^{-2}\right] \qquad (9.2.36)$$

Since the real and imaginary parts of $\overline{\underline{S}}$ are orthogonal, there is no decay in the direction of propagation, and therefore no power absorption or heating of the media. Beyond the critical angle θ_c the power is perfectly reflected. In the next section we shall see that the real and imaginary parts of $\overline{\underline{S}}$ are often neither orthogonal nor parallel.

9.2.4   Waves in lossy media

Sometimes one or both of the two media are conductive. This section explores the nature of waves propagating in such lossy media having conductivity $\sigma > 0$. Section 9.2.5 then discusses reflections from such media. Losses can also arise if ε or μ are complex. The quasistatic relaxation of charge, current, and field distributions in lossy media is discussed separately in Section 4.3.

We can determine the nature of waves in lossy media using the approach of Section 2.3.3 and including the conduction currents $\overline{J}$ in Ampere's law:

$$\nabla\times\overline{\underline{H}} = \overline{\underline{J}} + j\omega\varepsilon\overline{\underline{E}} = \sigma\overline{\underline{E}} + j\omega\varepsilon\overline{\underline{E}} = j\omega\underline{\varepsilon}_{\text{eff}}\overline{\underline{E}} \qquad (9.2.37)$$

where the effective complex permittivity $\underline{\varepsilon}_{\text{eff}}$ is:

$$\underline{\varepsilon}_{\text{eff}} = \varepsilon\left[1-\left(j\sigma/\omega\varepsilon\right)\right] \qquad (9.2.38)$$

The quantity $\sigma/\omega\varepsilon$ is called the *loss tangent* of the medium and indicates how fast waves decay. As we shall see, waves propagate well if $\sigma << \omega\varepsilon$, and decay rapidly if $\sigma > \omega\varepsilon$, sometimes within a fraction of a wavelength.

Substituting $\underline{\varepsilon}_{\text{eff}}$ for ε in $k^2 = \omega^2\mu\varepsilon$ yields the dispersion relation:

$$\underline{k}^2 = \omega^2\mu\varepsilon\left[1-\left(j\sigma/\omega\varepsilon\right)\right] = \left(k'-jk''\right)^2 \qquad (9.2.39)$$

where we define the complex wavenumber $\underline{k}$ in terms of its real and imaginary parts as:

$$\underline{k} = k' - jk'' \tag{9.2.40}$$

The form of the wave solution, following (2.3.26), is therefore:

$$\underline{E}(\bar{r}) = \hat{y} E_o e^{-jk'z - k''z} \quad \left[v\ m^{-1} \right] \tag{9.2.41}$$

This wave has wavelength λ', frequency ω, and phase velocity v_p inside the conductor related by:

$$k' = 2\pi/\lambda' = \omega/v_p \tag{9.2.42}$$

and the wave decays exponentially with z as $e^{-k''z} = e^{-z/\Delta}$. Note that the wave decays in the same direction as it propagates, corresponding to power dissipation, and that the 1/e *penetration depth* Δ is 1/k'' meters. Inside conductors, λ' and v_p are much less than their free-space values.

We now need to determine k' and k''. In general, matching the real and imaginary parts of (9.2.39) yields two equations that can be solved for k' and k'':

$$(k')^2 - (k'')^2 = \omega^2 \mu\varepsilon \tag{9.2.43}$$

$$2k'k'' = \omega\mu\sigma \tag{9.2.44}$$

However, in the limits of very high or very low values of the loss tangent $\sigma/\omega\varepsilon$, it is much easier to evaluate (9.2.39) directly.

In the low loss limit where $\sigma \ll \omega\varepsilon$, (9.2.39) yields:

$$\underline{k} = \omega\sqrt{\mu\varepsilon\left[1 - (j\sigma/\omega\varepsilon)\right]} \cong \omega\sqrt{\mu\varepsilon} - j\sigma\eta/2 \qquad (\sigma \ll \omega\varepsilon) \tag{9.2.45}$$

where the approximate wave impedance of the medium is $\eta = \sqrt{\mu/\varepsilon}$, and we have used the Taylor series approximation $\sqrt{1+\delta} \cong 1 + \delta/2$ for $\delta \ll 1$. In this limit we see from (9.2.45) that λ' and $v_p \cong c$ are approximately the same as they are for the lossless case, and that the 1/e penetration depth $\Delta \cong 2/\sigma\eta$, which becomes extremely large as $\sigma \to 0$.

In the high loss limit where $\sigma \gg \omega\varepsilon$, (9.2.39) yields:

$$\underline{k} = \omega\sqrt{\mu\varepsilon\left[1 - (j\sigma/\omega\varepsilon)\right]} \cong \sqrt{-j\mu\omega\sigma} \qquad (\sigma \gg \omega\varepsilon) \tag{9.2.46}$$

$$\cong \sqrt{\omega\mu\sigma}\sqrt{-j} = \pm\sqrt{\frac{\omega\mu\sigma}{2}}(1-j) \tag{9.2.47}$$

The real and imaginary parts of $\underline{k}$ have the same magnitudes, and the choice of sign determines the direction of propagation. The wave generally decays exponentially as it propagates, although exponential growth occurs in media with negative conductivity. The penetration depth is commonly called the *skin depth* δ in this limit ($\sigma \gg \omega\varepsilon$), where:

$$\delta = 1/k'' \cong (2/\omega\mu\sigma)^{0.5} \quad [m] \qquad\qquad \text{(skin depth)} \quad (9.2.48)$$

Because the real and imaginary parts of $\underline{k}$ are equal here, both the skin depth and the wavelength λ' inside the conductor are extremely small compared to the free-space wavelength λ; thus:

$$\lambda' = 2\pi/k' = 2\pi\delta \quad [m] \qquad\qquad \text{(wavelength in conductor)} \quad (9.2.49)$$

These distances δ and λ' are extremely short in common metals such as copper ($\sigma \cong 5.8 \times 10^7$, $\mu = \mu_o$) at frequencies such as 1 GHz, where $\delta \cong 2 \times 10^{-6}$ m and $\lambda' \cong 13 \times 10^{-6}$ m, which are roughly five orders of magnitude smaller than the 30-cm free space wavelength. The phase velocity v_p of the wave is reduced by the same large factor.

In the high conductivity limit, the wave impedance of the medium also becomes complex:

$$\underline{\eta} = \sqrt{\frac{\mu}{\varepsilon_{\text{eff}}}} = \sqrt{\frac{\mu}{\varepsilon(1 - j\sigma/\omega\varepsilon)}} \cong \sqrt{\frac{j\omega\mu}{\sigma}} = \sqrt{\frac{\omega\mu}{2\sigma}}(1 + j) \qquad (9.2.50)$$

where $+j$ is consistent with a decaying wave in a lossy medium. The imaginary part of $\underline{\eta}$ corresponds to power dissipation, and is non-zero whenever $\sigma \neq 0$.

Often we wish to shield electronics from unwanted external radiation that could introduce noise, or to ensure that no radiation escapes to produce *radio frequency interference* (RFI) that affects other systems. Although the skin depth effect shields electromagnetic radiation, high conductivity will reflect most incident radiation in any event. Conductors generally provide good *shielding* at higher frequencies for which the time intervals are short compared to the magnetic relaxation time (4.3.15) while remaining long compared to the charge relaxation time (4.3.3); Section 4.3.2 and Example 4.3B present examples of magnetic field diffusion into conductors.

Example 9.2C
A uniform plane wave propagates at frequency $f = c/\lambda = 1$ MHz in a medium characterized by ε_o, μ_o, and conductivity σ. If $\sigma \cong 10^{-3}\omega\varepsilon_o$, over what distance D would the wave amplitude decay by a factor of 1/e? What would be the 1/e wave penetration depth δ in a good conductor having $\sigma \cong 10^{11}\omega\varepsilon$ at this frequency?

Solution: In the low-loss limit where $\sigma \ll \omega\varepsilon$, $\underline{k} \cong \omega/c - j\sigma\eta/2$ (9.2.45), so $E \propto e^{-\sigma\eta z/2} = e^{-z/D}$
where $D = 2/\sigma\eta = 2(\varepsilon_o/\mu_o)^{0.5}/10^{-3}\omega\varepsilon_o = 2000c/\omega \cong 318\lambda$ [m]. In the high-loss limit $\underline{k}$

$\cong (1 \pm j)(\omega\mu\sigma/2)^{0.5}$ so $E \propto e^{-z/\delta}$, where $\delta = (2/\omega\mu\sigma)^{0.5} = (2\times10^{-11}/\omega^2\mu\varepsilon)^{0.5} = (2\times10^{-11})^{0.5}$ $\lambda/2\pi \cong 7.1\times10^{-7}\lambda = 0.21$ mm. This conductivity corresponds to typical metal and the resulting penetration depth is a tiny fraction of a free-space wavelength.

9.2.5   Waves incident upon good conductors

This section focuses primarily on waves propagating inside good conductors. The field distributions produced outside good conductors by the superposition of waves incident upon and reflected from them are discussed in Section 9.2.3.

Section 9.2.4 showed that uniform plane waves in lossy conductors decay as they propagate. The wave propagation constant $\underline{k}$ is then complex in order to characterize exponential decay with distance:

$$\underline{k} = k' - jk'' \tag{9.2.51}$$

The form of a uniform plane wave in lossy media is therefore:

$$\overline{\underline{E}}(\overline{r}) = \hat{y}\underline{E}_o e^{-jk'z - k''z} \; [v \, m^{-1}] \tag{9.2.52}$$

When a plane wave impacts a conducting surface at an angle, a complex wave propagation vector $\overline{\underline{k}}_t$ is required to represent the resulting transmitted wave. The real and imaginary parts of $\overline{\underline{k}}_t$ are generally at some angle to each other. The result is a *non-uniform plane wave* because its intensity is non-uniform across each phase front.

To illustrate how such transmitted non-uniform plane waves can be found, consider a lossy transmission medium characterized by ε, σ, and μ, where we can combine ε and σ into a single effective complex permittivity, as done in (9.2.38)[48]:

$$\varepsilon_{eff} \equiv \varepsilon(1 - j\sigma/\omega\varepsilon) \tag{9.2.53}$$

If we represent the electric field as $\overline{\underline{E}}_o e^{-j\overline{\underline{k}}\cdot\overline{r}}$ and substitute it into the wave equation $(\nabla^2 + \omega^2\mu\varepsilon)\overline{\underline{E}} = 0$, we obtain for non-zero $\overline{\underline{E}}$ the general *dispersion relation* for plane waves in isotropic lossy media:

$$\left[(-j\overline{\underline{k}}) \bullet (-j\overline{\underline{k}}) + \omega^2\mu\varepsilon_{eff}\right]\overline{\underline{E}} = 0 \tag{9.2.54}$$

$$\overline{\underline{k}} \bullet \overline{\underline{k}} = \omega^2\mu\varepsilon_{eff} \qquad \text{(dispersion relation)} \tag{9.2.55}$$

---

[48] $\nabla \times \overline{\underline{H}} = \overline{\underline{J}} + j\omega\overline{\underline{E}} = \sigma\overline{\underline{E}} + j\omega\overline{\underline{E}} = j\omega\varepsilon_{eff}\overline{\underline{E}}$

Once a plane of incidence such as the x-z plane is defined, this relation has four scalar unknowns—the real and imaginary parts for each of the x and z (in-plane) components of $\overline{k}$. At a planar boundary there are four such unknowns for each of the reflected and transmitted waves, or a total of eight unknowns. Each of these four components of $\overline{k}$ (real and imaginary, parallel and perpendicular) must satisfy a boundary condition, yielding four equations. The dispersion relation (9.2.55) has real and imaginary parts for each side of the boundary, thus providing four more equations. The resulting set of eight equations can be solved for the eight unknowns, and generally lead to real and imaginary parts for $\overline{k}_t$ that are neither parallel nor perpendicular to each other or to the boundary. That is, the real and imaginary parts of $\overline{k}$ and $\overline{S}$ can point in four different directions.

It is useful to consider the special case of reflections from planar conductors for which $\sigma \gg \omega\varepsilon$. In this limit the solution is simple because the transmitted wave inside the conductor propagates almost perpendicular to the interface, which can be shown as follows. Equation (9.2.47) gave the propagation constant $\underline{k}$ for a uniform plane wave in a medium with $\sigma \gg \omega\varepsilon$:

$$\underline{k} \cong \pm\sqrt{\frac{\omega\mu\sigma}{2}}(1-j) \tag{9.2.56}$$

The real part of such a $\overline{k}$ is so large that even for grazing angles of incidence, $\theta_i \cong 90^o$, the transmission angle θ_t must be nearly zero in order to match phases, as suggested by Figure 9.2.3(a) in the limit where k_t is orders of magnitude greater than k_i. As a result, the power dissipated in the conductor is essentially the same as for $\theta_i = 0^o$, and therefore depends in a simple way on the induced surface current and parallel surface magnetic field $\overline{H}_{//}$. $\overline{H}_{//}$ is simply twice that associated with the incident wave alone $(\overline{H}_\perp \cong 0)$; essentially all the incident power is reflected so the incident and reflected waves have the same amplitudes and their magnetic fields add.

The power density P_d [W m^{-2}] dissipated by waves traveling in the +z direction in conductors with an interface at $z = 0$ can be found using the Poynting vector:

$$P_d = \frac{1}{2}R_e\left\{\left(\overline{E}\times\overline{H}^*\right)\bullet\hat{z}\right\}\Big|_{z=0_+} = R_e\left\{\frac{|\underline{T}\underline{E}_i|^2}{2\underline{\eta}_t}\right\} = \frac{1}{2}R_e\left\{\frac{1}{\underline{\eta}_t}\right\}\eta_i^2\,|\underline{H}_i\underline{T}|^2 \tag{9.2.57}$$

The wave impedance $\underline{\eta}_t$ of the conductor ($\sigma \gg \omega\varepsilon$) was derived in (9.2.50), and (9.2.29) showed that $\underline{T} = 2\underline{\eta}_n'/\left(\underline{\eta}_n'+1\right) \cong 2\underline{\eta}_n'$ for TE waves and $\underline{\eta}_n' = \dfrac{\eta_t\cos\theta_i}{\eta_i\cos\theta_t} \ll 1$:

$$\underline{\eta}_t \cong \left(\omega\mu_t/2\sigma\right)^{0.5}(1+j) \tag{9.2.58}$$

$$\underline{T}_{TE}\left(\theta_i\right) \cong 2\underline{\eta}_n'/\left(\underline{\eta}_n'+1\right) \cong 2\underline{\eta}_n' \cong 2\underline{\eta}_t\cos\theta_t/\eta_i = \left(2\omega\mu_t\varepsilon_i/\mu_i\sigma\right)^{0.5}(1+j)\cos\theta_i \tag{9.2.59}$$

Therefore (9.2.57), (9.2.58), and (9.2.59) yield:

$$P_d \cong \sqrt{\frac{\sigma}{2\omega\mu_t}} \frac{\mu_i}{\varepsilon_i} |\underline{H}_i|^2 \frac{4\omega\mu_t\varepsilon_i}{2\mu_i\sigma} = |\overline{\underline{H}}(z=0)|^2 \sqrt{\frac{\omega\mu}{8\sigma}} \left[W/m^2\right] \qquad (9.2.60)$$

A simple way to remember (9.2.60) is to note that it yields the same dissipated power density that would result if the same surface current $\overline{J}_s$ flowed uniformly through a conducting slab having conductivity σ and a thickness equal to the skin depth $\delta = \sqrt{2/\omega\mu\sigma}$:

$$P_d = \frac{\delta}{2} R_e \left\{ \overline{\underline{E}} \bullet \overline{\underline{J}}^* \right\} = |\overline{\underline{J}}|^2 \frac{\delta}{2\sigma} = \frac{|\overline{\underline{J}}_s|^2}{2\sigma\delta} = |\overline{\underline{H}}(z=0)|^2 \sqrt{\frac{\omega\mu}{8\sigma}} \left[W/m^2\right] \qquad (9.2.61)$$

The significance of this result is that it simplifies calculation of power dissipated when waves impact conductors—we need only evaluate the surface magnetic field under the assumption the conductor is perfect, and then use (9.2.61) to compute the power dissipated per square meter.

Example 9.2D
What fraction of the 10-GHz power reflected by a satellite dish antenna is resistively dissipated in the metal if $\sigma = 5\times10^7$ Siemens per meter? Assume normal incidence. A wire of diameter D and made of the same metal carries a current $\underline{I}$. What is the approximate power dissipated per meter if the skin depth δ at the chosen frequency is much greater than D? What is this dissipation if $\delta \ll D$?

Solution: The plane wave intensity is $I = \eta_o |\underline{H}_+|^2 / 2 \, [W/m^2]$, and the power absorbed by a good

conductor is given by (9.2.61): $P_d \cong |2\underline{H}_+|^2 \sqrt{\omega\mu/4\sigma}$, where the magnetic field near a good conductor is twice the incident magnetic field due to the reflected wave. The fractional power absorbed is:

$P_d/I = 4\sqrt{\omega\mu/\sigma}/\eta_o = 4\sqrt{\omega\varepsilon_o/\sigma} \cong 4\left(2\pi10^{10}\times8.8\times10^{-12}/5\times10^7\right)^{0.5} = 4.2\times10^{-4}$. If

$\delta \gg D$, then a wire dissipates $|\underline{I}|^2 R/2$ watts $= 2|\underline{I}|^2/\sigma\pi D^2 \, [W/m^2]$. The magnetic field around a wire is: $\underline{H} = \underline{I}/\pi D$, and if $\delta \ll D$, then the power dissipated per meter

is: $\pi D |\underline{H}|^2 \sqrt{\omega\mu/4\sigma} = |\underline{I}|^2 \sqrt{\omega\mu/4\sigma}/\pi D \, [W/m^2]$, where the surface area for dissipation is $\pi D \, [m^2]$. Note that the latter dissipation is now increases with the square-root of frequency and is proportional to $1/\sqrt{\sigma}$, not $1/\sigma$.

9.2.6 Duality and TM waves at dielectric boundaries

Transverse magnetic (TM) waves reflect from planar surfaces just as do TE waves, except with different amplitudes as a function of angle. The angles of reflection and transmission are the

same as for TE waves, however, because both TE and TM waves must satisfy the same phase matching boundary condition (9.2.25).

The behavior of TE waves at planar boundaries is characterized by equations (9.2.14) and (9.2.15) for the incident electric and magnetic fields, (9.2.16) and (9.2.17) for the reflected wave, and (9.2.18) and (9.2.19) for the transmitted wave, supplemented by expressions for the complex reflection and transmission coefficients $\underline{\Gamma} = \underline{E}_r/\underline{E}_o$, (9.2.28), and $\underline{T} = \underline{E}_t/\underline{E}_o$, (9.2.29). Although the analogous behavior of TM waves could be derived using the same boundary-value problem solving method used in Section 9.2.2 for TE waves, the principle of *duality* can provide the same solutions with much less effort.

Duality works because Maxwell's equations without charges or currents are duals of themselves. That is, by transforming $\overline{E} \Rightarrow \overline{H}$, $\overline{H} \Rightarrow -\overline{E}$, and $\varepsilon \Leftrightarrow \mu$, the set of Maxwell's equations is unchanged:

$$\nabla \times \overline{E} = -\mu \partial \overline{H}/\partial t \qquad \rightarrow \qquad \nabla \times \overline{H} = \varepsilon \partial \overline{E}/\partial t \tag{9.2.62}$$

$$\nabla \times \overline{H} = \varepsilon \partial \overline{E}/\partial t \qquad \rightarrow \qquad -\nabla \times \overline{E} = \mu \partial \overline{H}/\partial t \tag{9.2.63}$$

$$\nabla \bullet \varepsilon \overline{E} = 0 \qquad \rightarrow \qquad \nabla \bullet \mu \overline{H} = 0 \tag{9.2.64}$$

$$\nabla \bullet \mu \overline{H} = 0 \qquad \rightarrow \qquad \nabla \bullet \varepsilon \overline{E} = 0 \tag{9.2.65}$$

The transformed set of equations on the right-hand side of (9.2.62) to (9.2.65) is the same as the original, although sequenced differently. As a result, any solution to Maxwell's equations is also a solution to the dual problem where the variables and boundary conditions are all transformed as indicated above.

The boundary conditions derived in Section 2.6 for a planar interface between two insulating uncharged media are that $\overline{E}_{//}$, $\overline{H}_{//}$, $\mu \overline{H}_\perp$, and $\varepsilon \overline{E}_\perp$ be continuous across the boundary. Since the duality transformation leaves these boundary conditions unchanged, they are dual too. However, duality cannot be used, for example, in the presence of perfect conductors that force $\overline{E}_{//}$ to zero, but not $\overline{H}_{//}$.
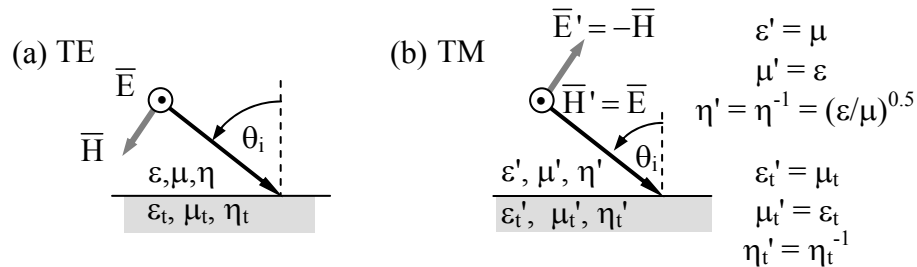
Figure 9.2.5   Dual TE and TM waves incident upon a dual planar boundary.

Figure 9.2.5(b) illustrates a TM plane wave incident upon a planar boundary where both the wave and the boundary conditions are dual to the TE wave illustrated in (a).

The behavior of TM waves at planar boundaries between non-conducting media is therefore characterized by duality transformations of Equations (9.2.62–65) for TE waves, supplemented by similar transformations of the expressions for the complex reflection and transmission coefficients $\underline{\Gamma} = \underline{E}_r/\underline{E}_o$, (9.2.28), and $\underline{T} = \underline{E}_t/\underline{E}_o$, (9.2.29). After the transformations $\overline{E} \Rightarrow \overline{H}$, $\overline{H} \Rightarrow -\overline{E}$, and $\varepsilon \Leftrightarrow \mu$, Equations (9.2.14–19) become:

$$\overline{H}_i = \hat{y}\underline{H}_o e^{jk_x x - jk_z z}\left[Am^{-1}\right]$$ (9.2.66)

$$\overline{E}_i = \left(\underline{H}_o \eta\right)\left(\hat{x}\sin\theta_i + \hat{z}\cos\theta_i\right)e^{jk_x x - jk_z z}\left[Vm^{-1}\right]$$ (9.2.67)

$$\overline{H}_r = \hat{y}\underline{H}_r e^{-jk_{rx} x - jk_z z}\left[Am^{-1}\right]$$ (9.2.68)

$$\overline{E}_r = \left(\underline{H}_r \eta\right)\left(\hat{x}\sin\theta_r - \hat{z}\cos\theta_r\right)e^{-jk_{rx} x - jk_z z}\left[Vm^{-1}\right]$$ (9.2.69)

$$\overline{H}_t = \hat{y}\underline{H}_t e^{jk_{tx} x - jk_z z}\left[Am^{-1}\right]$$ (9.2.70)

$$\overline{E}_t = \left(\underline{H}_t \eta_t\right)\left(\hat{x}\sin\theta_t + \hat{z}\cos\theta_t\right)e^{jk_{tx} x - jk_z z}\left[Vm^{-1}\right]$$ (9.2.71)

The complex reflection and transmission coefficients for TM waves are transformed versions of (9.2.28) and (9.2.29), where we define a new angle-dependent η_n by interchanging $\mu \leftrightarrow \varepsilon$ in η_n' in (9.2.28):

$$\underline{H}_r/\underline{H}_o = \left(\eta_n^{-1} - 1\right)\big/\left(\eta_n^{-1} + 1\right)$$ (9.2.72)

$$\underline{H}_t/\underline{H}_o = 2\eta_n^{-1}\big/\left(\eta_n^{-1} + 1\right)$$ (9.2.73)

$$\eta_n^{-1} \equiv \eta\cos\theta_i\big/\left(\eta_t\cos\theta_t\right)$$ (9.2.74)

These equations, (9.2.66) to (9.2.74), completely describe the TM case, once phase matching provides θ_r and θ_t.

It is interesting to compare the power reflected for TE and TM waves as a function of the angle of incidence θ_i. Power in uniform plane waves is proportional to both $|\overline{E}|^2$ and $|\overline{H}|^2$. Figure 9.2.6 sketches how the fractional power reflected or *surface reflectivity* varies with angle of incidence θ_i for both TE and TM waves for various impedance mismatches, assuming $\mu = \mu_t$

and $\sigma = 0$ everywhere. If the wave is incident upon a medium with $\varepsilon_t > \varepsilon$, then $|\underline{\Gamma}|^2 \to 1$ as $\theta_i \to 90°$, whereas $|\underline{\Gamma}|^2 \to 1$ at the critical angle θ_c if $\varepsilon_t < \varepsilon$, and remains unity for $\theta_c < \theta < 90°$ (this θ_c case is not illustrated).
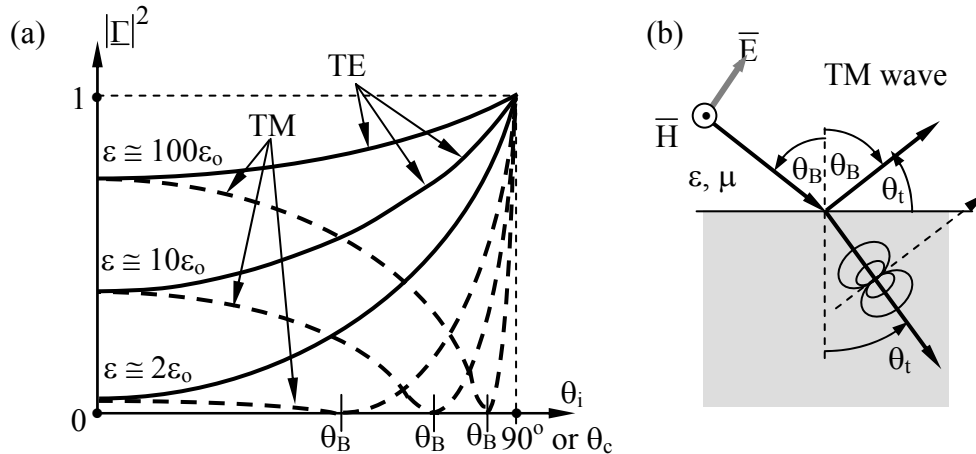
Figure 9.2.6   Power reflected from planar dielectric interfaces for $\mu = \mu_t$.

Figure 9.2.6 reveals an important phenomenon—there is perfect transmission at *Brewster's angle* θ_B for one of the two polarizations. In this case Brewster's angle occurs for the TM polarization because μ is the same everywhere and ε is not, and it would occur for TE polarization if μ varied across the boundary while ε did not. This phenomenon is widely used in glass *Brewster-angle windows* when even the slightest reflection must be avoided or when pure linear polarization is required (the reflected wave is pure).

We can compute θ_B by noting $\underline{H}_r/\underline{H}_o$ and, using (9.2.72), $\eta_n = 1$. If $\mu = \mu_t$, then (9.2.74) yields $\varepsilon_t^{0.5} \cos \theta_i = \varepsilon^{0.5} \cos \theta_t$. Snell's law for $\mu = \mu_t$ yields $\varepsilon^{0.5} \sin \theta_i = \varepsilon_t^{0.5} \sin \theta_t$. These two equations are satisfied if $\sin \theta_i = \cos \theta_t$ and $\cos \theta_i = \sin \theta_t$. Dividing this form of Snell's law by [$\cos \theta_i = \sin \theta_t$] yields: $\tan \theta_i = (\varepsilon_t/\varepsilon_i)^{0.5}$, or:

$$\theta_B = \tan^{-1} \sqrt{\varepsilon_t/\varepsilon_i} \qquad\qquad (9.2.75)$$

Moreover, dividing [$\sin \theta_i = \cos \theta_t$] by [$\cos \theta_i = \sin \theta_t$] yields $\tan \theta_B = \cos \theta_t$, which implies $\theta_B + \theta_t = 90°$. Using this equation it is easy to show that $\theta_B > 45°$ for interfaces where $\theta_t < \theta_i$, and when $\theta_t > \theta_i$, it follows that $\theta_B < 45°$.

One way to physically interpret Brewster's angle for TM waves is to note that at θ_B the polar axes of the electric dipoles induced in the second dielectric ε_t are pointed exactly at the angle of reflection mandated by phase matching, but dipoles radiate nothing along their polar axis; Figure 9.2.6(b) illustrates the geometry. That is, $\theta_B + \theta_t = 90°$. For magnetic media

magnetic dipoles are induced, and for TE waves their axes point in the direction of reflection at Brewster's angle.

Yet another way to physically interpret Brewster's angle is to note that perfect transmission can be achieved if the boundary conditions can be matched without invoking a reflected wave. This requires existence of a pair of incidence and transmission angles θ_i and θ_t such that the parallel components of both $\overline{E}$ and $\overline{H}$ for these two waves match across the boundary. Such a pair consistent with Snell's law always exists for TM waves at planar dielectric boundaries, but not for TE waves. Thus there is perfect impedance matching at Brewster's angle.

Example 9.2E

What is Brewster's angle θ_B if $\mu_2 = 4\mu_1$, and $\varepsilon_2 = \varepsilon_1$, and for which polarization would the phenomenon be observed?

Solution: If the permeabilities differ, but not the permittivities, then Brewster's angle is observed only for TE waves. At Brewster's angle $\theta_B + \theta_t = 90°$, and Snell's law

says $\dfrac{\sin\theta_t}{\sin\theta_B} = \sqrt{\dfrac{\mu}{\mu_t}}$. But $\sin\theta_t = \sin(90° - \theta_B) = \cos\theta_B$, so Snell's law becomes:

$\tan\theta_B = \sqrt{\mu_t/\mu} = 2$, and $\theta_B \cong 63°$.

9.3 *Waves guided within Cartesian boundaries*

9.3.1 Parallel-plate waveguides

We have seen in Section 9.2 that waves can be reflected at planar interfaces. For example, (9.2.14) and (9.2.16) describe the electric fields for a TE wave reflected from a planar interface at x = 0, and are repeated here:

$$\overline{E}_i = \hat{y}\underline{E}_o e^{jk_x x - jk_z z}\left[V\ m^{-1}\right] \qquad\qquad \text{(incident TE wave)} \qquad (9.3.1)$$

$$\overline{E}_r = \hat{y}\underline{E}_r{}^{-jk_x x - jk_z z} \qquad\qquad\qquad \text{(reflected TE wave)} \qquad (9.3.2)$$

Note that the subscripts i and r denote "incident" and "reflected", not "imaginary" and "real". These equations satisfy the phase-matching boundary condition that $k_z = k_{zi} = k_{zr}$. Therefore $|k_x|$ is the same for the incident and reflected waves because for both waves $k_x{}^2 + k_z{}^2 = \omega^2\mu\varepsilon$.

If the planar interface at x = 0 is a perfect conductor, then the total electric field there parallel to the conductor must be zero, implying $\underline{E}_r = -\underline{E}_o$. The superposition of these two incident and reflected waves is:

$$\overline{E}(x, z) = \hat{y}\underline{E}_o\left(e^{jk_x x} - e^{-jk_x x}\right)e^{-jk_z z} = \hat{y}2j\underline{E}_o \sin k_x x e^{-jk_z z} \qquad (9.3.3)$$

Thus $\overline{E} = 0$ in a series of parallel planes located at x = d, where:

$$k_x d = n\pi \ \text{ for } \ n = 0, 1, 2, ... \qquad \text{(guidance condition)} \qquad (9.3.4)$$

Because $k_x = 2\pi/\lambda_x$, these planes of electric-field nulls are located at $x_{nulls} = n\lambda_x/2$. A second perfect conductor could be inserted at any one of these x planes so that the waves would reflect back and forth and propagate together in the +z direction, trapped between the two conducting planes.

We can easily confirm that the boundary conditions are satisfied for the corresponding magnetic field $\overline{H}(x,z)$ by using Faraday's law:

$$\overline{H}(x,z) = -(\nabla \times \overline{E})/j\omega\mu = -(2\underline{E}_0/\omega\mu)(\hat{x}\, jk_z \sin k_x x + \hat{z}\, k_x \cos k_x x)e^{-jk_z z} \qquad (9.3.5)$$

At $x = n\lambda_x/2$ we find $\overline{H}_\perp = \underline{H}_x = 0$, so this solution is valid.

Equation (9.3.4), $k_x d = n\pi$ (n = 1, 2,...), is the *guidance condition* for parallel-plate waveguides that relates mode number to waveguide dimensions; d is the separation of the parallel plates. We can use this guidance condition to make the expressions (9.3.3) and (9.3.5) for $\overline{E}$ and $\overline{H}$ more explicit by replacing $k_x x$ with $n\pi x/d$, and k_z with $2\pi/\lambda_z$:

$$\overline{E}(x,z) = \hat{y}\, 2j\underline{E}_0 \sin\left(\frac{n\pi x}{d}\right)e^{-\frac{j2\pi z}{\lambda_z}} \qquad (\overline{E} \text{ for TE}_n \text{ mode}) \qquad (9.3.6)$$

$$\overline{H}(x,z) = -\frac{2\underline{E}_0}{\omega\mu}\left[\hat{x}\, j\frac{2\pi}{\lambda_z}\sin\left(\frac{n\pi x}{d}\right) + \hat{z}\, \frac{n\pi}{d}\cos\left(\frac{n\pi x}{d}\right)\right]e^{-\frac{j2\pi z}{\lambda_z}} \qquad (9.3.7)$$

These electric and magnetic fields correspond to a *waveguide mode* propagating in the +z direction, as illustrated in Figure 9.3.1(a) for a waveguide with plates separated by distance d. The direction of propagation can be inferred from the Poynting vector $\overline{S} = \overline{E} \times \overline{H}^*$. Because this is a TE wave and there is only one half wavelength between the two conducting plates (n = 1), this is designated the *TE$_1$ mode* of a *parallel-plate waveguide*. Because Maxwell's equations are linear, several propagating modes with different values of n can be active simultaneously and be superimposed.

These fields are periodic in both the x and z directions. The wavelength λ_z along the z axis is called the *waveguide wavelength* and is easily found using $k_z = 2\pi/\lambda_z$ where $k^2 = k_x^2 + k_z^2 = \omega^2\mu\varepsilon$:

$$\lambda_z = 2\pi\left(k^2 - k_x^2\right)^{-0.5} = 2\pi\left[(\omega/c)^2 - (n\pi/d)^2\right]^{-0.5} \qquad \text{(waveguide wavelength)} \qquad (9.3.8)$$

(a) TE$_1$ mode   (b) TM$_1$ mode
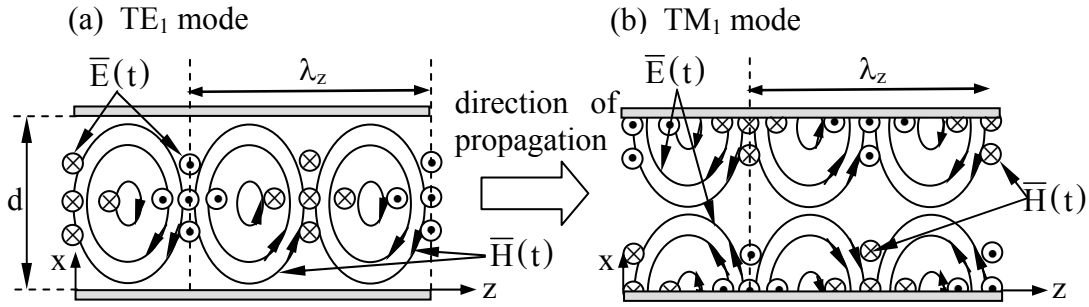
Figure 9.3.1   TE$_1$ and TM$_1$ modes of parallel-plate waveguides.

Only those TE$_n$ modes having $n < \omega d / c\pi = 2d/\lambda$ have non-imaginary waveguide wavelengths λ_z and propagate. Propagation ceases when the mode number n increases to the point where $n\pi/d \geq k \equiv 2\pi/\lambda \equiv \omega/c$.   This propagation requirement can also be expressed in terms of a minimum frequency ω of propagation, or *cut-off frequency*, for any TE mode:

$$\omega_{TEn} = n\pi c/d \qquad\qquad \text{(cut-off frequency for TE}_n \text{ mode)} \qquad (9.3.9)$$

Thus each TE$_n$ mode has a minimum frequency ω_{TEn} for which it can propagate, as illustrated in Figure 9.3.2.
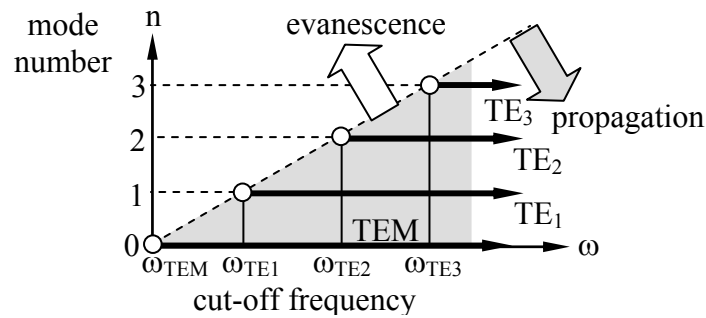
Figure 9.3.2   Propagation frequencies of TE$_n$ modes in parallel-plate waveguides.

The TE$_0$ mode has zero fields everywhere and therefore does not exist. The TM$_0$ mode can propagate even DC signals, but is identical to the TEM mode. This same relationship can also be expressed in terms of the maximum free-space wavelength λ_{TEn} that can propagate:

$$\lambda_{TE_n} = 2d/n \qquad\qquad \text{(cut-off wavelength } \lambda_{TEn} \text{ for TE}_n \text{ mode)} \qquad (9.3.10)$$

For a non-TEM wave the longest free-space wavelength λ that can propagate in a parallel-plate waveguide of width d is 2d/n.

If $\omega < \omega_{TEn}$, then we have an *evanescent wave*; k_z, (9.3.6), and (9.3.7) become:

$$k_z{}^2 = \omega^2\mu\varepsilon - k_x{}^2 < 0 \;, \quad \underline{k}_z = \pm j\alpha \tag{9.3.11}$$

$$\underline{\overline{E}}(x,z) = \hat{y}2j\underline{E}_o \sin\left(n\pi x/d\right)e^{-\alpha z} \qquad \left(\underline{\overline{E}} \text{ for } TE_n \text{ mode, } \omega <_{TEn}\right) \tag{9.3.12}$$

$$\underline{\overline{H}}(x,z) = -\left(\nabla\times\underline{\overline{E}}\right)\big/j\omega\mu = -\left(2\underline{E}_o/\omega\mu\right)\left(\hat{x}\alpha\sin k_x x + \hat{z}k_x \cos k_x x\right)e^{-\alpha z} \tag{9.3.13}$$

Such evanescent waves propagate no time average power, i.e., $R_e\{\overline{\underline{S}}_z\} = 0$, because the electric and magnetic fields are 90 degrees out of phase everywhere and decay exponentially toward zero as z increases, as illustrated in Figure 9.3.3.
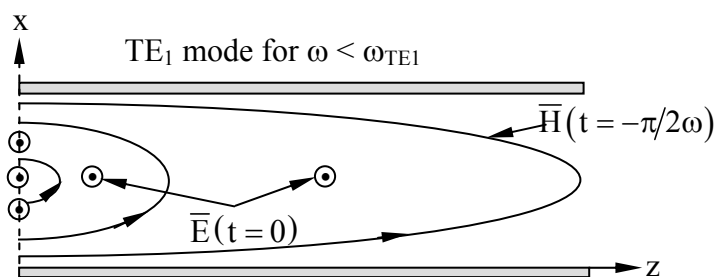
Figure 9.3.3   Evanescent TE_1 mode in a parallel-plate waveguide.

If we reflect a TM wave from a perfect conductor there again will be planar loci where additional perfectly conducting plates could be placed without violating boundary conditions, as suggested in Figure 9.3.1(b) for the TM_1 mode. Note that the field configuration is the same as for the TE_1 mode, except that $\overline{E}$ and $\overline{H}$ have been interchanged (allowed by duality) and phase shifted in the lateral direction to match boundary conditions. Between the plates the TE and TM field solutions are dual, as discussed in Section 9.2.6. Also note that TEM = TM_{01}.

Evaluation of Poynting's vector reveals that the waves in Figure 9.3.1 are propagating to the right. If this waveguide mode were superimposed with an equal-strength wave traveling to the left, the resulting field pattern would be similar, but the magnetic and electric field distributions $\overline{E}(t)$ and $\overline{H}(t)$ would be shifted relative to one another by $\lambda_z/4$ along the z axis, and they would be 90° out of phase in time; the time-average power flow would be zero, and the reactive power $I_m\{\overline{S}\}$ would alternate between inductive (+j) and capacitive (-j) at intervals of $\lambda_z/2$ down the waveguide.

Because k_z/ω is frequency dependent, the shapes of waveforms evolve as they propagate. If the signal is narrowband, this evolution can be characterized simply by noting that the envelope of the waveform propagates at the "group velocity" v_g, and the modulated sinusoidal wave inside the envelope propagates at the "phase velocity" v_p, as discussed more fully in Section 9.5.2. These velocities are easily found from $k(\omega)$: $v_p = \omega/k$ and $v_g = (\partial k/\partial\omega)^{-1}$.
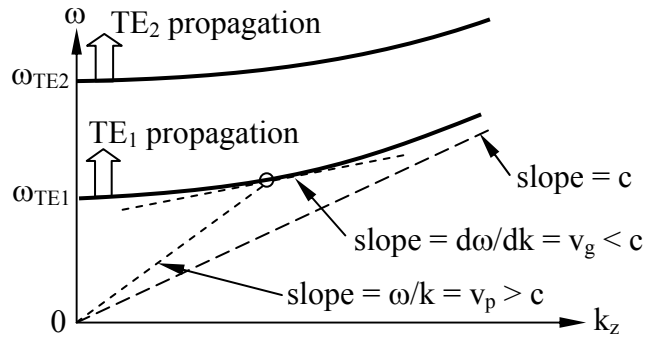
Figure 9.3.4   Dispersion relation and v_p and v_g for a parallel-plate waveguide.

The phase and group velocities of waves in parallel-plate waveguides do not equal c, as can be deduced the dispersion relation plotted in Figure 9.3.4:

$$k_z^2 = \omega^2 \mu\varepsilon - k_x^2 = \omega^2/c^2 - \left(n\pi/d\right)^2 \quad \text{(waveguide dispersion relation)} \quad (9.3.14)$$

The phase velocity v_p of a wave, which is the velocity at which the field distribution pictured in Figure 9.3.1 moves to the right, equals ω/k, which approaches infinity as ω approaches the cut-off frequency ω_{TEn} from above. The group velocity v_g, which is the velocity of energy or information propagation, equals $d\omega/dk$, which is the local slope of the dispersion relation $\omega(k)$ and can never exceed c. Both v_p and v_g approach c = $(\mu\varepsilon)^{-0.5}$ as $\omega \to \infty$.

Example 9.3A
A TM$_2$ mode is propagating in the +z direction in a parallel-plate waveguide with plate separation d and free-space wavelength λ_o. What are $\overline{E}$ and $\overline{H}$? What is λ_z? What is the decay length α_z^{-1} when the mode is evanescent and $\lambda_o = 2\lambda_{cut\,off}$?

Solution: We can superimpose incident and reflected TM waves or use duality to yield $\overline{\overline{H}}(x,y)$ = $\hat{y}\,H_o \cos k_x x\ e^{-jk_z z}$, analogous to (9.3.3), where $k_x d = 2\pi$ for the TM$_2$ mode. Therefore $k_x = 2\pi/d$ and $k_z = 2\pi/\lambda_z = (k_o^2 - k_x^2)^{0.5} = [(2\pi/\lambda_o)^2 - (2\pi/d)^2]^{0.5}$. Ampere's law yields:

$$\overline{E} = \nabla \times \overline{H}/j\omega\varepsilon = \left(\hat{x}\partial H_y/\partial z + \hat{z}H_y/\partial x\right)/j\omega\varepsilon$$

$$= \left(\hat{x}k_z \cos k_x x + \hat{z}jk_x \sin k_x x\right)\left(H_o/\omega\varepsilon\right)e^{-jk_z z}$$

The waveguide wavelength $\lambda_z = 2\pi/k_z = (\lambda_o^{-2} - d^{-2})^{0.5}$. Cutoff occurs when $k_z = 0$, or $k_o = k_x = 2\pi/d = 2\pi/\lambda_{cut\,off}$. Therefore $\lambda_o = 2\lambda_{cut\,off} \Rightarrow \lambda_o = 2d$, and $\alpha_o^{-1} = (-jk_z) = (k_x^2 - k_o^2)^{0.5} = [d^{-2} - (2d)^{-2}]^{-0.5}/2\pi = d/(3^{0.5}\pi)$ [m].

9.3.2   Rectangular waveguides

Waves can be trapped within conducting cylinders and propagate along their axis, rectangular and cylindrical waveguides being the most common examples. Consider the *rectangular waveguide* illustrated in Figure 9.3.5. The fields inside it must satisfy the wave equation:

$$\left(\nabla^2 + \omega^2\mu\varepsilon\right)\underline{\overline{E}} = 0 \qquad\qquad \left(\nabla^2 + \omega^2\mu\varepsilon\right)\underline{\overline{H}} = 0 \qquad\qquad (9.3.15)$$

where $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$.
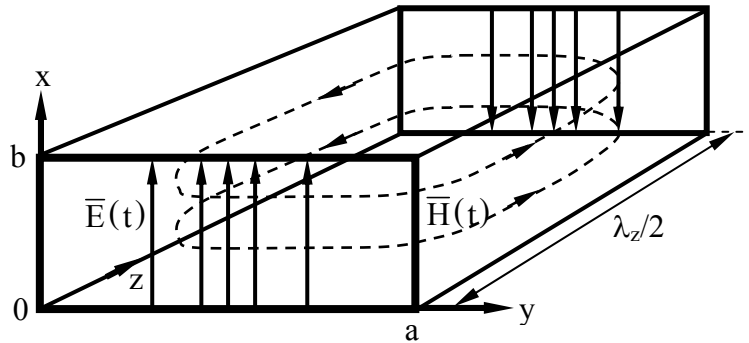
Figure 9.3.5   Dominant mode in rectangular waveguide (TE$_{10}$).

Since the wave equation requires that the second spatial derivative of $\underline{\overline{E}}$ or $\underline{\overline{H}}$ equal $\underline{\overline{E}}$ or $\underline{\overline{H}}$ times a constant, these fields must be products of sinusoids or exponentials along each of the three cartesian coordinates, or sums of such products. For example, the wave equation and boundary conditions $\left(\overline{E}_{//} = 0 \text{ at } x = 0 \text{ and } y = 0\right)$ are satisfied by:

$$E_x = \sin k_y y \left(A \sin k_x x + B \cos k_x x\right) e^{-jk_z z} \qquad\qquad (9.3.16)$$

$$E_y = \sin k_x x \left(C \sin k_y y + D \cos k_y y\right) e^{-jk_z z} \qquad\qquad (9.3.17)$$

provided that the usual dispersion relation for uniform media is satisfied:

$$k_x{}^2 + k_y{}^2 + k_z{}^2 = k_o{}^2 = \omega^2\mu\varepsilon \qquad\qquad \text{(dispersion relation)} \quad (9.3.18)$$

These field components must also satisfy the boundary condition $\overline{E}_{//} = 0$ for x = b and y = a, which leads to the guidance conditions:

$$k_x b = n\pi \qquad\qquad (9.3.19)$$

$$k_y a = m\pi \qquad\qquad \text{(guidance conditions)} \quad (9.3.20)$$

We can already compute the *cut-off frequency* of propagation ω_{mn} for each mode, assuming our conjectured solutions are valid, as shown below. Cut-off occurs when a mode becomes evanescent, i.e., when k_o^2 equals zero or becomes negative. At cut-off for the m,n mode:

$$\omega_{mn}\mu\varepsilon = k_y^2 + k_x^2 = (m\pi/a)^2 + (n\pi/b)^2 \tag{9.3.21}$$

$$\omega_{mn} = \left[(m\pi c/a)^2 + (n\pi c/b)^2\right]^{0.5} \qquad \text{(cut-off frequencies)} \tag{9.3.22}$$

Since $a \geq b$ by definition, and the 0,0 mode cannot have non-zero fields, as shown later, the lowest frequency that can propagate is:

$$\omega_{10} = \pi c/a \tag{9.3.23}$$

which implies that the longest wavelength that can propagate in rectangular waveguide is $\lambda_{max} = 2\pi c/\omega_{10} = 2a$. If $a > 2b$, the second lowest cut-off frequency is $\omega_{20} = 2\omega_{10}$, so such waveguides can propagate only a single mode over no more than an octave (factor of 2 in frequency) before another propagating mode is added. The parallel-plate waveguides of Section 9.3.1 exhibited similar properties.

Returning to the field solutions, $\underline{E}$ must also satisfy Gauss's law, $\nabla \bullet \varepsilon \underline{E} = 0$, within the waveguide, where ε is constant and $E_z \equiv 0$ for TE waves. This implies:

$$\nabla \bullet \underline{E} = 0 = \partial E_x/\partial x + \partial E_y/\partial y + \partial E_z/\partial z = \left[k_x \sin k_y y \left(A \cos k_x x\right.\right.$$
$$\left.\left. - B \sin k_x x\right) + k_y \sin k_x x \left(C \cos k_y y - D \sin k_y y\right)\right] e^{-jk_z z} \tag{9.3.24}$$

The only way (9.3.24) can be zero for all combinations of x and y is for:

$$A = C = 0 \tag{9.3.25}$$

$$k_y D = -k_x B \tag{9.3.26}$$

The electric field for TE modes follows from (9.3.17), (9.3.18), (9.3.25), and (9.3.26):

$$\underline{E} = (E_o/k_o)\left(\hat{x} k_y \sin k_y y \cos k_x x - \hat{y} k_x \sin k_x x \cos k_y y\right) e^{-jk_z z} \tag{9.3.27}$$

where the factor of k_o^{-1} was introduced so that $\underline{E}_o$ would have its usual units of volts/meter. Note that since $k_x = k_y = 0$ for the TE_{00} mode, $\underline{E} = 0$ everywhere and this mode does not exist. The corresponding magnetic field follows from $\underline{H} = -(\nabla \times \underline{E})/j\omega\mu$:

$$\overline{H} = \left(E_o/\eta k_o^2\right)\left(\hat{x}k_x k_z \sin k_x x \cos k_y y + \hat{y}k_y k_z \cos k_x x \sin k_y y\right.$$
$$\left. -j\hat{z}\,k_x^2 \cos k_x x \cos k_y y\right)e^{-jk_z z}$$

(9.3.28)

A similar procedure yields the fields for the TM modes; their form is similar to the TE modes, but with $\overline{E}$ and $\overline{H}$ interchanged and then shifted spatially to match boundary conditions. The validity of field solutions where $\overline{E}$ and $\overline{H}$ are interchanged also follows from the principle of duality, discussed in Section 9.2.6.

The most widely used rectangular waveguide mode is TE$_{10}$, often called the *dominant mode*, where the first digit corresponds to the number of half-wavelengths along the wider side of the guide and the second digit applies to the narrower side. For this mode the guidance conditions yield $k_x = 0$ and $k_y = \pi/a$, where $a \geq b$ by convention. Thus the fields (9.3.27) and (9.3.28) become:

$$\overline{E}_{TE10} = \underline{E}_o\hat{x}\left(\sin k_y y\right)e^{-jk_z z} \qquad \text{(dominant mode)} \qquad (9.3.29)$$

$$\overline{H}_{TE10} = \left(\underline{E}_o/\omega\mu\right)\left[\hat{y}k_z \sin\left(\pi y/a\right) - j\hat{z}\left(\pi/a\right)\cos\left(\pi y/a\right)\right]e^{-jk_z z} \qquad (9.3.30)$$

The fields for this mode are roughly sketched in Figure 9.3.5 for a wave propagating in the plus-z direction. The electric field varies as the sine across the width a, and is uniform across the height b of the guide; at any instant it also varies sinusoidally along z. H_y varies as the sine of y across the width b, while H_x varies as the cosine; both are uniform in x and vary sinusoidally along z.

The forms of evanescent modes are easily found. For example, the electric and magnetic fields given by (9.3.29) and (9.3.30) still apply even if $k_z = (k_o^2 - k_x^2 - k_y^2)^{0.5} \equiv -j\alpha$ so that e^{-jkz} becomes $e^{-\alpha z}$. For frequencies below cutoff the fields for this mode become:

$$\overline{E}_{TE10} = \hat{x}\underline{E}_o\left(\sin k_y y\right)e^{-\alpha z} \qquad (9.3.31)$$

$$\overline{H}_{TE10} = -j\left(\pi\underline{E}_o/\eta a k_o^2\right)\left[\hat{y}\alpha \sin\left(\pi y/a\right) + \hat{z}(\pi/a)\cos(\pi y/a)\right]e^{-\alpha z} \qquad (9.3.32)$$

The main differences are that for the evanescent wave: 1) the field distribution at the origin simply decays exponentially with distance z and the fields lose their wave character since they wax and wane in synchrony at all positions, 2) the electric and magnetic fields vary 90 degrees out of phase so that the total energy storage alternates twice per cycle between being purely electric and purely magnetic, and 3) the energy flux becomes purely reactive since the real (time average) power flow is zero. The same differences apply to any evanscent TE$_{mn}$ or TM$_{mn}$ mode.

Example 9.3B

What modes have the four lowest cutoff frequencies for a rectangular waveguide having the dimensions a = 1.2b? For the TE$_{10}$ mode, where can we cut thin slots in the waveguide walls such that they transect no currents and thus have no deleterious effect?

Solution: The cut-off frequencies (9.3.22) are: $\omega_{mn} = [(m\pi c/a)^2 + (n\pi c/b)^2]^{0.5}$, so the lowest cut-off is for TE$_{mn}$ = TE$_{10}$, since TE$_{00}$, TM$_{00}$, TM$_{01}$, and TM$_{10}$ do not exist. Next comes TE$_{11}$ and TM$_{11}$, followed by TE$_{20}$. The wall currents are perpendicular to $\overline{H}$, which has no x component for the dominant mode (9.3.30); see Figure 9.3.5. Therefore thin slots cut in the x direction in the sidewalls (y = 0,a) will never transect current or perturb the TE$_{10}$ mode. In addition, the figure and (9.3.30) show that the z-directed currents at the center of the top and bottom walls are also always zero, so thin z-directed slots at those midlines do not perturb the TE$_{10}$ mode either. Small antennas placed through thin slots or holes in such waveguides are sometimes used to introduce or extract signals.

9.3.3 Excitation of waveguide modes

Energy can be radiated inside waveguides and resonators by antennas. We can compute the energy radiated into each waveguide or resonator mode using modal expansions for the fields and matching the boundary conditions imposed by the given source current distribution $\overline{J}$.

Consider a waveguide of cross-section a×b and uniform in the z direction, where a ≥ b. If we assume the source current $\overline{J}_s$ is confined at z = 0 to a wire or current sheet in the x,y plane, then the associated magnetic fields $\overline{H}_+$ and $\overline{H}_-$ at $z = 0 \pm \delta$, respectively ($\delta\rightarrow0$), must satisfy the boundary condition (2.1.11):

$$\overline{H}_+ - \overline{H}_- = \overline{J}_s \times \hat{z} \qquad\qquad (9.3.33)$$

Symmetry dictates $\overline{H}_-(x,y) = -\overline{H}_+(x,y)$ for the x-y components of the fields on the two sides of the boundary at z = 0, assuming there are no other sources present, so:

$$\hat{z}\times\left(\overline{H}_+ - \overline{H}_-\right) = \hat{z}\times\left(\overline{J}_s \times \hat{z}\right) = \overline{J}_s = 2\hat{z}\times\overline{H}_+(x,y) \qquad\qquad (9.3.34)$$

To illustrate the method we restrict ourselves to the simple case of TE$_{m,0}$ modes, for which $\overline{E}$ and $\overline{J}_s$ are in the x direction. The total magnetic field (9.3.28) summed over all TE$_{m,0}$ modes and orthogonal to $\hat{z}$ for forward propagating waves is:

$$\overline{H}_{+total} = \hat{y} \sum_{m=0}^{\infty} \left[\underline{E}_{m,0}m\pi/\left(\eta a k_o{}^2\right)\right]k_{zm}\sin(m\pi y/a) = \left(\overline{J}_s/2\right)\times\hat{z} \qquad (9.3.35)$$

- 286 -

where $\underline{E}_m$ is the complex amplitude of the electric field for the $TE_{m,0}$ mode, and k_y has been replaced by $m\pi/a$. We can multiply both right-hand sides of (9.3.35) by $\sin(n\pi y/a)$ and integrate over the x-y plane to find:

$$\sum_{m=0}^{\infty} \left[\underline{E}_{m,0} m\pi \middle/ \left(\eta a k_o{}^2 \right) \right] k_{zm} \oiint_A \sin\left(m\pi y/a\right) \sin\left(n\pi y/a\right) \, dx \, dy$$

$$= 0.5 \oiint_A \underline{J}_s\left(x, y\right) \sin\left(n\pi y/a\right) \, dx \, dy \tag{9.3.36}$$

Because sine waves of different frequencies are orthogonal when integrated over an integral number of half-wavelengths at each frequency, the integral on the left-hand side is zero unless m = n. Therefore we have a simple way to evaluate the phase and magnitude of each excited mode:

$$\underline{E}_{n,0} = \left[\eta k_o{}^2 \middle/ n b \pi k_{zn} \right] \iint_A \underline{J}_s(x,y) \sin(n\pi y/a) \, dx \, dy \tag{9.3.37}$$

Not all excited modes propagate real power, however. Modes n with cutoff frequencies above ω are evanescent, so $k_{zn} = [k_o{}^2 - (n\pi/a)^2]^{0.5}$ is imaginary. The associated magnetic field remains in phase with $\underline{J}_s$ and real, and therefore the power in each evanescent wave is imaginary. Since all modes are orthogonal in space, their powers add; for evanescent modes the imaginary power corresponds to net stored magnetic or electric energy. The reactance at the input to the wires driving the current J_s is therefore either capacitive or inductive, depending on whether the total energy stored in the reactive modes is predominantly electric or magnetic, respectively.

A more intuitive way to understand modal excitation is to recognize that the power P delivered to the waveguide by a current distribution $\bar{J}_s$ is:

$$P = \oiiint_V \overline{E} \bullet \bar{J}_s{}^* dv \; [V] \tag{9.3.38}$$

and therefore any mode for which the field distribution $\overline{E}$ is orthogonal to $\bar{J}_s$ will not be excited, and vice versa. For example, a straight wire in the x direction across a waveguide carrying current at some frequency ω will excite all TE_{n0} modes that have non-zero $\overline{E}$ at the position of the wire; modes with cutoff frequencies above ω will contribute only reactance to the current source, while the propagating modes will contribute a real part. Proper design of the current distribution $\bar{J}_s$ can permit any combination of modes to be excited, while not exciting the rest.

9.4   Cavity resonators

9.4.1   Rectangular cavity resonators

Rectangular *cavity resonators* are hollow rectangular conducting boxes of width a, height b, and length d, where d ≥ a ≥ b by convention.   Since they are simply rectangular waveguides terminated at both ends by conducting walls, and the electric fields must still obey the wave equation, $(\nabla^2 + \omega^2\mu\varepsilon)\underline{\overline{E}} = 0$, therefore $\underline{\overline{E}}$ for TE modes must have the form of the TE waveguide fields (9.3.27), but with a sinusoidal z dependence that matches the boundary conditions at z = 0 and z = d; for example, equal forward- and backward-propagating waves would form the standing wave:

$$\overline{E} = \left(\underline{E}_o/k_o\right)\left(\hat{x}k_y\sin k_y y\cos k_x x - \hat{y}k_x\sin k_x x\cos k_y y\right)\left(\underline{A}\sin k_z z + \underline{B}\cos k_z z\right) \qquad (9.4.1)$$

where B = 0 ensures $\overline{E}_{//} = 0$ at z = 0, and $k_z = p\pi/d$ ensures it for z = d, where p = 1, 2, ...

Unlike rectangular waveguides that propagate any frequency above cut-off for the spatial field distribution (mode) of interest, cavity resonators operate only at specific *resonant frequencies* or combinations of them in order to match all boundary conditions.   The *resonant frequencies* ω_{mnp} for a rectangular cavity resonator follow from the dispersion relation:

$$\omega_{mnp}{}^2\mu\varepsilon = k_y{}^2 + k_x{}^2 + k_z{}^2 = \left(m\pi/a\right)^2 + \left(n\pi/b\right)^2 + \left(p\pi/d\right)^2 \qquad (9.4.2)$$

$$\omega_{mnp} = \left[\left(m\pi c/a\right)^2 + \left(n\pi c/b\right)^2 + \left(p\pi c/d\right)^2\right]^{0.5} \left[r\ s^{-1}\right] \qquad \text{(cavity resonances)} \qquad (9.4.3)$$

The *fundamental mode* for a cavity resonator is the lowest frequency mode. Since boundary conditions can not be met unless at least two of the quantum numbers m, n, and p are non-zero, the lowest resonant frequency is associated with the two longest dimensions, d and a. Therefore the lowest resonant frequency is:

$$\omega_{101} = \left[\left(\pi c/a\right)^2 + \left(\pi c/d\right)^2\right]^{0.5} \ \text{[radians/sec]} \qquad \text{(lowest resonance)} \qquad (9.4.4)$$

Cavity resonators are therefore sometimes filled with dielectrics or magnetic materials to reduce their resonant frequencies by reducing c.

The fields for the fundamental mode of a rectangular cavity resonator, TE_{101}, follow from (9.4.1) and Faraday's law:

$$\overline{E} = \hat{x}E_o \sin\left(\pi y/a\right)\sin\left(\pi z/d\right) \qquad \text{(fundamental waveguide mode)} \qquad (9.4.5)$$

$$\overline{H} = jE_o \left(\pi \omega c^2 / \eta \right) \left[\hat{y} \sin \left(\pi y / a \right) \cos \left(\pi z / d \right) / d - \hat{z} \cos \left(\pi y / a \right) \sin \left(\pi z / d \right) / a \right] \qquad (9.4.6)$$

The total energy w [J] = $w_e(t)$ + $w_m(t)$ in each mode m,n,p of a cavity resonator can be calculated using (2.7.28) and (2.7.29), and will decay exponentially at a rate that depends on total power dissipation P_d [W] due to losses in the walls and in any insulator filling the cavity interior:

$$w(t) \cong w_o e^{-P_d t/w} = w_o e^{-\omega t/Q} \qquad (9.4.7)$$

Wall losses and any dissipation in insulators can be estimated by integrating (9.2.60) and (2.7.30), respectively, over the volume of the cavity resonator. The energy stored, power dissipation, and Q can be quite different for different modes, and are characterized by w_{mnp}, $P_{d,mnp}$, and Q_{mnp}, respectively, as defined by either (3.5.23) or (7.4.43):

$$Q_{mnp} = \omega w_{mnp} / P_{d_{mnp}} \qquad (9.4.8)$$

Example 9.4A
What are the lowest resonant frequency and its Q for a perfectly conducting metallic cavity of dimensions a, b, d if it is filled with a medium characterized by ε, μ, and σ. Assume Q >> 1.

Solution: The lowest resonant frequency ω_{101} is given by (9.4.4), where c = $(\mu\varepsilon)^{-0.5}$: ω_{101} = $\pi(\mu\varepsilon)^{-0.5}(a^{-2} + d^{-2})^{0.5}$. Q_{101} = $\omega_{101}w_{T101}/P_{d101}$ where the total energy stored w_{T101} is twice the average electric energy stored since the total electric and magnetic energy storages are equal. At each point in the resonator the time-average electric energy density stored is $\langle W_e \rangle = \varepsilon |\overline{E}|^2 / 4$ [J m^{-3}] and the time-average power dissipated is $\sigma |\overline{E}|^2 / 2$, [W m^{-3}] so the electric-energy/dissipation density ratio everywhere is $\varepsilon/2\sigma$, and thus $w_{T101}/P_{d101} = \varepsilon/\sigma$, so $Q_{101} = \pi\varepsilon (\mu\varepsilon)^{-0.5} (a^{-2} + d^{-2})^{0.5} / \sigma$.

9.4.2   Perturbation of resonator frequencies

Often we would like to tune a resonance to some nearby frequency. This can generally be accomplished by changing the shape of the resonator slightly. Although the relationship between shape and resonant frequency can be evaluated using Maxwell's equations, a simpler and more physical approach is taken here.

The energy stored in a resonator can be regarded as a population of N trapped photons at frequency f bouncing about inside. Since the energy E per photon is hf (1.1.10), the total energy in the resonator is:

$$w_T = Nhf \quad [\ ] \qquad (9.4.9)$$

If we force the walls of a resonator to move slowly toward its new shape, they will move either opposite to the forces imposed by the electromagnetic fields inside, or in the same direction, and thereby do positive or negative work, respectively, on those fields. If we do positive work, then the total electromagnetic energy w_T must increase. Since the number of photons remains constant if the shape change is slow compared to the frequency, positive work on the fields results in increased electromagnetic energy and frequency f. If the resonator walls move in the direction of the applied electromagnetic forces, the externally applied work on the fields is negative and the energy and resonant frequency decrease.

The paradigm above leads to a simple expression for the change in resonant frequency of any resonator due to small physical changes. Consider the case of an air-filled metallic cavity of any shape that is perturbed by pushing in or out the walls slightly in one or more places. The electromagnetic force on a conductor has components associated with both the attractive electric and repulsive magnetic pressures on conductors given by (4.1.15) and (4.1.23), respectively. For sinusoidal waves these pressures are:

$$P_e = -\varepsilon_o \left|\underline{E}_o\right|^2 / 4 \left[N\ m^{-2}\right] \qquad \text{(electric pressure)} \qquad (9.4.10)$$

$$P_m = \mu_o \left|\underline{H}_o\right|^2 / 4 \left[N\ m^{-2}\right] \qquad \text{(magnetic pressure)} \qquad (9.4.11)$$

But these pressures, except for the negative sign of P_e (corresponding to attraction), are the electric and magnetic energy densities [J m^{-3}].

The work Δw done in moving the cavity boundary slightly is the pressure $P_{e/m}$ applied, times the area over which it is applied, times the distance moved perpendicular to the boundary. For example, Δw equals the inward electromagnetic pressure ($\pm$ energy density) times the increase in volume added by the moving boundary. But this increase in total stored electromagnetic energy is simply:

$$\Delta w_T = Nh\,\Delta f = -\left(P_e + P_m\right)\Delta v_{\text{olume}} = \Delta w_e - \Delta w_m \qquad (9.4.12)$$

The signs for the increases in electric and magnetic energy storage Δw_e and Δw_m and pressures P_e and P_m are different because the pressures P_e and P_m are in opposite directions, where $\Delta w_e = W_e \Delta v_{ol}$, and $\Delta w_m = -P_m \Delta v_{ol} = -W_m \Delta v_{ol}$. Δw_e is defined as the electric energy stored in the increased volume of the cavity, Δv_{ol}, assuming the electric field strength remains constant as the wall moves slightly; Δw_m is defined similarly. The main restriction here is that the walls cannot be moved so far that the force density on the walls changes, nor can their shape change abruptly for the same reason. For example, a sharp point concentrates electric fields and would violate this constraint.

Dividing (9.4.12) by $w_T = Nhf$ yields the frequency perturbation equation:

$$\Delta w_T / w_T = \Delta f / f = \left(\Delta w_e - \Delta w_m \right) / w_T = \Delta v_{ol} \left(W_e - W_m \right) / w_T$$

$$\text{(frequency perturbation)} \tag{9.4.13}$$

A simple example illustrates its use. Consider a rectangular cavity resonator operating in the TE_{101} mode with the fields given by (5.4.37) and (5.4.38). If we push in the center of the top or bottom of the cavity where $\overline{\underline{H}} \cong 0$ and $\overline{E} \neq 0$ we are reducing the volume allocated to electric energy storage, so Δw_e is negative and the resonant frequency will drop in accord with (9.4.13). If we push in the sides, however, the resonant frequency will increase because we are reducing the volume where magnetic energy is stored and Δw_m is negative; the electric energy density at the sidewalls is zero. In physical terms, pushing in the top center where the electric fields pull inward on the wall means that those fields are doing work on the moving wall and therefore lose energy and frequency. Pushing in where the magnetic fields are pushing outward does work on the fields, increasing their energy and frequency. This technique can be used to determine experimentally the unknown resonant mode of a cavity as well as tuning it.

9.5   Waves in complex media

9.5.1   Waves in anisotropic media

There are many types of media that can be analyzed simply using Maxwell's equations, which characterize media by their permittivity ε, permeability μ, and conductivity σ. In general ε, η, and σ can be complex, frequency dependent, and functions of field direction. They can also be functions of density, temperature, field strength, and other quantities. Moreover they can also couple $\overline{E}$ to $\overline{B}$, and $\overline{H}$ to $\overline{D}$. In this section we only treat the special cases of anisotropic media (Section 9.5.1), dispersive media (Section 9.5.2), and plasmas (Section 9.5.3). Lossy media were treated in Sections 9.2.4 and 9.2.5.

Anisotropic media, by definition, have permittivities, permeabilities, and/or conductivities that are functions of field direction. We can generally represent these dependences by 3×3 matrices (tensors), i.e.:

$$\overline{D} = \overline{\overline{\varepsilon}} \overline{E} \tag{9.5.1}$$

$$\overline{B} = \overline{\overline{\mu}} \overline{H} \tag{9.5.2}$$

$$\overline{J} = \overline{\overline{\sigma}} \overline{E} \tag{9.5.3}$$

For example, (9.5.1) says:

$$
\begin{aligned}
D_x &= \varepsilon_{xx} E_x + \varepsilon_{xy} E_y + \varepsilon_{xz} E_z \\
D_y &= \varepsilon_{yx} E_x + \varepsilon_{yy} E_y + \varepsilon_{yz} E_z \\
D_z &= \varepsilon_{zx} E_x + \varepsilon_{zy} E_y + \varepsilon_{zz} E_z
\end{aligned}
\tag{9.5.4}
$$

Most media are symmetric so that $\varepsilon_{ij} = \varepsilon_{ji}$; in this case the matrix $\overline{\overline{\varepsilon}}$ can always be diagonalized by rotating the coordinate system to define new directions x, y, and z that yield zeros off-axis:

$$\overline{\overline{\varepsilon}} = \begin{pmatrix} \varepsilon_x & 0 & 0 \\ 0 & \varepsilon_y & 0 \\ 0 & 0 & \varepsilon_z \end{pmatrix} \qquad (9.5.5)$$

These new axes are called the *principal axes* of the medium. The medium is isotropic if the permittivities of these three axes are equal, *uniaxial* if only two of the three axes are equal, and *biaxial* if all three differ. For example, tetragonal, hexagonal, and rhombohedral crystals are uniaxial, and orthorhombic, monoclinic, and triclinic crystals are biaxial. Most constitutive tensors are symmetric (they equal their own transpose), the most notable exception being permeability tensors for magnetized media like plasmas and ferrites, which are hermetian[49] and not discussed in this text.

One immediate consequence of anisotropic permittivity and (9.5.4) is that $\overline{D}$ is generally no longer parallel to $\overline{E}$, as suggested in Figure 9.5.1 for a uniaxial medium. When $\varepsilon_{xx} \neq \varepsilon_{zz}$, $\overline{E}$ and $\overline{D}$ are parallel only if they lie along one of the principal axes. As explained shortly, this property of uniaxial or biaxial media can be used to convert any wave polarization into any other.
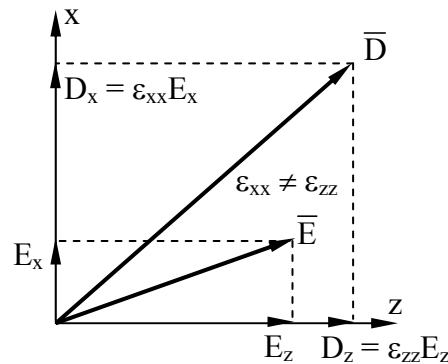
Figure 9.5.1 $\overline{D}$ and $\overline{E}$ in an anisotropic medium.

The origins of anisotropy in media are easy to understand in terms of simple models for crystals. For example, an isotropic cubic lattice becomes uniaxial if it is compressed or stretched along one of those axes, as illustrated in Figure 9.5.2(a) for z-axis compression. That such compressed columns act to increase the effective permittivity in their axial direction can be understood by noting that each of these atomic columns functions like columns of dielectric between capacitor plates, as suggested in Figure 9.5.2(b). Parallel-plate capacitors were

---

[49] Hermetian matrices equal the complex conjugate of their transpose.

discussed in Section 3.1.3. Alternatively the same volume of dielectric could be layered over one of the capacitor plates, as illustrated in Figure 9.5.2(c).
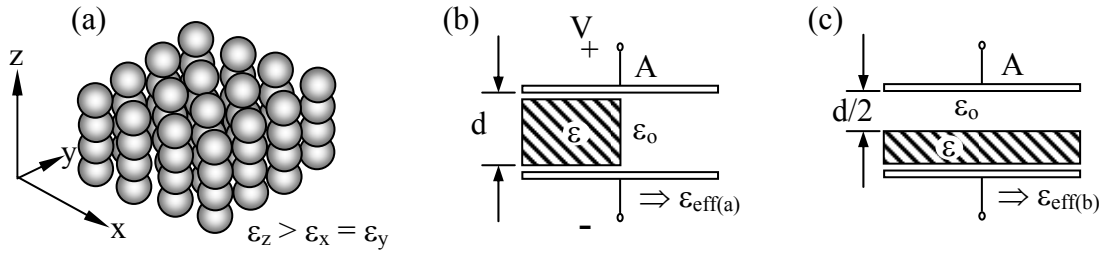
Figure 9.5.2   Uniaxial crystal and anisotropically filled capacitors.

Even though half the volume between the capacitor plates is occupied by dielectric in both these cases, the capacitance for the columns [Figure 9.5.2(b)] is greater, corresponding to a greater effective permittivity ε_{eff}. This can be shown using Equation (3.1.10), which says that a parallel plate capacitor has $C = \varepsilon_{eff}A/d$, where A is the plate area and d is the distance between the plates. The capacitances C_a and C_b for Figures 9.5.2(b) and (c) correspond to two capacitors in parallel and series, respectively, where:

$$C_a = \left(\varepsilon A/2d\right) + \left(\varepsilon_o A/2d\right) = \left(\varepsilon + \varepsilon_o\right)A/2d = \varepsilon_{eff(a)}A/d \tag{9.5.6}$$

$$C_b = \left[\left(\varepsilon A2/d\right)^{-1} + \left(\varepsilon_o A2/d\right)^{-1}\right]^{-1} = \left[\varepsilon\varepsilon_o/\left(\varepsilon + \varepsilon_o\right)\right]2A/d = \varepsilon_{eff(b)}A/d \tag{9.5.7}$$

In the limit where $\varepsilon \gg \varepsilon_o$ the permittivity ratio $\varepsilon_{eff(a)}/\varepsilon_{eff(b)} \rightarrow \varepsilon/4\varepsilon_o > 1$. In all compressive cases $\varepsilon_{eff(a)} \geq \varepsilon_{eff(b)}$. If the crystal were stretched rather than compressed, this inequality would be reversed. Exotic complex materials can exhibit inverted behavior, however.

Since the permittivity here interacts directly only with $\overline{\overline{E}}$, not $\overline{\overline{H}}$, the velocity of propagation $c = 1/\sqrt{\mu\varepsilon}$ depends only on the permittivity in the direction of $\overline{\overline{E}}$. We therefore expect slower propagation of waves linearly polarized so that $\overline{\overline{E}}$ is parallel to an axis with higher values of ε. We can derive this behavior from the source-free Maxwell's equations and the matrix constitutive relation (9.5.4).

$$\nabla \times \overline{\overline{E}} = -j\omega\mu\overline{\overline{H}} \tag{9.5.8}$$

$$\nabla \times \overline{\overline{H}} = j\omega\overline{\overline{D}} \tag{9.5.9}$$

$$\nabla \cdot \overline{\overline{D}} = 0 \tag{9.5.10}$$

$$\nabla \cdot \overline{\overline{B}} = 0 \tag{9.5.11}$$

Combining the curl of Faraday's law (9.5.8) with Ampere's law (9.5.9), as we did in Section 2.3.3, yields:

$$\nabla \times \left(\nabla \times \overline{E} \right) = \nabla \left(\nabla \bullet \overline{E} \right) - \nabla^2 \overline{E} = \omega^2 \mu \overline{D} \tag{9.5.12}$$

We now assume, and later prove, that $\nabla \bullet \overline{E} = 0$, so (9.5.12) becomes:

$$\nabla^2 \overline{E} + \omega^2 \mu \overline{D} = 0 \tag{9.5.13}$$

This expression can be separated into independent equations for each axis. Waves propagating in the z direction are governed by the x and y components of (9.5.13):

$$\left[\left(\partial^2 / \partial z^2 \right) + \omega^2 \mu \varepsilon_x \right] \underline{E}_x = 0 \tag{9.5.14}$$

$$\left[\left(\partial^2 / \partial z^2 \right) + \omega^2 \mu \varepsilon_y \right] \underline{E}_y = 0 \tag{9.5.15}$$

The wave equation (9.5.14) characterizes the propagation of x-polarized waves and (9.5.15) characterizes y-polarized waves; their wave velocities are $(\mu\varepsilon_x)^{-0.5}$ and $(\mu\varepsilon_y)^{-0.5}$, respectively. If $\varepsilon_x \neq \varepsilon_y$ then the axis with the lower velocity is called the "slow" axis, and the other is the "fast" axis. This dual-velocity phenomenon is called *birefringence*. That our assumption $\nabla \bullet \overline{E} = 0$ is correct is easily seen by noting that the standard wave solution for both x- and y- polarized waves satisfies these constraints. Since ∇ is distributive, the equation is also satisfied for arbitrary linear combinations of x- and y- polarized waves, which is the most general case here.

If a wave has both x- and y-polarized components, the polarization of their superposition will evolve as they propagate along the z axis at different velocities. For example, a linearly wave polarized at 45 degrees to the x and y axes will evolve into elliptical and then circular polarization before evolving back into linear polarization orthogonal to the input.

This ability of a birefringent medium to transform polarization is illustrated in Figure 9.5.3. In this case we can represent the linearly polarized wave at z = 0 as:

$$\overline{E}(z = 0) = E_o \left(\hat{x} + \hat{y} \right) \tag{9.5.16}$$

If the wave numbers for the x and y axes are k_x and k_y, respectively, then the wave at position z will be:

$$\overline{E}(z) = E_o e^{-jk_x z} \left(\hat{x} + \hat{y} e^{j(k_x - k_y)z} \right) \tag{9.5.17}$$

The phase difference between the x- and y-polarized components of the electric field is therefore $\Delta\phi = (k_x - k_y)z$. As suggested in the figure, circular polarization results when the two

components are 90 degrees out of phase ($\Delta\phi = \pm 90^o$), and the orthogonal linear polarization results when $\Delta\phi = 180^o$.
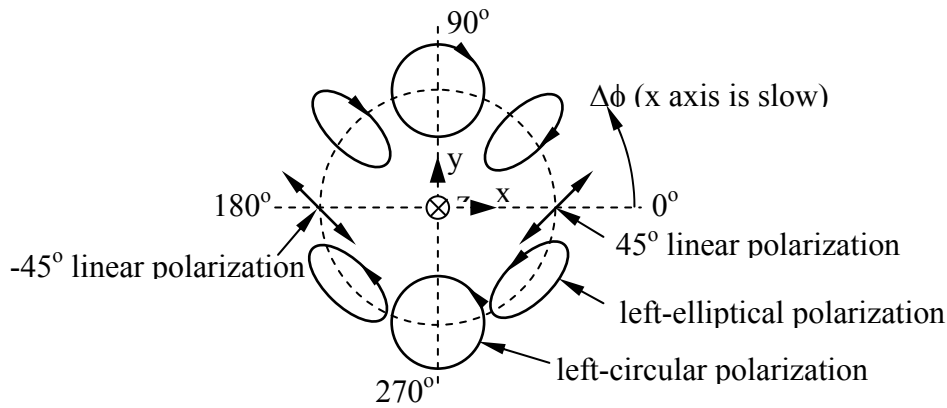
Figure 9.5.3   Polarization conversion in a birefringent medium.

Polarization conversion is commonly used in optical systems to convert linear polarization to circular, or vice-versa, via a *quarter-wave plate* for which $\Delta\phi$ is 90°, equivalent to a quarter wavelength. A *half-wave plate* ($\Delta\phi = 180^o$) reverses the sense of any polarization.

Example 9.5A

A certain birefringent medium is characterized by μ_o, $\varepsilon_x = 2\varepsilon_o$, $\varepsilon_y = 2.002\varepsilon_o$. How thick D must a quarter-wave plate be if $\lambda = 5 \times 10^{-7}$[m] in free space (visible light)? At what thickness D' might this same plate rotate appropriate linear polarization 90 degrees?

Solution: The phase lags along the x and y axes arise from $e^{-jk_x D}$ and $e^{-jk_y D}$, respectively, and the difference is $\pi/2 = (k_y - k_x)D$ for a quarter-wave plate. But $k_i = \omega(\mu_o\varepsilon_i)^{0.5}$, so $(k_y - k_x)D = \omega(\mu_o\varepsilon_x)^{0.5}[(1 + 1.001)^{0.5} - 1] D \cong (\omega/c_x)^{0.5}\, 0.0005\, D = \pi/2$. Since $\omega/c_x = 2\pi/\lambda_x$, therefore $D = 2000\lambda_x/4$ where $\lambda_x = 5\times10^{-7}(\varepsilon_o/\varepsilon_x)^{0.5}$. Thus $D = 500\lambda_x = 0.18$ mm, which is approximately the thickness of a Vu-Graph transparency that acts as a quarter-wave plate. A differential phase lag of π yields 90° polarization rotation for waves linearly polarized at an angle 45° from the principal axes x and y, so the thickness would be doubled to ~0.36 mm.

9.5.2   Waves in dispersive media

Dispersive media have wave velocities that are frequency dependent due to the frequency dependence of μ, ε, or σ. These frequency dependencies arise in all materials because of the non-instantaneous physical responses of electrons to fields. Often these time lags are so brief that only at optical frequencies do they become a significant fraction of a period, although propagation over sufficiently long paths can introduce significant cumulative differences in effects across any frequency band or gap. Only vacuum is essentially non-dispersive.

The principal consequence of dispersion is that narrowband pulse signals exhibit two velocities, the *phase velocity* v_p of the sinusoids within the pulse envelope, and the *group velocity* v_g at which the pulse envelope, energy, and information propagate. Because energy and information travel at the group velocity, it never exceeds the velocity of light although phase velocity frequently does.

A simple way to reveal this phenomenon is to superimpose two otherwise identical sinusoidal waves propagating at slightly different frequencies, $\omega \pm \Delta\omega$; superposition is valid because Maxwell's equations are linear. The corresponding wave numbers are $k \pm \Delta k$, where $\Delta k \ll k$ and $\Delta\omega \ll \omega$. Such a superposition for two sinusoids propagating in the +z direction is:

$$\begin{aligned} E(t,z) &= E_o \cos[(\omega+\Delta\omega)t-(k+\Delta k)z]+E_o \cos[(\omega-\Delta\omega t)-(k-\Delta k)z] \\ &= E_o 2\cos(\omega t-kz)\cos(\Delta\omega t-\Delta kz) \end{aligned}$$

(9.5.18)

where we used the identity $\cos\alpha + \cos\beta = 2\cos[(\alpha+\beta)/2]\cos[(\alpha-\beta)/2]$. The first factor on the right-hand side of (9.5.18) is a sine wave propagating at the center frequency ω at the phase velocity:

$$v_p = \omega/k \qquad \text{(phase velocity)} \qquad (9.5.19)$$

The second factor is the low-frequency, long-wavelength modulation envelope that propagates at the group velocity $v_g = \Delta\omega/\Delta k$, which is the slope of the $\omega(k)$ *dispersion relation*:

$$v_g = \partial\omega/\partial k = (\partial k/\partial\omega)^{-1} \qquad \text{(group velocity)} \qquad (9.5.20)$$

Figure 9.5.4(a) illustrates the original sinusoids plus their superposition at two points in time, and Figure 9.5.4(b) illustrates the corresponding dispersion relation.
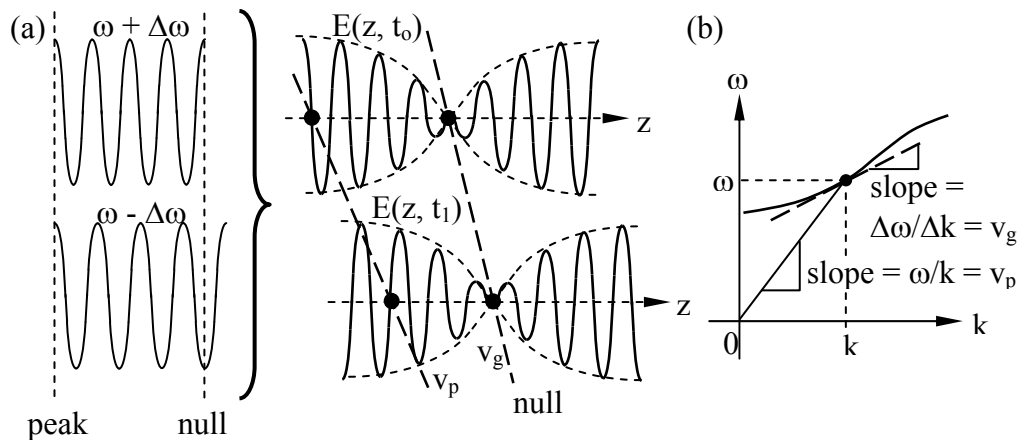
Figure 9.5.4   Phase and group velocity for two superimposed sinusoids.

Note that this dispersion relation has a phase velocity that approaches infinity at the lowest frequencies, which is what happens in plasmas near the plasma frequency, as discussed in the next section.

Communications systems employ finite-duration pulses with Fourier components at all frequencies, so if such pulses travel sufficiently far even the envelope with its finite bandwidth will become distorted. As a result dispersive media are either avoided or compensated in most communications system unless the bandwidths are sufficiently narrow. Compensation is possible because dispersion is a linear process so inverse filters are readily designed. Section 12.2.2 discusses dispersion further in the context of optical fibers.

Example 9.5B

When $\omega = c_o^{0.5}$, what are the phase and group velocities v_p and v_g in a medium having the dispersion relation $k = \omega^2/c_o$?

Solution: $v_p = \omega/k = c_o/\omega = c_o^{0.5}$ [m s^{-1}]. $v_g = (\partial k/\partial\omega)^{-1} = c_o/2\omega = c_o^{0.5}/2$ [m s^{-1}].

9.5.3   Waves in plasmas

A *plasma* is a charge-neutral gaseous assembly of atoms or molecules with enough free electrons to significantly influence wave propagation. Examples include the ionosphere[50], the sun, interiors of fluorescent bulbs or nuclear fusion reactors, and even electrons in metals or electron pairs in superconductors. We can characterize fields in plasmas once we know their permittivity ε at the frequency of interest.

To compute the permittivity of a non-magnetized plasma we recall (2.5.8) and (2.5.13):

$$\overline{D} = \varepsilon\overline{E} = \varepsilon_o\overline{E} + \overline{P} = \varepsilon_o\overline{E} + nq\overline{d} \qquad (9.5.21)$$

where $q = -e$ is the electron charge, $\overline{d}$ is the mean field-induced displacement of the electrons from their equilibrium positions, and n_3 is the number of electrons per cubic meter. Although positive ions are also displaced, these displacements are generally negligible in comparison to those of the electrons because the electron masses m_e are so much less. We can take the mass m_i of the ions into account simply by replacing m_e in the equations by m_r, the reduced mass of the electrons, where it can be shown that $m_r = m_e m_i/(m_e + m_i) \cong m_e$.

To determine ε in (9.5.21) for a collisionless plasma, we merely need to solve Newton's law for $\overline{d}(t)$, where the force $\overline{f}$ follows from (1.2.1):

$$\overline{f} = q\overline{E} = m\overline{a} = m(j\omega)^2\overline{d} \qquad (9.5.22)$$

---

[50] The terrestrial ionosphere is a partially ionized layer at altitudes ~50-5000 km, depending primarily upon solar ionization. Its peak electron density is ~10^{12} electrons m^{-3} at 100-300 km during daylight.

Solving (9.5.22) for $\overline{d}$ and substituting it into the expression for $\overline{P}$ yields:

$$\overline{P} = nq\overline{d} = -ne\overline{d} = \overline{E}ne^2 m^{-1}(j\omega)^{-2} \qquad (9.5.23)$$

Combining (9.5.21) and (9.5.23) yields:

$$\overline{D} = \varepsilon_o \overline{E} + \overline{P} = \varepsilon_o \left[1 + ne^2 / m(j\omega)^2 \varepsilon_o\right]\overline{E} = \varepsilon_o\left[1 - \omega_p^2 / \omega^2\right]\overline{E} = \varepsilon\overline{E} \qquad (9.5.24)$$

where ω_p is defined as the *plasma frequency*:

$$\omega_p \equiv \left(ne^2 / m\varepsilon_o\right)^{0.5} \quad \left[\text{radians s}^{-1}\right] \qquad (9.5.25)$$

The plasma frequency is the natural frequency of oscillation of a displaced electron or cluster of electrons about their equilibrium location in a neutral plasma, and we shall see that the propagation of waves above and below this frequency is markedly different.

The dispersion relation for a collisionless non-magnetic plasma is thus:

$$k^2 = \omega^2 \mu\varepsilon = \omega^2 \mu_o \varepsilon_o \left(1 - \omega_p^2 / \omega^2\right) \qquad (9.5.26)$$

which is plotted as $\omega(k)$ in Figure 9.5.5 together with the slopes representing the phase and group velocities of waves in plasmas.
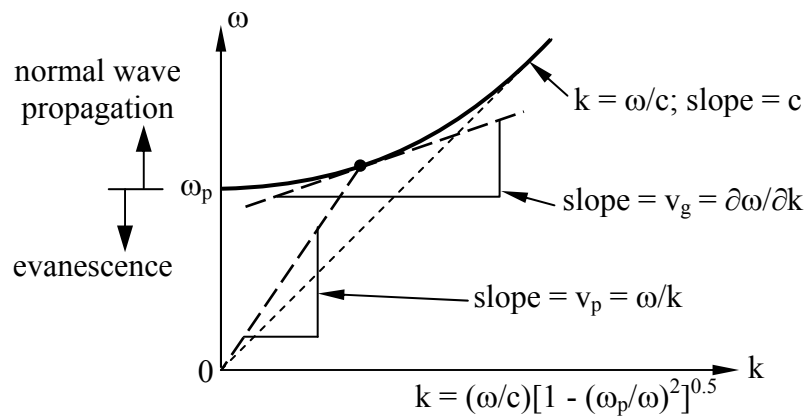
Figure 9.5.5   Dispersion relation and velocities for a simple plasma.

Using the expressions (9.5.19) and (9.5.20) for phase and group velocity we find for plasmas:

$$v_p = \omega/k = c\left[1 - \left(\omega_p / \omega\right)^2\right]^{-0.5} \qquad (9.5.27)$$

- 298 -

$$v_g = (\partial k/\partial\omega)^{-1} = c\left[1-\left(\omega_p/\omega\right)^2\right]^{0.5} \tag{9.5.28}$$

Since $v_p v_g = c^2$, and since $v_g \leq c$, it follows that v_p is alway equal to or greater than c. However, for $\omega < \omega_p$ we find v_p and v_g become imaginary because normal wave propagation is replaced by another behavior.

From (9.5.26) we see that when $\omega < \omega_p$:

$$\underline{k} = \frac{\omega}{c}\sqrt{1-(\omega_p/\omega)^2} = \pm j\alpha \tag{9.5.29}$$

$$\underline{\eta} = \sqrt{\frac{\mu}{\varepsilon}} = \eta_o \,/\, \sqrt{1-\left(\omega_p/\omega\right)^2} = \mp j\frac{\mu_o\omega}{\alpha} \tag{9.5.30}$$

Therefore an x-polarized wave propagating in the +z direction would be:

$$\overline{E} = \hat{x}E_o e^{-\alpha z} \tag{9.5.31}$$

$$\overline{H} = \hat{y}\underline{\eta}^{-1}E_o e^{-\alpha z} = j\hat{y}\left(\alpha/\omega\mu_o\right)E_o e^{-\alpha z} \tag{9.5.32}$$

where the sign of $\pm j\alpha$ was chosen (-) to correspond to exponential decay of the wave rather than growth. We find from (9.5.32) that H(t) is delayed 90° behind E(t), that both decay exponentially with z, and that the Poynting vector $\overline{S}$ is purely imaginary:

$$\overline{S} = \overline{E} \times \overline{H}^* = -j\hat{z}\left(\alpha|\underline{E}_o|^2\big/\omega\mu_o\right)e^{-2\alpha z} \tag{9.5.33}$$

Such an *evanescent wave* decays exponentially and carry only reactive power and no time-average power because of the time orthogonality of E and H. Reactive power implies that below ω_p the average energy stored is predominantly electric, but in this case the stored energy is actually dominated by the kinetic energy of the electrons. It is this extra energy that allows the permittivity ε to become negative below ω_p although μ_o remains constant. The frequency ω_p below which conversion from propagation to evanescence occurs is called the *cut-off frequency*, which is the plasma frequency here.

Example 9.5C
What is the plasma frequency f_p [Hz]of the ionosphere when $n_e = 10^{12}$ m^{-3}?

Solution: $f_p = \omega_p/2\pi = (10^{12}e^2/m\varepsilon_o)^{0.5}/2\pi \cong 9.0$ MHz.

Chapter 10: Antennas and Radiation

10.1   Radiation from charges and currents

10.1.1   Introduction to antennas and radiation

An antenna is a device that couples currents to electromagnetic waves for purposes of radiation or reception. The process by which antennas radiate can be easily understood in terms of the way in which accelerating charged particles or time-varying currents radiate, which is discussed in Section 10.1. The expressions for radiated electromagnetic fields derived in Section 10.1.4 are simple extensions of those derived in Sections 10.1.2 and 10.1.3 for the fields produced by static charges and currents, respectively.

Using the basic expressions for radiation derived in Section 10.1, simple short dipole antennas are shown in Section 10.2 to have stable directional properties far from the antenna (the antenna far field), and different directional properties closer than $\sim\lambda/2\pi$ (the antenna near field). In Section 10.3 these properties are related to basic metrics that characterize each antenna, such as gain, effective area, and impedance. These metrics are then related to the performance of various communications systems. Antenna arrays are discussed in Section 10.4, followed by aperture and more complicated wire antennas in Sections 11.1 and 11.2, respectively.

10.1.2   Electric fields around static charges

One simple way to generate electromagnetic waves is to vibrate electric charges, creating time-varying current. The equation characterizing this radiation is very similar to that characterizing the electric fields produced by a single static charge, which is developed below. Section 10.1.3 extends this result to magnetic fields produced by moving charges.

Faraday's and Gauss's laws for static charges in vacuum are:

$$\nabla \times \overline{E} = 0 \tag{10.1.1}$$

$$\nabla \bullet \varepsilon_o \overline{E} = \rho \tag{10.1.2}$$

Since the curl of $\overline{E}$ is zero, $\overline{E}$ can be the gradient of any arbitrary scalar function $\Phi(\bar{r})$ and still satisfy (10.1.1). That is:

$$\overline{E}(\bar{r}) = -\nabla\Phi(\bar{r}) \tag{10.1.3}$$

where Φ is the *scalar electric potential* and is in units of *Volts*. The negative sign is consistent with $\overline{E}$ pointing away from regions of high potential and toward lower potentials. Note that (10.1.3) satisfies (10.1.1) because $\nabla \times (-\nabla\Phi) \equiv 0$ is an identity, and that a simple three-dimensional scalar field Φ fully characterizes the three-dimensional vector electric field $\overline{E}(\bar{r})$. It

is therefore often easiest to find the electric potential $\Phi(\bar{r})$ before computing the electric field produced by static source charges.

If the charge q [Coulombs] is spherically symmetric, both Φ and $\overline{E}$ must also be spherically symmetric. The only way a vector field can be spherically symmetric is for it to be directed radially, so:

$$\overline{E} = \hat{r}E_r(r) \tag{10.1.4}$$

where r is the radius from the origin where the charge is centered and $E_r(r)$ is the radial field. We can now relate $\overline{E}$ to q by applying Gauss's divergence theorem (2.4.6) to the volume integral of Gauss's law (10.1.2):

$$\oiiint_V \left(\nabla \bullet \varepsilon_o \overline{E} \right) dv = \oiiint_V \rho \, dv = q = \oiint_A \varepsilon_o \overline{E} \bullet \hat{n} \, da$$
$$= \oiint_A \varepsilon_o \hat{n} \bullet \hat{r} E_r(r) \, da = 4\pi r^2 \varepsilon_o E_r(r) \tag{10.1.5}$$

Therefore the electric field produced by a charge q at the origin is:

$$\overline{E}(\bar{r}) = \hat{r}E_r(r) = \hat{r}q / 4\pi\varepsilon_o r^2 = -\nabla\Phi \quad \left[Vm^{-1} \right] \tag{10.1.6}$$

To find the associated scalar potential Φ we integrate (10.1.6) using the definition of the *gradient operator*:

$$\nabla\Phi \equiv \left(\hat{x}\frac{\partial}{\partial x} + \hat{y}\frac{\partial}{\partial y} + \hat{z}\frac{\partial}{\partial z} \right)\Phi \qquad \text{(gradient in Cartesian coordinates)} \quad (10.1.7)$$

$$= \left[\hat{r}\frac{\partial}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial}{\partial\phi} \right]\Phi \qquad \text{(gradient in spherical coordinates)} \quad (10.1.8)$$

Since the spherically symmetric potential Φ (10.1.6) is independent of θ and ϕ, it follows that $\partial/\partial\theta = \partial/\partial\phi = 0$ and Equation (10.1.8) becomes:

$$\nabla\Phi = \hat{r}\partial\Phi/\partial r \tag{10.1.9}$$

This mathematical simplification occurs only in spherical coordinates, not Cartesian. Substitution of (10.1.9) into (10.1.6), followed by integration of (10.1.6) with respect to radius r, yields:

$$\Phi(\bar{r}) = \int \left(q / 4\pi\varepsilon_o r^2 \right) dr = \Phi_o + q / 4\pi\varepsilon_o r = q / \left(4\pi\varepsilon_o |\bar{r}| \right) \tag{10.1.10}$$

where we define as zero the electric potential Φ_o contributed by any charge infinitely far away.

The solution for the electric potential Φ due to charge q at some position $\bar{r}_q$ other than the origin follows from (10.1.10):

$$\Phi(\bar{r}) = q\big/\left(4\pi\varepsilon_o \left|\bar{r} - \bar{r}_q\right|\right) = q\big/\left(4\pi\varepsilon_o r_{pq}\right) \quad [V] \tag{10.1.11}$$

which can alternatively be written using subscripts p and q to refer to the locations $\bar{r}_p$ and $\bar{r}_q$ of the person (or observer) and the charge, respectively, and r_{pq} to refer to the distance $\left|\bar{r}_p - \bar{r}_q\right|$ between them.

If we replace the charge q with a charge density ρ_q in the infinitesimal volume dv, then we can integrate (10.1.11) over the source region to obtain the total static *electric potential* produced by an arbitrary charge distribution ρ_q:

$$\Phi_p = \oiiint_{V_q} \left[\rho_q\big/\left(4\pi\varepsilon_o r_{pq}\right)\right] dv \quad [V] \qquad \text{(scalar Poisson integral)} \tag{10.1.12}$$

This integration to find Φ_p can be performed because Maxwell's equations are linear so that superposition applies. Thus we have a simple way to compute Φ_p and $\overline{E}$ for any arbitrary static charge density distribution ρ_q. This *scalar Poisson integral* for the potential function Φ is similar to that found for dynamic charge distributions in the next section. The integral (10.1.12) is also a solution to the *Poisson equation*:

$$\nabla^2 \Phi = -\rho/\varepsilon_o \qquad \text{(Poisson equation)} \tag{10.1.13}$$

which follows from computing the divergence of Gauss's law:

$$\nabla \bullet \{\nabla \Phi = -\overline{E}\} \Rightarrow \nabla^2 \Phi = -\nabla \bullet \overline{E} = -\rho/\varepsilon_o \tag{10.1.14}$$

Poisson's equation reduces to *Laplace's equation,* $\nabla^2\Phi = 0$, when $\rho = 0$.

10.1.3   Magnetic fields around static currents

Maxwell's equations governing static magnetic fields in vacuum are:

$$\nabla \times \overline{H} = \overline{J} \qquad \text{(static Ampere's law)} \tag{10.1.15}$$

$$\nabla \bullet \mu_o \overline{H} = 0 \qquad \text{(Gauss's law)} \tag{10.1.16}$$

Because the divergence of $\overline{H}$ is always zero, we can define the magnetic flux density in vacuum as being:

- 303 -

$$\overline{B} = \mu_o \overline{H} = \nabla \times \overline{A} \qquad\qquad (10.1.17)$$

where $\overline{A}$ is defined as the *magnetic vector potential*, which is a vector analog to Φ. This very general expression for $\overline{B}$ always satisfies Gauss's law: $\nabla \bullet (\nabla \times \overline{A}) \equiv 0$.

Substituting (10.1.17) into Ampere's law (10.1.15) results in:

$$\nabla \times (\nabla \times \overline{A}) = \mu_o \overline{J} \qquad\qquad (10.1.18)$$

This can be simplified using the vector identity:

$$\nabla \times (\nabla \times \overline{A}) \equiv \nabla (\nabla \bullet \overline{A}) - \nabla^2 \overline{A} \qquad\qquad (10.1.19)$$

where we note that $\nabla \bullet \overline{A}$ is arbitrary and does not impact any of our prior equations; therefore we set it equal to zero. Then (10.1.18) becomes the *vector Poisson equation*:

$$\nabla^2 \overline{A} = -\mu_o \overline{J} \qquad\qquad \text{(vector Poisson equation)} \quad (10.1.20)$$

The three vector components of (10.1.20) are each scalar Poisson equations identical to (10.1.13) except for the constant, so the solution is nearly identical to (10.1.12) once the constants have been reconciled; this solution is:

$$\overline{A}_p = \oiiint_{V_q} \left[\mu_o \overline{J}_q / (4\pi r_{pq}) \right] dv \ \left[\text{V s m}^{-1} \right] \qquad\qquad (10.1.21)$$

Thus we have a simple way to compute $\overline{A}$ and therefore $\overline{B}$ for any arbitrary static current distribution $\overline{J}_q$.

10.1.4   Electromagnetic fields produced by dynamic charges

In the static case of Section 10.1.2 it was very helpful to define the potential functions $\overline{A}$ and Φ, and the time-dependent Maxwell's equations for vacuum permit us to do so again:

$$\nabla \times \overline{E} = -\partial \overline{B} / \partial t \qquad\qquad \text{(Faraday's law)} \quad (10.1.22)$$

$$\nabla \times \overline{H} = \overline{J} + \partial \overline{D} / \partial t \qquad\qquad \text{(Ampere's law)} \quad (10.1.23)$$

$$\nabla \bullet \overline{E} = \rho / \varepsilon_o \qquad\qquad \text{(Gauss's law)} \quad (10.1.24)$$

$$\nabla \bullet \overline{B} = 0 \qquad\qquad \text{(Gauss's law)} \quad (10.1.25)$$

Although the curl of $\overline{E}$ is no longer zero so that $\overline{E}$ no longer equals the gradient of some potential Φ, we can satisfy $\nabla \cdot \overline{B} = 0$ if we define a vector potential $\overline{A}$ such that:

$$\overline{B} = \nabla \times \overline{A} = \mu_o \overline{H} \tag{10.1.26}$$

This definition of $\overline{A}$ always satsifies Gauss's law: $\nabla \cdot (\nabla \times \overline{A}) \equiv 0$. Substituting $\nabla \times \overline{A}$ for $\overline{B}$ in Faraday's law yields:

$$\nabla \times \overline{E} = -\partial (\nabla \times \overline{A}) / \partial t \tag{10.1.27}$$

Rearranging terms yields:

$$\nabla \times (\overline{E} + \partial \overline{A} / \partial t) = 0 \tag{10.1.28}$$

which implies that the quantity $(\overline{E} + \partial \overline{A} / \partial t)$ can be the gradient of any potential function Φ:

$$\overline{E} + \partial \overline{A} / \partial t = -\nabla \Phi \tag{10.1.29}$$

$$\overline{E} = -(\partial \overline{A} / \partial t + \nabla \Phi) \tag{10.1.30}$$

Thus dynamic electric fields have two components—one due to the instantaneous value of $\Phi(t)$, and one proportional to the time derivative of $\overline{A}$.

We can now use the vector identity (10.1.19) to simplify Ampere's law after $(\nabla \times \overline{A}) / \mu_o$ replaces $\overline{H}$:

$$\nabla \times (\nabla \times \overline{A}) = \mu_o (\overline{J} + \partial \overline{D} / \partial t) = \nabla (\nabla \cdot \overline{A}) - \nabla^2 \overline{A} \tag{10.1.31}$$

By substituting (10.1.30) in (10.1.31) for $\overline{D} = \varepsilon_o \overline{E}$ and grouping terms we obtain:

$$\nabla^2 \overline{A} - \nabla (\nabla \cdot \overline{A} + \mu_o \varepsilon_o \partial \Phi / \partial t) - \mu_o \varepsilon_o \partial^2 \overline{A} / \partial t^2 = -\mu_o \overline{J} \tag{10.1.32}$$

In the earlier static case we let $\nabla \cdot \overline{A} = 0$ because specifying the curl of a vector field $(\overline{B} = \nabla \times \overline{A})$ does not constrain its divergence, which can be independently chosen[51]. Here we can let:

$$\nabla \cdot \overline{A} = -\mu_o \varepsilon_o \partial \Phi / \partial t \tag{10.1.33}$$

---

[51] Let $\overline{A} = \nabla \Phi + \nabla \times \overline{N}$; then $\nabla \times \overline{A} = \nabla \times (\nabla \times \overline{N})$ and $\nabla \cdot \overline{A} = \nabla^2 \Phi$, so $\nabla \times \overline{A}$ and $\nabla \cdot \overline{A}$ can be chosen independently simply by choosing $\overline{N}$ and Φ independently.

This reduces (10.1.32) to a simple equation by eliminating its second term, yielding:

$$\nabla^2 \overline{A} - \mu_o \varepsilon_o \partial^2 \overline{A} \big/ \partial t^2 = -\mu_o \overline{J} \qquad (10.1.34)$$

which is called the inhomogeneous vector *Helmholtz equation* (the homogeneous version has no source term on the right hand side; $\overline{J} = 0$). It is a wave equation for $\overline{A}$ driven by the current source $\overline{J}$.

A similar inhomogeneous wave equation relating the electric potential Φ to the charge distribution ρ can also be derived. Substituting (10.1.30) into Gauss's law (10.1.24) yields:

$$\nabla \bullet \overline{E} = -\nabla \bullet \left(\partial \overline{A} \big/ \partial t + \nabla \Phi \right) = -\partial (\nabla \bullet \overline{A}) \big/ \partial \tau - \nabla^2 \Phi \qquad (10.1.35)$$

Replacing $\nabla \bullet \overline{A}$ using (10.1.33) then produces:

$$\nabla \bullet \overline{E} = \mu_o \varepsilon_o \partial^2 \Phi \big/ \partial t^2 - \nabla^2 \Phi = \rho \big/ \varepsilon_o \qquad (10.1.36)$$

which is more commonly written as the inhomogeneous scalar Helmholtz equation:

$$\nabla^2 \Phi - \mu_o \varepsilon_o \partial^2 \Phi \big/ \partial t^2 = -\rho \big/ \varepsilon_o \qquad (10.1.37)$$

analogous to the vector version (10.1.34) for $\overline{A}$. These inhomogeneous scalar and vector Helmholtz equations, (10.1.34) and (10.1.37), permit us to calculate the electric and magnetic potentials and fields produced anywhere in vacuum as a result of arbitrary source charges and currents, as explained below.

The solutions to the Helmholtz equations must reduce to: a) the traveling-wave solutions [e.g., (2.2.9)] for the wave equation [e.g., (2.2.7)] when the source terms are zero, and b) the static solutions (10.1.10) and (10.1.21) when $\partial/\partial t = 0$. The essential feature of solutions to wave equations is that their separate dependences on space and time must have the same form because their second derivatives with respect to space and time are identical within a constant multiplier. These solutions can therefore be expressed as an arbitrary function of a single argument that sums time and space, e.g. (z - ct) or (t - r_{pq}/c). The solutions must also have the form of the static solutions because they reduce to them when the source is static. Thus the solutions to the Helmholtz inhomogeneous equations are the static solutions expressed in terms of the argument (t - r_{pq}/c):

$$\Phi_p = \iiint_{V_q} \left[\rho_q \left(t - r_{pq} \big/ c \right) \big/ \left(4 \pi \varepsilon_o r_{pq} \right) \right] dv \quad [V] \qquad (10.1.38)$$

$$\overline{A}_p = \iiint_{V_q} \left[\mu_o \overline{J}_q \left(t - r_{pq} \big/ c \right) \big/ \left(4 \pi r_{pq} \right) \right] dv \quad \left[Vsm^{-1} \right] \qquad (10.1.39)$$

These solutions are the dynamic scalar *Poisson integral* and the dynamic vector Poisson integral, respectively. Note that Φ_p and $\overline{A}_p$ depend on the state of the sources at some time in the past, not on their instantaneous values. The delay r_{pq}/c is the ratio of the distance r_{pq} between the source and observer, and the velocity of light c. That is, r_{pq}/c is simply the propagation time between source and observer.

10.2   Short dipole antennas

10.2.1   Radiation from Hertzian dipoles

Since Maxwell's equations are linear, superposition applies and therefore the electromagnetic field produced by an arbitrary current distribution is simply the integral of the fields produced by each infinitesimal element. Thus the electromagnetic field response to an infinitesimal current element is analogous to the impulse response of a linear circuit, and comparably useful for calculating responses to arbitrary stimuli.

The simplest infinitesimal radiating element, called a *Hertzian dipole*, is a current element of length d carrying I(t) amperes. Conservation of charge requires charge reservoirs at each end of the current element containing $\pm q(t)$ coulombs, where I = dq/dt, as illustrated in Figure 10.2.1(a). The total charge is zero. If we align the z axis with the direction of the current and assume the cross-sectional area of the current element is A_c [m^2], then the current density within the element is:

$$\overline{J}_q(t) = \hat{z} I(t) \Big/ A_c \quad \left[\text{Am}^{-2} \right]$$

(10.2.1)

Substituting this current density into the expression (10.1.39) for vector potential yields:

$$\overline{A}_p = \hat{z} \oiiint_V \left[\mu_o I \left(t - \frac{r_{pq}}{c} \right) \Big/ \left(A_c 4\pi r_{pq} \right) \right] dv = \hat{z} \frac{\mu_o d}{4\pi r_{pq}} I \left(t - \frac{r_{pq}}{c} \right) \quad [\text{Vs/m}]$$

(10.2.2)

where integration over the volume V of the current element yielded a factor of $A_c d$.

To obtain simple expressions for the radiated electric and magnetic fields we must now switch to: 1) time-harmonic representations because radiation is frequency dependent, and 2) polar coordinates because the symmetry of the radiation is polar, not cartesian, as suggested in Figure 10.2.1(b). The time harmonic form of I(t - r_{pq}/c) is $Ie^{-jkr_{pq}}$ and the polar form of $\hat{z}$ is $\hat{r}\cos\theta - \hat{\theta}\sin\theta$, so (10.2.2) becomes:

$$\underline{\overline{A}}_p = \left(\hat{r}\cos\theta - \hat{\theta}\sin\theta \right) \mu_o \underline{I} d e^{-jkr_{pq}} \Big/ 4\pi r_{pq} \quad \left[\text{Vsm}^{-1} \right]$$

(10.2.3)
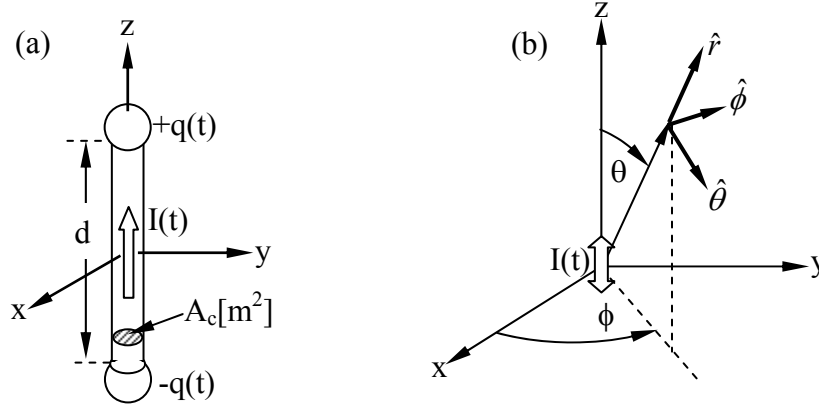
Figure 10.2.1   Hertzian dipole in spherical coordinates.

To find $\overline{H}$ and $\overline{E}$ radiated by this current element, we need to compute the curl of $\overline{A}$ in spherical coordinates:

$$\overline{H} = \left(\nabla \times \overline{A}\right)/\mu_o = \left(\mu_o r^2 \sin\theta\right)^{-1} \det \begin{vmatrix} \hat{r} & r\hat{\theta} & r\sin\theta\hat{\phi} \\ \partial/\partial r & \partial/\partial\theta & \partial/\partial\phi \\ A_r & rA_\theta & r\sin\theta A_\phi \end{vmatrix} \qquad (10.2.4)$$

Since $\overline{A}$ is independent of position ϕ (so $\partial/\partial\phi = 0$) and has no ϕ component, (10.2.4) becomes:

$$\overline{H} = \hat{\phi}\left(jk\underline{I}d/4\pi r\right)e^{-jkr}\left[1 + \left(jkr\right)^{-1}\right]\sin\theta \qquad (10.2.5)$$

After some computation the radiated electric field can be found from (10.2.5) using Ampere's law (2.3.17):

$$
\begin{aligned}
\overline{E} &= \left(\nabla \times \overline{H}\right)/j\omega\varepsilon_o \\
&= j\frac{k\underline{I}d\eta_o}{4\pi r}e^{-jkr}\left\{\hat{r}\left[\frac{1}{jkr} + \frac{1}{\left(jkr\right)^2}\right]2\cos\theta + \hat{\theta}\left[1 + \frac{1}{jkr} + \frac{1}{\left(jkr\right)^2}\right]\sin\theta\right\}
\end{aligned} \qquad (10.2.6)
$$

These solutions (10.2.5–6) for the Hertzian dipole are fundamental because they permit us to calculate easily the radiation from arbitrary current sources. It suffices to know the source current distribution because it uniquely determines the charge distribution via conservation of charge (2.1.19), and therefore the charge does not radiate independently.

These solutions for $\overline{E}$ and $\overline{H}$ are polynomials in $1/jkr$, so they have two asymptotes—one for large values of kr and one for small values. When kr is very large the lowest order terms dominate, so kr = $2\pi r/\lambda \gg 1$, or:

$$r \gg \lambda/2\pi \qquad\qquad\qquad\qquad\qquad \text{(far field)} \quad (10.2.7)$$

which we call the *far field* of the dipole.

In the far field the expressions (10.2.5) and (10.2.6) for $\overline{H}$ and $\overline{E}$ simplify to:

$$\overline{E} = \hat{\theta}\,\frac{jk\underline{I}d\eta_o}{4\pi r}\,e^{-jkr}\sin\theta \qquad\qquad \text{(far-field electric field)} \quad (10.2.8)$$

$$\overline{H} = \hat{\phi}\left(jk\underline{I}de^{-jkr}\sin\theta\right)\!\big/4\pi r \qquad\qquad \text{(far-field magnetic field)} \quad (10.2.9)$$

These expressions are identical, except that $\overline{E}$ points in the θ direction while $\overline{H}$ points in the orthogonal ϕ direction; the radial components are negligible in the far field. Also:

$$|\overline{E}| = |\overline{H}|\,\eta_o \qquad\qquad\qquad\qquad\qquad\qquad (10.2.10)$$

where the impedance of free space $\eta_o = \sqrt{\mu_o/\varepsilon_o} \cong 377$ ohms. We found similar orthogonality and proportionality for uniform plane waves in Sections 2.3.2 and 2.3.3.

We can calculate the radiated intensity in the far field using (2.7.41) and the field expressions (10.2.8) and (10.2.9):

$$\langle \overline{S}(t) \rangle = 0.5\mathrm{R}_e\left[\underline{\overline{S}}\right] = 0.5\mathrm{R}_e\left[\overline{E} \times \overline{H}^*\right] \qquad\qquad (10.2.11)$$

$$\langle \overline{S}(t) \rangle = \hat{r}\,|\overline{E}_o|^2\big/2\eta_o = \hat{r}\,(\eta_o/2)\,|kId/4\pi r|^2 \sin^2\theta \ \left[\mathrm{W\ m^{-2}}\right] \qquad (10.2.12)$$

The *radiation pattern* $\langle S(t,\theta) \rangle$ for a Hertzian dipole is therefore a donut-shaped figure of revolution about its z axis, as suggested in Figure 10.2.2(b).
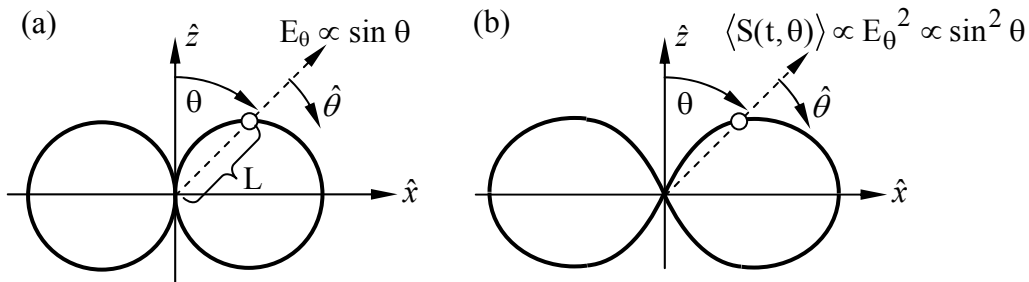
Figure 10.2.2 Electric field strength E_θ and power $\langle S(t,\theta) \rangle$ radiated by a Hertzian dipole.

Hertzian dipoles preferentially radiate laterally, with zero radiation along their axis. The electric field strength E_θ varies as $\sin\theta$, which yields a circle in a polar plot, as illustrated in Figure 10.2.2(a). The distance L in the plot corresponds to $|\underline{E}_\theta|$. Since $\sin^2 45^o = 0.5$, the width of the beam between half-power points in the θ direction is 90°.

The total power P_R radiated by a Hertzian dipole can be calculated by integrating the radial component $\langle S_r \rangle$ of $\langle \overline{S}(t,\theta) \rangle$ over all directions:

$$P_R = \int_0^{2\pi} d\phi \int_0^\pi \langle S_r \rangle r^2 \sin\theta \, d\theta = \pi\eta_o \left| k\underline{I}d/4\pi \right|^2 \int_0^\pi \sin^3\theta \, d\theta$$

$$= \left(\eta_o/12\pi \right) \left| k\underline{I}d \right|^2 \cong 395 \, (Id/\lambda)^2 \quad [W] \tag{10.2.13}$$

Thus the radiated fields increase linearly with $\underline{I}d/\lambda$ and the total radiated power increases as the square of this factor, i.e. as $\left| \underline{I}d/\lambda \right|^2$. Since the electric and magnetic fields are in phase with each other in the far field, the imaginary power $I_m \{\underline{\overline{S}}\} = 0$ there.

Example 10.2A
Equation (10.2.13) says the current I input to a Hertzian dipole radiates P_R watts. What value resistor R_r would dissipate the same power for the same I?

Solution: $P_R = \left(\eta_o/12\pi \right) \left| k\underline{I}d \right|^2 = \left| \underline{I} \right|^2 R_r \big/ 2 \Rightarrow R_r = \left(2\pi\eta_o/3 \right)(d/\lambda)^2$ ohms; this quantity is often called the radiation resistance of the radiator.

10.2.2   Near fields of a Hertzian dipole

If we examine the *near fields* radiated by a Hertzian dipole close to the origin where $kr << 1$, then the $(kr)^{-2}$ terms in the expression (10.2.7) for $\overline{E}$ dominate all other terms for both $\overline{E}$ (10.2.6) and $\overline{H}$ (10.2.5), so we are left primarily with $\overline{E}$:

$$\overline{\underline{E}} \cong \left(\underline{I}d/j\omega 4\pi\varepsilon_o r^3 \right)\left(\hat{r}2\cos\theta + \hat{\theta}\sin\theta \right) \tag{10.2.14}$$

This is simply the electric field produced by a static *electric dipole* of length d with a *dipole moment* of $\underline{p} = \underline{q}d$, where the charges at the ends of the dipole are $\pm q$ and $\underline{I} = j\omega\underline{q}$ to conserve charge.

Substituting this definition of $\underline{p}$ into (10.2.14) yields:

$$\overline{\underline{E}} \cong \left(\underline{p}/4\pi\varepsilon_o r^3 \right)\left(\hat{r}2\cos\theta + \hat{\theta}\sin\theta \right) \qquad \text{(near-field electric field)} \tag{10.2.15}$$

The dominant term for $\overline{H}$ in the near field is:

$$\overline{H} \cong \hat{\phi}\left(j\omega p/4\pi r^2\right)\sin\theta \qquad \text{(near-field magnetic field)} \quad (10.2.16)$$

Because $\overline{\overline{S}} = \overline{E} \times \overline{H}^*$ is purely negative imaginary in the near fields, these fields correspond to reactive power and stored electric energy. Integrating the exact expressions for $\overline{\overline{S}}$ over 4π steradians yields a real part that is independent of r; that is, the total power radiated is the same (10.2.13) regardless of the radius r at which we integrate, even in the near field.

A simple expression for $\overline{H}$ in the near field of the source is called the *Biot-Savart law*; it easily follows from the expression (10.2.5) for magnetic fields close to a current element $\underline{I}d\hat{z}$ when kr << 1:

$$\overline{H} = \hat{\phi}(\underline{I}d\sin\theta)/4\pi r^2 \qquad (10.2.17)$$

The Biot-Savart law relates arbitrary current distributions $\overline{J}(\overline{r}',t)$ to $\overline{H}(\overline{r},t)$ when the distance $r = |\overline{r}' - \overline{r}| << \lambda/2\pi$; it therefore applies to static current distributions as well. But $\underline{I}d$ here is simply $\int_V \underline{J}\ dx\ dy\ dz$, where the current $\underline{I}$ and current density $\underline{J}$ are in the z direction and V is the volume of the sources of $\overline{r}'$. With these substitutions (10.2.17) becomes:

$$\overline{H}(\overline{r}) = \iiint_V \frac{\underline{J}\hat{\phi}\sin\theta}{4\pi|\overline{r}-\overline{r}'|^2}\ dx\ dy\ dz \qquad (10.2.18)$$

We can also use the definition of vector cross product to replace $\hat{\phi}\underline{J}\sin\theta$:

$$\hat{\phi}\underline{J}\sin\theta = \frac{\overline{J}\times(\overline{r}-\overline{r}')}{|\overline{r}-\overline{r}'|} \qquad (10.2.19)$$

Substituting (10.2.19) into (10.2.18) in their time-domain forms yields the Biot-Savart law:

$$\overline{H}(\overline{r},t) = \iiint_{V'} \frac{\overline{J}\times(\overline{r}-\overline{r}')}{4\pi|\overline{r}-\overline{r}'|^3}\ dx\ dy\ dz \qquad (10.2.20)$$

Equation (10.2.20) has been generalized to $\overline{H}(\overline{r},t)$ because $\overline{H}$ is independent of frequency if $|r - r'| << \lambda/2\pi$.

10.2.3 Short dipole antennas

Antennas transform freely propagating electromagnetic waves into circuit voltages for reception, and also transform such voltages into free-space waves for transmission. They are used for wireless communications, power transmission, or surveillance at wavelengths ranging from micrometers (infrared and visible wavelengths) to hundreds of kilometers. Their sophistication and performance continue to increase as improved computational and fabrication methods are developed, although simple structures still dominate today.

Determining the fields and currents associated with a given antenna can be difficult using traditional approaches to boundary value problems because many waves must usually be superimposed in order to match boundary conditions, even when well chosen orthogonal wave expansions other than plane-waves are used. Fortunately modern computer software tools can handle most antenna problems. Here we take a traditional alternative approach to antenna analysis that yields acceptable solutions for most common configurations by making one key assumption—that the current distribution on the antenna is known. Determination of antenna current distributions is discussed in Sections 11.1–2.

Arbitrary antenna current distributions can be approximated by superimposing infinitesimal Hertzian dipole radiators that have constant current $\bar{I}$ over an infinitesimal length d. The electric far fields each dipole radiates are given by (10.2.8), and the total radiated field $\bar{E}$ is simply the sum of these differential contributions. Superposition of these fields is valid because Maxwell's equations are linear. $\bar{H}$ can then be readily found using Faraday's law or direct integration. This is the approach taken here; the far fields of the short dipole antenna are found by integrating the contributions from each infinitesimal element of that dipole. From these fields the antenna gain, effective area, and circuit properties can then be found, as discussed in Section 10.3. Many practical antennas, such as those used in many cars for the ~1-MHz Amplitude-Modulated (AM) band, are approximately short-dipole antennas with lengths less than a few percent of the associated wavelength λ. Their simple behavior provides an easy introduction to antenna analysis.

Consider the *short-dipole antenna* illustrated in Figure 10.2.3. It has length d $<< \lambda$ and is driven by a current source with $\underline{I}_o$ amperes at angular frequency ω [radians/second]. Complex notation is used here for simplicity because antenna characteristics are frequency-dependent. As discussed later, the wires, or "feed-lines", providing current to the dipole do not alter the dipole's radiated fields because those wires are perpendicular to the antenna fields and also do not radiate.

The electric far field $\bar{\underline{E}}_{ff}{}'$ radiated by an infinitesimal current element $\underline{I}$ of length δ is given by (10.2.8):

$$\bar{\underline{E}}_{ff}{}' = \hat{\theta}\frac{jk\underline{I}\delta\eta_o}{4\pi r}e^{-jkr}\sin\theta \qquad\qquad \text{(far-field electric field)} \quad (10.2.21)$$

This expression can be integrated over the contributions from all infinitesimal elements of the current distribution on the short dipole to find the total radiated electric far field $\bar{\underline{E}}_{ff}$:

$$\overline{E}_{ff} = \int_{-d/2}^{d/2} \overline{E}_{ff}'(r',\theta)dz = \frac{jk\eta_o}{4\pi} \int_{-d/2}^{d/2} \hat{\theta}' \left[\underline{I}(z) e^{-jkr'} \frac{\sin\theta'}{r'} \right] dz \qquad (10.2.22)$$
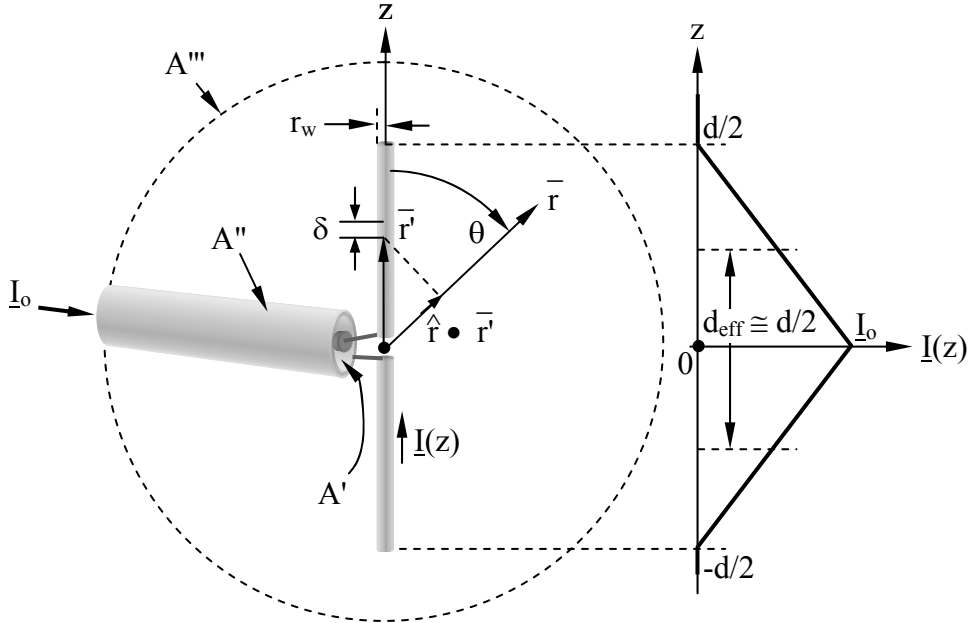
Figure 10.2.3   Short-dipole antenna.

This integral can be simplified if the observer is far from the antenna relative to its length d so that $\theta' = \theta$, $\hat{\theta}' \cong \hat{\theta}$, and $r'^{-1} \cong r^{-1}$. In addition, if $d << \lambda/2\pi$ for all z, then:

$$e^{-jkr'} \cong e^{-jkr} \qquad (10.2.23)$$

$$\begin{aligned}
\overline{E}_{ff} &\cong \hat{\theta} j\frac{k\eta_o}{4\pi r} \sin\theta e^{-jkr} \int_{-d/2}^{d/2} \underline{I}(z)dz \\
&= \hat{\theta} j\frac{k\eta_o \underline{I}_o d_{eff}}{4\pi r} \sin\theta\ e^{-jkr}
\end{aligned} \qquad \text{(far-field radiation)} \quad (10.2.24)$$

where the *effective length* of the dipole d_{eff} is illustrated in Figure 10.2.3 and is defined as:

$$d_{eff} \equiv \underline{I}_o^{-1} \int_{-d/2}^{d/2} \underline{I}(z)dz \qquad \text{(effective length of short dipole)} \quad (10.2.25)$$

Because both these far fields $\overline{E}_{ff}$ and the near fields are perpendicular to the x-y plane where the feed line is located, they are consistent with the boundary conditions associated with a sufficiently small conducting feed line located in that plane, and no reflected waves are

produced. Moreover, if the feed line is a coaxial cable with a conducting sheath, Poynting's vector on its outer surface is zero so it radiates no power.

The far fields radiated by a short dipole antenna are thus radially propagating θ-polarized plane waves with ϕ-directed magnetic fields $\overline{H}$ of magnitude $|\overline{E}|/\eta_o$. The time-average intensity $\overline{P}$ of these radial waves is given by Poynting's vector:

$$\overline{P} = \frac{1}{2} R_e \{\overline{S}\} = \frac{1}{2} R_e \{\overline{E} \times \overline{H}^*\} = \hat{r} \frac{|\overline{E}_{ff}|^2}{2\eta_o} \quad [\text{Wm}^{-2}] \tag{10.2.26}$$

$$\overline{P} = \hat{r} \frac{\eta_o}{2} \left(\frac{k|\underline{I}_o d_{eff}|}{4\pi r} \right)^2 \sin^2 \theta = \hat{r} \frac{\eta_o}{2} \left| \frac{\underline{I}_o d_{eff}}{\lambda 2r} \right|^2 \sin^2 \theta \tag{10.2.27}$$

This angular distribution of radiated power is illustrated in Figure 10.2.4.
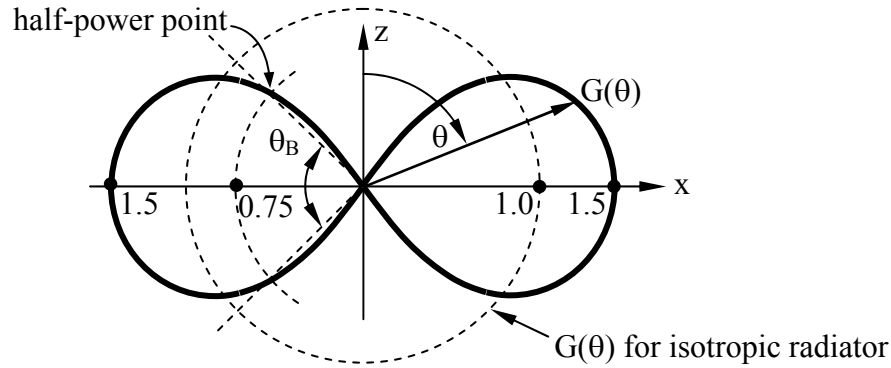
Figure 10.2.4   Antenna gain G(θ) for a short dipole or Hertzian antenna.

The total power radiated is the integral of this intensity over 4π steradians[52]:

$$P_T = \int_{4\pi} \left[\overline{P}(r,\theta) \bullet \hat{r} \right] r^2 \sin \theta \, d\theta \, d\phi = \frac{\eta_o \pi}{3} \left| \frac{\underline{I}_o d_{eff}}{\lambda} \right|^2 \quad [\text{W}] \qquad \text{(radiated power)} \tag{10.2.28}$$

10.3   Antenna gain, effective area, and circuit properties

10.3.1   Antenna directivity and gain

The far-field intensity $\overline{P}(r,\theta)$ [W m^{-2}] radiated by any antenna is a function of direction, as given for a short dipole antenna by (10.2.27) and illustrated in Figure 10.2.4. *Antenna gain*

---

[52] Recall $\int_0^{\pi/2} \sin^n x \, dx = [2\bullet4\bullet6...(n-1)]/[1\bullet3\bullet5...(n)]$ for n odd; $(\pi/2)[1\bullet3\bullet5...(n-1)]/[2\bullet4\bullet6...(n)]$ for n even.

$G(\theta,\phi)$ is defined as the ratio of the intensity $P(\theta,\phi,r)$ to the intensity [Wm^{-2}] that would result if the same total power available at the antenna terminals, P_A [W], were radiated isotropically over 4π steradians. $G(\theta,\phi)$ is often called "gain over isotropic" where:

$$G(\theta,\phi) \equiv \frac{P(r,\theta,\phi)}{\left(P_A/4\pi r^2\right)} \qquad \text{(antenna gain definition)} \qquad (10.3.1)$$

A related quantity is *antenna directivity* $D(\theta,\phi)$, which is normalized to the total power radiated P_T rather than to the power P_A available at the antenna terminals:

$$D(\theta,\phi) \equiv \frac{P(r,\theta,\phi)}{\left(P_T/4\pi r^2\right)} \qquad \text{(antenna directivity definition)} \qquad (10.3.2)$$

The transmitted power is less than the available power if the antenna is mismatched or lossy. Since the total power radiated is $P_T = r^2 \int_{4\pi} P(r,\theta,\phi)\sin\theta \, d\theta \, d\phi$, a useful relation follows from (10.3.2):

$$\oint_{4\pi} D(\theta,\phi)\sin\theta \, d\theta \, d\phi = 4\pi \qquad (10.3.3)$$

Equation (10.3.3) says that if the directivity or gain is large in one direction, it must be correspondingly diminished elsewhere, as suggested in Figure 10.2.4, where the pattern is plotted relative to an isotropic radiator and exhibits its "main lobe" in the direction $\theta = 90°$. This pattern is independent of ϕ. The half-power *antenna beamwidth* in the θ direction is the angle θ_B between two directions where the radiated power is half that radiated at the peak, as illustrated. Thus (10.3.3) and the figure also suggest that high directivity antennas have narrower beamwidths θ_B, or are more "directive".

The ratio P_T/P_A is that fraction of the power available at the antenna terminals (P_A) that is radiated; it is defined as the *radiation efficiency* η_R:

$$\eta_R \equiv P_T/P_A \qquad \text{(radiation efficiency)} \qquad (10.3.4)$$

$$G(\theta,\phi) \equiv \eta_R D(\theta,\phi) \qquad (10.3.5)$$

The radiation efficiency is usually near unity because the resistive losses and the reflective losses due to impedance mismatches are small in most systems. Typical exceptions to the rule $\eta_R \cong 1$ include most short dipoles and antennas that are used over bandwidths much greater than an octave; their impedances are difficult to match.

The directivity of a short dipole antenna is given by substituting (10.2.27) and (10.2.28) into (10.3.2):

$$D(\theta,\phi) = \frac{(\eta_o/2)|\underline{I}_o d/\lambda 2r|^2 \sin^2\theta}{(\eta_o\pi/3)|\underline{I}_o d/\lambda|^2/4\pi r^2} = 1.5\sin^2\theta \quad \text{(short dipole directivity)} \quad (10.3.6)$$

Lossless matched short dipole antennas have gain:

$$G(\theta,\phi) = 1.5\sin^2\theta \qquad\qquad \text{(short-dipole antenna gain)} \quad (10.3.7)$$

Example 10.3A

What is the maximum solid angle Ω_B [steradians] over which a lossless matched antenna can have constant gain $G_o = 40$ dB? If the beam is circular, approximately what is its diameter θ_B? How much transmitter power P_T is required to yield $\underline{E}_o = 1$ volt per meter at 10 kilometers?

Solution: Since $G(\theta,\phi) = D(\theta,\phi)$ for a lossless matched antenna, and $\int_{4\pi} D(\theta,\phi)d\Omega = 4\pi$, it

follows that $G_o\Omega_B = 4\pi$ since the maximum gain results when all sidelobes have $G = 0$. Therefore $\Omega_B = 4\pi\times10^{-4}$, corresponding to $\pi\theta_B^2/4 \cong \Omega_B \Rightarrow \theta_B \cong 2(\Omega_B/\pi)^{0.5} \cong 2(4\pi\times10^{-4}/\pi)^{0.5} \cong 0.04$ radians $\cong 2.4°$. $G_oP_T/4\pi r^2 = |\underline{E}_o|^2/2\eta_o \Rightarrow P_T = 4\pi r^2|\underline{E}_o|^2/2\eta_o G_o = 4\pi(10^4)^2\times1^2/(2\times377\times10^4) \cong 166$ [W].

10.3.2   Circuit properties of antennas

Antennas connect to electrical circuits, and therefore it is important to understand the circuit properties of antennas. The linearity of Maxwell's equations applies to antennas, so they can therefore be modeled by a *Thevenin equivalent circuit* consisting of a *Thevenin equivalent impedance* $\underline{Z}_A$ in series with a *Thevenin voltage source* $\underline{V}_{Th}$. This section evaluates the Thevenin equivalent impedance $\underline{Z}_A$, and Section 10.3.3 evaluates $\underline{V}_{Th}$. The frequency dependence of these circuit equivalents usually does not map neatly into that of inductors, capacitors, and resistors, and so we simply use complex notation and a generalized $\underline{Z}_A(\omega)$ instead, where:

$$\underline{Z}_A(\omega) = R(\omega) + jX(\omega) \qquad\qquad (10.3.8)$$

$R(\omega)$ is the resistive part of the impedance corresponding to the total power dissipated and radiated, and $X(\omega)$ is the reactive part, corresponding to near-field energy storage.

To find $\underline{Z}_A(\omega)$ we can use the integral form of Poynting's theorem (2.7.23) for a volume V bounded by surface area A to relate the terminal voltage $\underline{V}$ and current $\underline{I}$ to the near and far fields of any antenna:

$$\oiint_A (\overline{E}\times\overline{H}^*)\bullet\hat{n}\,da = -\iiint_V \{\overline{E}\bullet\overline{J}^* + j\omega(\overline{H}^*\bullet\overline{B} - \overline{E}\bullet\overline{D}^*)\}dv \qquad (10.3.9)$$

For example, the short dipole antenna in Figure 10.2.3 is shown surrounded by a surface area A = A' + A" + A"', where A' is the cross-sectional area of the TEM feed line, A" is the outer surface of the coaxial feed line, and A"' is far from the antenna and intercepts only radiated fields.

These three contributions (A', A", and A"') to the surface integral on the left-hand side of (10.3.9) are given by the next three equations:

$$\frac{1}{2}\iint_{A'}\left(\overline{E}\times\overline{H}^*\right)\bullet\hat{n}d\ a = -\frac{1}{2}\underline{V}\underline{I}^* = -\frac{1}{2}\underline{Z}|\underline{I}_o|^2\ \ [W] \tag{10.3.10}$$

Equation (10.3.10) simply expresses in two different ways the power flowing away from the antenna through the TEM feed line; the negative sign results because Poynting's vector here is oriented outward and the current flow $\underline{I}$ is oriented inward. Because no power flows perpendicular to the conducting sheath of the feed line, we have:

$$\iint_{A''}\left(\overline{E}\times\overline{H}^*\right)\bullet\hat{n}d\ a = 0 \tag{10.3.11}$$

The third integral over the far fields A"' captures the total power radiated by the antenna, which must equal the real power into the antenna associated with radiation, or $R_r|\underline{I}_o|^2/2$, where (10.3.12) defines the *radiation resistance* R_r of an antenna. In the far field the left-hand side is purely real:

$$\frac{1}{2}\iint_{A'''}\left(\overline{E}\times\overline{H}^*\right)\bullet\hat{n}d\ a = P_T \equiv \frac{1}{2}|\underline{I}_o|^2\ R_r\ \ [W] \qquad \text{(radiation resistance)} \tag{10.3.12}$$

By combining the expression for $\underline{Z}(\omega)$ in (10.3.10) with equations (10.3.9–12) we obtain:

$$\underline{Z}(\omega) = R + jX = R_r + \iiint_V \left\{\left[\overline{E}\bullet\overline{J}^* + j\omega\left(\overline{H}^*\bullet\overline{B} - \overline{E}\bullet\overline{D}^*\right)\right]\Big/|\underline{I}_o|^2\right\}dv \tag{10.3.13}$$

$$R(\omega) = R_r + \iiint_V jR_e\left\{\left[\overline{E}\bullet\overline{J}^* + \omega\left(\overline{H}^*\bullet\overline{B} - \overline{E}\bullet\overline{D}^*\right)\right]\Big/|\underline{I}_o|^2\right\}dv = R_r + R_d \tag{10.3.14}$$

$$X(\omega) = \iiint_V I_m\left\{\left[\overline{E}\bullet\overline{J}^* + j\omega\left(\overline{H}^*\bullet\overline{B} - \overline{E}\bullet\overline{D}^*\right)\right]\Big/|\underline{I}_o|^2\right\}dv \tag{10.3.15}$$

$X(\omega)$ is the antenna reactance, and the integral in (10.3.14) is the dissipative component $R_d(\omega)$ of antenna resistance $R(\omega)$. If the average near-field magnetic energy storage exceeds the electric energy storage, then the *antenna reactance* X is positive and inductive; if the energy stored is predominantly electric, then X is negative and capacitive. In practice the real part of the $j\omega$ term in (10.3.14) is usually zero, as is the imaginary part of the $\overline{E}\bullet\overline{J}^*$ term in (10.3.15), but there can be exceptions. The R and X of antennas are seldom computed analytically, but are usually determined by experiment or computational tools.

The radiation resistance R_r of short dipole antennas can be estimated using (10.3.12) and (10.2.28); the dissipative resistance R_d in short wires given by (10.3.14) is usually negligible:

$$R_r = \frac{2P_T}{|\underline{I}_o|^2} = \frac{2\eta_o\pi}{3}\left(\frac{d_{eff}}{\lambda}\right)^2 \text{ ohms} \qquad \text{(radiation resistance, short dipole)} \quad (10.3.16)$$

The effective length d_{eff} of a short dipole is approximately half its physical length [see (10.2.25) and Figure 10.2.3].

The reactance X of a short dipole antenna can be found using (10.3.15); it results primarily from the energy stored in the near fields. The near-field energy for short or Hertzian dipoles is predominantly electric, since the near-field $\overline{E} \propto r^{-3}$ (10.2.15) while the near-field $\overline{H} \propto r^{-2}$ (10.2.16), and r→ 0. Since the electric term of (10.3.15) is much greater than the magnetic term, X is negative.

Example 10.3B

A certain matched antenna radiates one watt (P_r) when driven with voltage $\underline{V}_o$ = 10 volts. What is the antenna radiation resistance R_r?

Solution: $P_r = |\underline{V}_o|^2/2R_r \Rightarrow R_r = |\underline{V}_o|^2/2P_r = 10^2/(2\times1) = 50\Omega$.

10.3.3 Receiving properties of antennas

Because Maxwell's equations are linear in field strength, antennas have equivalent circuits consisting of a Thevenin equivalent impedance $\underline{Z}_A(\omega)$, given by (10.3.13), in series with a Thevenin voltage source $\underline{V}_{Th}(\omega)$ that we can now evaluate. Non-zero voltages appear when antennas receive signals, where these voltages depend upon the direction, polarization, and strength of the intercepted waves.

Figure 10.3.1(a) illustrates the Thevenin equivalent circuit for any antenna, and Figure 10.3.1(b) illustrates the electric fields and equipotentials associated with a short dipole antenna intercepting a uniform plane wave polarized parallel to the dipole axis. When the wavelength λ greatly exceeds d and other local dimensions of interest, i.e. $\lambda \to \infty$, then Maxwell's equations become:

$$\nabla \times \overline{E} = -j(2\pi c/\lambda)\overline{B} \to 0 \qquad \text{for } \lambda \to \infty \tag{10.3.17}$$

$$\nabla \times \overline{H} = \overline{J} + j(2\pi c/\lambda)\overline{D} \to \overline{J} \quad \text{for } \lambda \to \infty \tag{10.3.18}$$

But these limits are the equations of electrostatics and magnetostatics. Therefore we can quickly sketch the electric field lines near the short dipole of Figure 10.3.1 using a three-dimensional version of the quasistatic field mapping technique of Section 4.6.2.
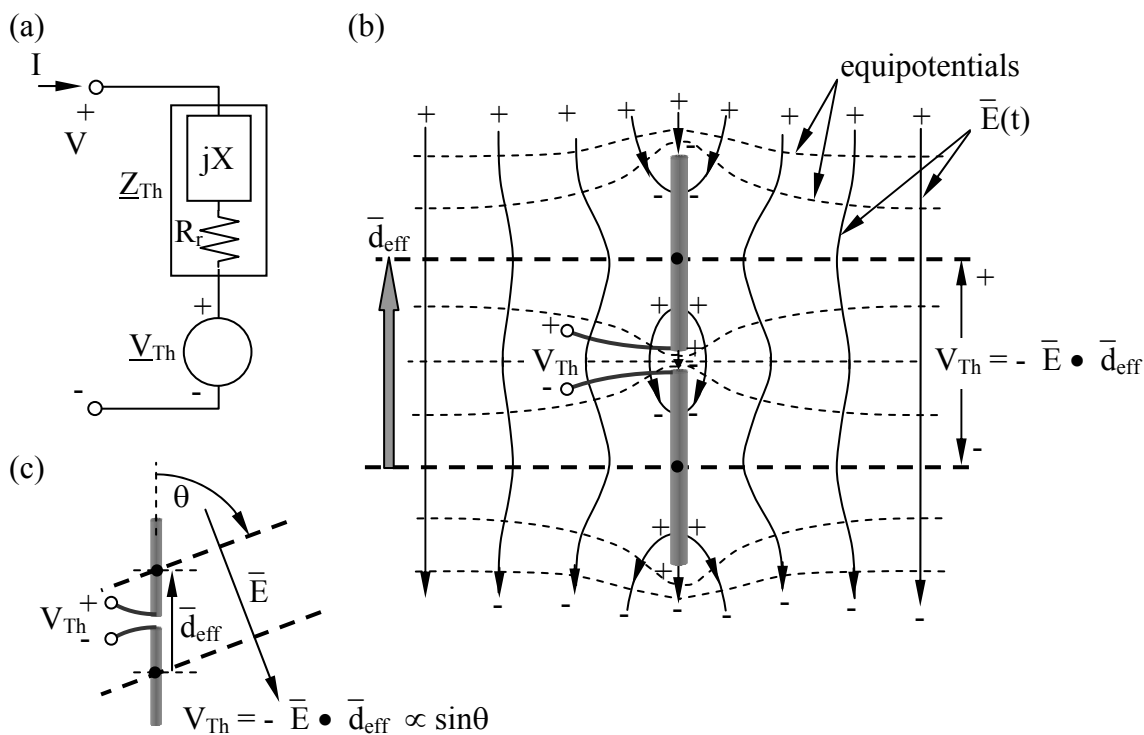
Figure 10.3.1 Thevenin voltage induced on a short dipole antenna.

Far from the dipole the field lines $\overline{E}$ in Figure 10.3.1(b) are those of the quasistatic incident plane wave, i.e., uniform and parallel to the dipole. Close to the conducting dipole $\overline{E}$ is distorted to match the boundary conditions: 1) $\overline{E}_{||} = 0$, and 2) each half of the dipole is an equipotential, intercepting only one equipotential line (boldface, dashed). If the wires comprising the short dipole are very thin, the effects of each wire on the other are negligible. Under these assumptions symmetry dictates the form for three of the equipotentials in Figure 10.3.1—the equipotentials through the center of the dipole and through each of its two halves are straight lines. The other equipotentials sketched with dashed lines curve around the conductors. The field lines $\overline{E}$ are sketched with solid lines locally perpendicular to the equipotentials. The field lines terminate at charges on the surface of the conductors and possibly at infinity, as governed by Gauss's law: $\hat{n} \bullet \overline{D} = \sigma_S$.

Figures 10.3.1(b) and (c) suggest why the open-circuit voltage V_{Th} of the short dipole antenna equals the potential difference between the centers of the two halves of this ideal dipole:

$$V_{Th} \equiv -\overline{E} \bullet \overline{d}_{eff} \qquad \text{(voltage induced on dipole antenna)} \quad (10.3.19)$$

- 319 -

The effective length of the dipole, $\overline{d}_{eff}$, is defined by (10.3.19), and is the same as the effective length defined in terms of the current distribution (10.2.25) for infinitesimally thin straight wires of length d $\ll \lambda$. Generally $d_{eff} \cong d/2$, which is the distance between the centers of the two conductors. Each conductor is essentially sampling the electrostatic potential in its vicinity and conveying that to the antenna terminals. The orientation of $\overline{d}_{eff}$ is that of the dipole current flow that would be driven by external sources having the defined terminal polarity.

The maximum power an antenna can deliver to an external circuit of impedance $\underline{Z}_L$ is easily computed once the antenna equivalent circuit is known. To maximize this transfer it is first necessary to add an external load reactance, $-jX_L$, in series to cancel the antenna reactance $+jX$ (X is negative for a short dipole antenna because it is capacitive). Then the resistive part of the load R_L must match that of the antenna, i.e., $R_L = R_r$. Maximum power transfer occurs when the impedances match so incident waves are not reflected. In this conjugate-match case ($Z_L = Z_A^*$), the antenna Thevenin voltage $\underline{V}_{Th}$ is divided across the two resistors R_r and R_L so that the voltage across R_L is $\underline{V}_{Th}/2$ and the power received by the short dipole antenna is:

$$P_r = \frac{1}{2R_r}\left|\frac{\underline{V}_{Th}}{2}\right|^2 \ [W] \qquad\qquad \text{(received power)} \quad (10.3.20)$$

Substitution into (10.3.20) of R_r (10.3.16) and V_{Th} (10.3.19) yields the received power:

$$P_r = \frac{3}{4\eta_o\pi(d/\lambda)^2}\left|\frac{\overline{E}d_{eff}\sin\theta}{2}\right|^2 = \frac{|\overline{E}|^2}{2\eta_o}\frac{\lambda^2}{4\pi}(1.5\sin^2\theta) \qquad\qquad (10.3.21)$$

$$P_r = I(\theta,\varphi)\frac{\lambda^2}{4\pi}G(\theta,\varphi) = I(\theta,\varphi)A(\theta,\varphi) \ [W] \qquad\qquad \text{(power received)} \quad (10.3.22)$$

where $I(\theta,\phi)$ is the power intensity [Wm^{-2}] of the plane wave arriving from direction (θ,ϕ), $G(\theta,\phi) = D(\theta,\phi) = 1.5\sin^2\theta$ is the antenna gain of a lossless short-dipole antenna (10.3.7), and $A(\theta,\phi)$ is the *antenna effective area* as defined by the equation $P_r \equiv I(\theta,\phi) A(\theta,\phi)$ [W] for the power received. Section 10.3.4 proves that the simple relation between gain $G(\theta,\phi)$ and effective area $A(\theta,\phi)$ proven in (10.3.22) for a short dipole applies to essentially all[53] antennas:

$$A(\theta,\varphi) = \frac{\lambda^2}{4\pi}G(\theta,\varphi) \ \left[m^2\right] \qquad\qquad \text{(antenna effective area)} \quad (10.3.23)$$

Equation (10.3.23) says that the effective area of a matched short-dipole antenna is equivalent to a square roughly $\lambda/3$ on a side, independent of antenna length. A small wire structure ($\ll \lambda/3$) can capture energy from this much larger area if it has a conjugate match, which generally

---

[53] This expression requires that all media near the antenna be reciprocal, which means that no magnetized plasmas or ferrites should be present so that the permittivity and permeabiliy matrices ε and μ everywhere equal their own transposes.

requires a high-Q resonance, large field strengths, and high losses. In practice, short-dipole antennas generally have a reactive mismatch that reduces their effective area below optimum.

10.3.4 Generalized relation between antenna gain and effective area

Section 10.3.3 proved for a short-dipole antenna the basic relation (10.3.23) between antenna gain $G(\theta,\phi)$ and antenna effective area $A(\theta,\phi)$:

$$A(\theta,\phi) = \frac{\lambda^2}{4\pi}G(\theta,\phi) \qquad\qquad (10.3.24)$$

This relation can be proven for any arbitrary antenna provided all media in and near the antenna are *reciprocal media,* i.e., their complex permittivity, permeability, and conductivity matrices $\underline{\varepsilon}$, $\underline{\mu}$, and $\underline{\sigma}$ are all symmetric:

$$\underline{\varepsilon} = \underline{\varepsilon}^t, \quad \underline{\mu} = \underline{\mu}^t, \quad \underline{\sigma} = \underline{\sigma}^t \qquad\qquad (10.3.25)$$

where we define the transpose operator t such that $\underline{A}^t_{ij} = \underline{A}_{ji}$. Non-reciprocal media are rare, but include magnetized plasmas and magnetized ferrites; they are not discussed in this text. Media characterized by matrices are discussed in Section 9.5.1.

To prove (10.3.24) we characterize a general linear 2-port network by its impedance matrix:

$$\overline{\overline{Z}} = \begin{bmatrix} \underline{Z}_{11} & \underline{Z}_{12} \\ \underline{Z}_{21} & \underline{Z}_{22} \end{bmatrix} \qquad\qquad \text{(impedance matrix)} \quad (10.3.26)$$

$$\overline{V} = \overline{\overline{Z}}\,\overline{I} \qquad\qquad (10.3.27)$$

where $\overline{V}$ and $\overline{I}$ are the two-element voltage and current vectors $[\underline{V}_1, \underline{V}_2]$ and $[\underline{I}_1, \underline{I}_2]$, and $\underline{V}_i$ and $\underline{I}_i$ are the voltage and current at terminal pair i. This matrix $\overline{\overline{Z}}$ does not depend on the network to which the 2-port is connected. If the 2-port system is a *reciprocal network,* then $\overline{\overline{Z}} = \overline{\overline{Z}}^t$, so $\underline{Z}_{12} = \underline{Z}_{21}$.

Since Maxwell's equations are linear, $\underline{V}$ is linearly related to $\underline{I}$, and we can define an antenna impedance $\underline{Z}_{11}$ consisting of a real part (10.3.14), typically dominated by the radiation resistance R_r (10.3.12), and a reactive part jX (10.3.15). Thus $\underline{Z}_{11} = R_1 + jX_1$, where R_1 equals the sum of the dissipative resistance R_{d1} and the radiation resistance R_{r1}. For most antennas $R_d \ll R_r$.

Figure 10.3.2 illustrates an unknown reciprocal antenna (1) that communicates with a short-dipole test antenna (2) that is aimed at antenna (1). Because the relations between the voltages and currents at the terminals are determined by electromagnetic waves governed by the linear Maxwell equations, the two antennas constitute a two-port network governed by (10.3.26) and

(10.3.27) and the complex impedance matrix $\bar{\bar{Z}}$. Complex notation is appropriate here because antennas are frequency dependent. This impedance representation easily introduces the reciprocity constraint to the relation between $G(\theta,\phi)$ and $A(\theta,\phi)$. We assume each antenna is matched to its load $\underline{Z}_L = R_r - jX$ so as to maximize power transfer.
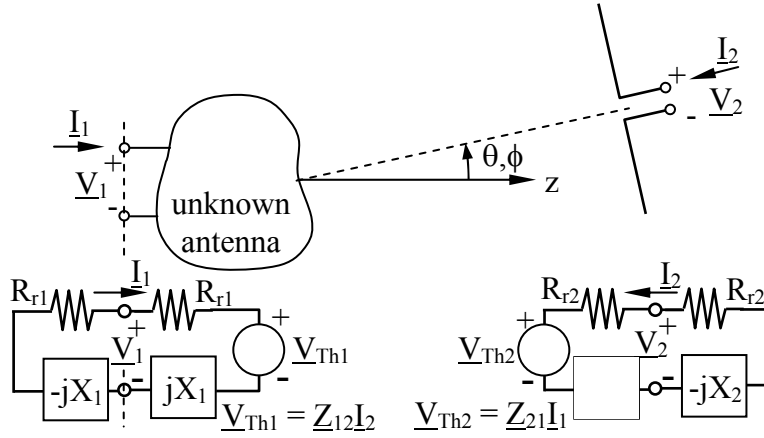
Figure 10.3.2   Coupled reciprocal antennas for relating $G(\theta,\phi)$ to $A(\theta,\phi)$.

The power P_r received by each antenna and dissipated in the load can be expressed in two equivalent ways—in terms of antenna *mutual impedance* $\underline{Z}_{ij}$ and in terms of antenna gain and effective area:

$$P_{r1} = \frac{\left|\underline{V}_{Th1}\right|^2}{8R_{r1}} = \frac{\left|\underline{Z}_{12}\underline{I}_2\right|^2}{8R_{r1}} = \frac{G_2 P_{t2}}{4\pi r^2} A_1 \qquad (10.3.28)$$

$$P_{r2} = \frac{\left|\underline{V}_{Th2}\right|^2}{8R_{r2}} = \frac{\left|\underline{Z}_{21}\underline{I}_1\right|^2}{8R_{r2}} = \frac{G_1 P_{t1}}{4\pi r^2} A_2 \qquad (10.3.29)$$

Taking the ratio of these two equations in terms of G and A yields:

$$\frac{P_{r2}}{P_{r1}} = \frac{G_1 A_2 P_{t1}}{G_2 A_1 P_{t2}} \qquad (10.3.30)$$

$$\therefore \frac{A_1}{G_1} = \frac{A_2}{G_2} \frac{P_{t1} P_{r1}}{P_{t2} P_{r2}} \qquad (10.3.31)$$

But the ratio of the same equations in terms of $\underline{Z}_{ij}$ also yields:

$$\frac{P_{r1}}{P_{r2}} = \frac{\left|\underline{Z}_{12}\underline{I}_2\right|^2 R_{r2}}{\left|\underline{Z}_{21}\underline{I}_1\right|^2 R_{r1}} = \frac{\left|\underline{Z}_{12}\right|^2 P_{t2}}{\left|\underline{Z}_{21}\right|^2 P_{t1}} \qquad (10.3.32)$$

Therefore if reciprocity applies, so that $|\underline{Z}_{12}|^2 = |\underline{Z}_{21}|^2$, then (10.3.23) for a short dipole and substitution of (10.3.32) into (10.3.31) proves that all reciprocal antennas obey the same A/G relationship:

$$\frac{A_1(\theta,\phi)}{G_1(\theta,\phi)} = \frac{A_2}{G_2} = \frac{\lambda^2}{4\pi} \qquad \text{(generalized gain-area relationship)} \quad (10.3.33)$$

10.3.5 Communication links

We now can combine the transmitting and receiving properties of antennas to yield the power that can be transmitted from one place to another. For example, the intensity $I(\theta,\phi)$ at distance r that results from transmitting P_t watts from an antenna with gain $G_t(\theta,\phi)$ is:

$$I(\theta,\phi) = G(\theta,\phi)\frac{P_t}{4\pi r^2} \quad \left[W/m^2\right] \qquad \text{(radiated intensity)} \quad (10.3.34)$$

The power received by an antenna with effective area $A(\theta,\phi)$ in the direction θ,ϕ from which the signal arrives is:

$$P_r = I(\theta,\phi)A(\theta,\phi) \quad [W] \qquad \text{(received power)} \quad (10.3.35)$$

where use of the same angles θ,ϕ for the transmission and reception implies here that the same ray is being both transmitted and received, even though the transmitter and receiver coordinate systems are typically distinct. Equation (10.3.33) says:

$$A(\theta,\phi) = \frac{\lambda^2}{4\pi}G_r(\theta,\phi) \qquad\qquad (10.3.36)$$

where G_r is the gain of the receiving antenna, so the power received (10.3.35) becomes:

$$P_r = \frac{P_t}{4\pi r^2}G_t(\theta,\phi)\frac{\lambda^2}{4\pi}G_r(\theta,\phi) = P_tG_t(\theta,\phi)G_r(\theta,\phi)\left(\frac{\lambda}{4\pi r}\right)^2 \quad [W] \qquad (10.3.37)$$

Although (10.3.37) suggests the received power becomes infinite as $r \rightarrow 0$, this would violate the far-field assumption that $r \gg \lambda/2\pi$.

Example 10.3C
Two wireless phones with matched short dipole antennas having d_{eff} equal one meter communicate with each other over a ten kilometer unobstructed path. What is the maximum power P_A available to the receiver if one watt is transmitted at $f = 1$ MHz? At 10 MHz? What is P_A at 1 MHz if the two dipoles are 45° to each other?

Solution: $P_A = AI$, where A is the effective area of the receiving dipole and I is the incident wave intensity [W m^{-2}]. $P_A = A\left(P_t G_t / 4\pi r^2\right)$ where $A = G_r \lambda^2 / 4\pi$ and $G_t \leq 1.5$; $G_r \leq 1.5$.   Thus $P_A = \left(G_r \lambda^2 / 4\pi\right)\left(P_t G_t / 4\pi r^2\right) = P_t \left(1.5\lambda / 4\pi r\right)^2 = P_t \left(1.5c / 4\pi rf\right)^2$ $= 1\left(1.5 \times 3 \times 10^8 / 4\pi 10^4 \times 10^6\right)^2 \cong 1.3 \times 10^{-5} [\text{W}]$. At 10 MHz the available power out is $\sim 1.3 \times 10^{-7}$ [W]. If the dipoles are 45° to each other, the receiving cross section is reduced by a factor of $\sin^2 45° = 0.5 \Rightarrow P_A \cong 6.4 \times 10^{-6} [\text{W}]$.

Example 10.3D

In terms of the incident electric field $\underline{E}_o$, what is the maximum Thevenin equivalent voltage source $\underline{V}_{Th}$ for a small N-turn loop antenna operating at frequency f? A loop antenna is made by winding N turns of a wire in a flat circle of diameter D, where D << λ. If N = 1, what must D be in order for this loop antenna to have the same maximum $\underline{V}_{Th}$ as a short dipole antenna with effective length d_{eff}?

Solution: The open-circuit voltage $\underline{V}_{Th}$ induced at the terminals of a small wire loop (D << λ) follows from Ampere's law: $\underline{V}_{Th} = \int_C \overline{E} \bullet d\overline{s} = -N \iint j\omega\mu_o \overline{H} \bullet d\overline{a}$ $= -Nj\omega\mu_o \underline{H}\pi D^2 / 4 = -Nj\omega\mu_o \underline{E}\pi D^2 / 4\eta_o$.   But $\omega\mu_o \pi / 4\eta_o = f\pi^2 / 2c$, so $\left|\underline{V}_{Th}\right| = Nf\pi^2 \left|\underline{E}_o\right| D^2 / 2c$.   For a short dipole antenna the maximum $\left|\underline{V}_{Th}\right| = d_{eff} \left|\underline{E}_o\right|$, so $D = \left(2cd_{eff} / f\pi^2 N\right)^{0.5} = \left(2\lambda d_{eff} / \pi^2 N\right)^{0.5} \cong 0.45\left(d_{eff}\lambda / N\right)^{0.5}$.

10.4   Antenna arrays

10.4.1   Two-dipole arrays

Although some communications services such as mobile phones use nearly omnidirectional electric or magnetic dipole antennas (short-dipole and loop antennas), most fixed services such as point-to-point, broadcast, and satellite services benefit from larger antenna gains. Also, some applications require rapid steering of the antenna beam from one point to another, or even the ability to observe or transmit in multiple narrow directions simultaneously. *Antenna arrays* with two or more dipoles can support all of these needs. Arrays of other types of antennas can similarly boost performance.

Since the effective area of an antenna, A(θ,ϕ), is simply related by (10.3.36) to antenna gain G(θ,ϕ), the gain of a dipole array fully characterizes its behavior, which is determined by the array current distribution. Sometimes some of the dipoles are simply mirrored images of others.

In every case the total radiated field is simply the superposition of the fields radiated by each contributing dipole in proportion to its strength, and delayed in proportion to its distance from the observer. For two-dipole arrays, the differential path length to the receiver can lead to reinforcement if the two sinusoidal waves are in phase, cancellation if they are 180° out of phase and equal, and intermediate strength otherwise.

It is convenient to represent the signals as phasors since the patterns are frequency dependent, so the total observed electric field $\overline{E} = \sum_i \overline{E}_i$, where $\overline{E}_i$ is the observed contribution from short-dipole i, including its associated phase lag of γ_i radians due to distance traveled. Consider first the two-dipole array in Figure 10.4.1(a), where the dipoles are z-axis oriented, parallel, fed in phase, and spaced distance L apart laterally in the y direction. Any observer in the x-z plane separating the dipoles receives equal in-phase contributions from each dipole, thereby doubling the observed far-field $\overline{E}_{ff}$ and quadrupling the power intensity P [Wm^{-2}] radiated in that direction θ relative to what would be transmitted by a single dipole.
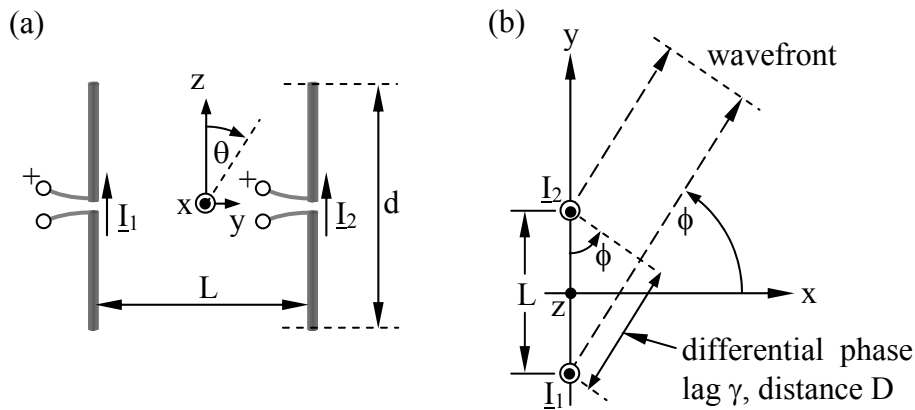
Figure 10.4.1   Two-dipole array.

The radiated power $P(r, \phi)$ in Figure 10.4.1 depends on the differential phase lag γ between the contributions from the two antennas. When the two dipoles are excited equally $\left(\underline{I}_1 = \underline{I}_2 = \underline{I} \right)$ and are spaced $L = \lambda/2$ apart, the two rays add in phase everywhere in the x-z plane perpendicular to the array axis, but are $\lambda/2$ (180°) out of phase and cancel along the array (y) axis. The resulting $G(\phi)$ is sketched in Figure 10.4.2(a) for the x-y plane. If $L = \lambda$ as illustrated in Figure 10.4.2(b), then the two rays add in phase along both the x-z plane and the y axis, but cancel in the x-y plane at $\phi_{null} = 30^\circ$ where the differential delay between the two rays is $\lambda/2$, as suggested by the right triangle in the figure.

Figure 10.4.2(c) illustrates how a non-symmetric pattern can be synthesized by exciting the two dipoles out of phase. In this case the lower dipole leads the upper dipole by 90 degrees, so that the total phase difference between the two rays propagating in the negative y direction is 180 degrees, producing cancellation; this phase difference is zero degrees for radiation in the +y direction, so the two rays add. Along the ±x axis the two rays are 90 degrees out of phase so the total E is $\sqrt{2}$ greater than from a single dipole, and the intensity is doubled. When the two phasors are in phase the total E is doubled and the radiated intensity is 4 times that of a single dipole; thus the intensity radiated along the ±x axis is half that radiated along the +y axis. Figure 10.4.2(d) illustrates how a null-free pattern can be synthesized with non-equal excitation of the two dipoles. In this case the two dipoles are driven in phase so that the radiated phase difference is 180 degrees along the ±y axis due to the $\lambda/2$ separation of the dipoles. Nulls are avoided by

exciting either dipole with a current that is ~40 percent of the other so that the ratio of maximum gain to minimum gain is ~$[(1 + 0.4)/(1 - 0.4)]^2 = 5.44$, and the pattern is vaguely rectangular.
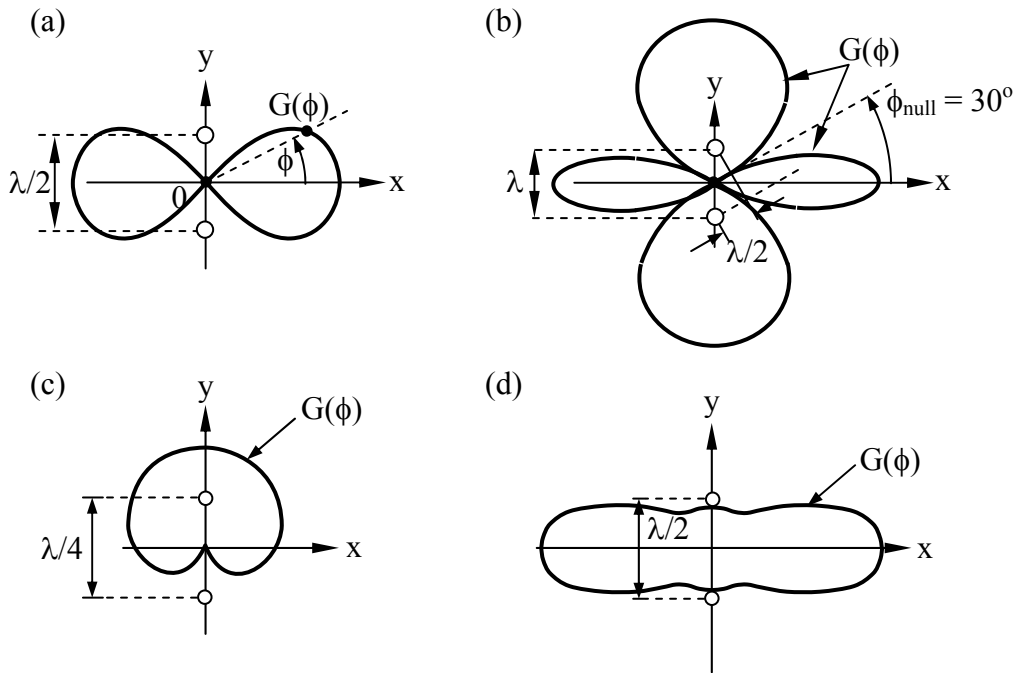
Figure 10.4.2   Gain $G(\phi)$ in the x-y plane orthogonal to two z-oriented dipoles.

A mathematical expression for the gain pattern can also be derived. Superimposing (10.2.8) for $\underline{I}_1$ and $\underline{I}_2$ yields:

$$\overline{\underline{E}}_{ff} \cong \hat{\theta} j \left(k \eta_o d_{eff}/4\pi r \right) \sin \theta \left(\underline{I}_1 e^{-jkr_1} + \underline{I}_2 e^{-jkr_2} \right)$$

$$\cong \hat{\theta} j \left(\eta_o d_{eff}/2\lambda r \right) \sin \theta \, \underline{I} \, e^{-jkr} \left(e^{+0.5jkL \sin \phi} + e^{-0.5jkL \sin \phi} \right) \tag{10.4.1}$$

$$\cong \hat{\theta} j \left(\eta_o \underline{I} d_{eff}/\lambda r \right) \sin \theta \, e^{-jkr} \cos \left(\pi L \lambda^{-1} \sin \phi \right) \tag{10.4.2}$$

where we have used the identities $e^{j\alpha} + e^{-j\alpha} = 2 \cos\alpha$ and $k = 2\pi/\lambda$.

Example 10.4A

If the two dipoles of Figure 10.4.1 are fed in phase and their separation is $L = 2\lambda$, at what angles ϕ in the x-y plane are there nulls and peaks in the gain $G(\phi)$? Are these peaks equal? Repeat this analysis for $L = \lambda/4$, assuming the voltage driving the dipole at $y > 0$ has a 90° phase lag relative to the other dipole.

Solution: Referring to Figure 10.4.1, there are nulls when the phase difference γ between the two rays arriving at the receiver is π or 3π, or equivalently, $D = \lambda/2$ or $3\lambda/2$, respectively.   This   happens   at   the   angles ϕ = $\pm$ $\sin^{-1}(D/L)$ =

- 326 -

$\pm\sin^{-1}[(\lambda/2)/2\lambda] = \pm\sin^{-1}(0.25) \cong \pm14°$, and $\phi = \pm\sin^{-1}(0.75) \cong \pm49°$. There are also nulls, by symmetry, at angles 180° away, or at $\phi \cong \pm194°$ and $\pm229°$. There are gain peaks when the two rays are in phase ($\phi = 0°$ and 180°) and when they differ in phase by 2π or 4π, which happens when $\phi = \pm\sin^{-1}(\lambda/2\lambda) = \pm30°$ and $\phi = \pm210°$, or when $\phi = \pm90°$, respectively. The gain peaks are equal because they all correspond to the two rays adding coherently with the same magnitudes. When $L = \lambda/4$ the two rays add in phase at $\phi = 90°$ along the +y axis because in that direction the phase lag balances the 90° delay suffered by the ray from the dipole on the -y axis. At $\phi = 270°$ these two 90-degree delays add rather than cancel, so the two rays cancel in that direction, producing a perfect null.

10.4.2   Array antennas with mirrors

One of the simplest ways to boost the gain of a short dipole antenna is to place a mirror behind it to as to reinforce the radiation in the desired forward direction and cancel it behind. Figure 10.4.3 illustrates how a short current element I placed near a perfectly conducting planar surface will behave as if the mirror were replaced by an image current an equal distance behind the mirror and pointed in the opposite direction parallel to the mirror but in the same direction normal to the mirror. The fields in front of the mirror are identical with and without the mirror if it is sufficiently large. Behind the mirror the fields approach zero, of course. Image currents and charges were discussed in Section 4.2.

Figure 10.4.3(a) illustrates a common way to boost the forward gain of a dipole antenna by placing it $\lambda/4$ in front of a planar mirror and parallel to it.
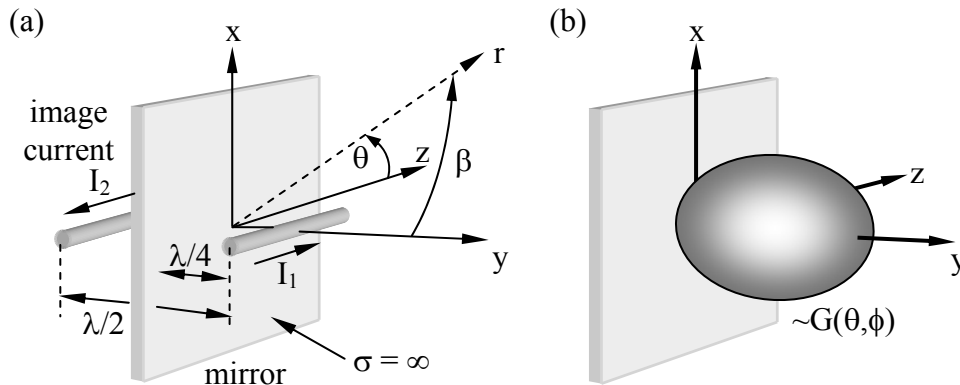
Figure 10.4.3   Half-wave dipole antenna $\lambda/4$ in front of a mirror.

The image current is 180 degrees out of phase, so the $\lambda/2$ delay suffered by the image ray brings it into phase coherence with the direct ray, effectively doubling the far field E_{ff} and quadrupling the intensity and gain G_o relative to the absence of the mirror. In all directions more nearly parallel to the mirror the source and image are more nearly out of phase, so the gain in those directions is diminished relative to the absence of the mirror. The resulting antenna gain $G(\theta,\phi)$ is sketched in Figure 10.4.3(b), and has no backlobes.

For the case where the dipole current $\underline{I}_2 = -\underline{I}_1$ and $kr_1 = kr - (\pi/2)\cos\beta$, the far-field in the forward direction is the sum of the contributions from $\underline{I}_1$ and $\underline{I}_2$, as given by (10.4.1):

$$
\begin{aligned}
\overline{E}_{ff} &= \hat{\theta}\left(j\eta_o d_{eff}/2\lambda r\right)\sin\ \theta\left(\underline{I}_1 e^{-jkr_1} + \underline{I}_2 e^{-jkr_2}\right) \\
&= \hat{\theta}\left(j\eta_o d_{eff}/2\lambda r\right)\sin\ \theta e^{-jkr}\underline{I}_1\left(e^{j(\pi/2)\cos\beta} - e^{-j(\pi/2)\cos\beta}\right)
\end{aligned}
\tag{10.4.3}
$$

$$
= -\hat{\theta}\left(\eta_o d_{eff}/\lambda r\right)\sin\ \theta\ e^{-jkr}\underline{I}_1\sin\left[(\pi/2)\cos\beta\right]
\tag{10.4.4}
$$

This expression reveals that the antenna pattern has no sidelobes and is pinched somewhat more in the θ direction than in the β direction (these directions are not orthogonal). An on-axis observer will receive a z-polarized signal.

Mirrors can also be parabolic and focus energy at infinity, as discussed further in Section 11.1. The sidelobe-free properties of this dipole-plus-mirror make it a good antenna feed for radiating energy toward much larger parabolic reflectors.

Example 10.4B

Automobile antennas often are thin metal rods ~1-meter long positioned perpendicular to an approximately flat metal surface on the car; the rod and flat surface are electrically insulated from each other. The rod is commonly fed by a coaxial cable, the center conductor being attached to the base of the rod and the sheath being attached to the adjacent car body. Approximately what is the radiation resistance and pattern in the 1-MHz radio broadcast band, assuming the flat plate is infinite?

Solution: Figure 4.2.3 shows how the image of a current flowing perpendicular to a conducting plane flows in the same direction as the original current, so any current flowing in the rod has an image current that effectively doubles the length of this antenna. The wavelength at 1 MHz is ~300 meters, much longer than the antenna, so the short-dipole approximation applies and the current distribution on the rod and its image resembles that of Figure 10.2.3; thus $d_{eff} \cong 1$ meter and the pattern above the metal plane is the top half of that illustrated in Figure 10.2.4. The radiation resistance of a normal short dipole antenna (10.3.16) is $R_r = 2P_T/|I_o|^2 = 2\eta_o\pi(d_{eff}/\lambda)^2/3 = 0.0088$ ohms for $d_{eff} = 1$ meter. Here, however, the total power radiated P_T is half that radiated by a short dipole of length 2 meters because there is no power radiated below the conducting plane, so $R_r = 0.0044$ ohms. The finite size of an automobile effectively warps and shortens both the image current and the effective length of the dipole, although the antenna pattern for a straight current is always dipolar above the ground plane.

10.4.3 Element and array factors

The power radiated by dipole arrays depends on the directional characteristics of the individual dipole antennas as well as on their spacing relative to wavelength λ. For example, (10.4.3) can be generalized to N identically oriented but independently positioned and excited dipoles:

$$\overline{E}_{ff} \cong \left[\hat{\theta} \frac{j\eta_o d_{eff}}{2\lambda r} \sin \theta \right] \left[\sum_{i=1}^{N} \overline{I}_i e^{-jkr_i} \right] = \overline{\varepsilon}(\theta,\phi) \underline{F}(\theta,\phi) \tag{10.4.5}$$

The *element factor* $\overline{\varepsilon}(\theta,\phi)$ for the dipole array represents the behavior of a single element, assuming the individual elements are identically oriented. The *array factor,* $\underline{F}(\theta,\phi) = \sum_i^N \overline{I}_i e^{-jkr_i}$, represents the effects of the relative strengths and placement of the elements. The distance between the observer and each element i of the array is r_i, and the phase lag $kr_i = 2\pi r_i/\lambda$.

Consider the element factor in the x-y plane for the two z-oriented dipoles of Figure 10.4.4(a).
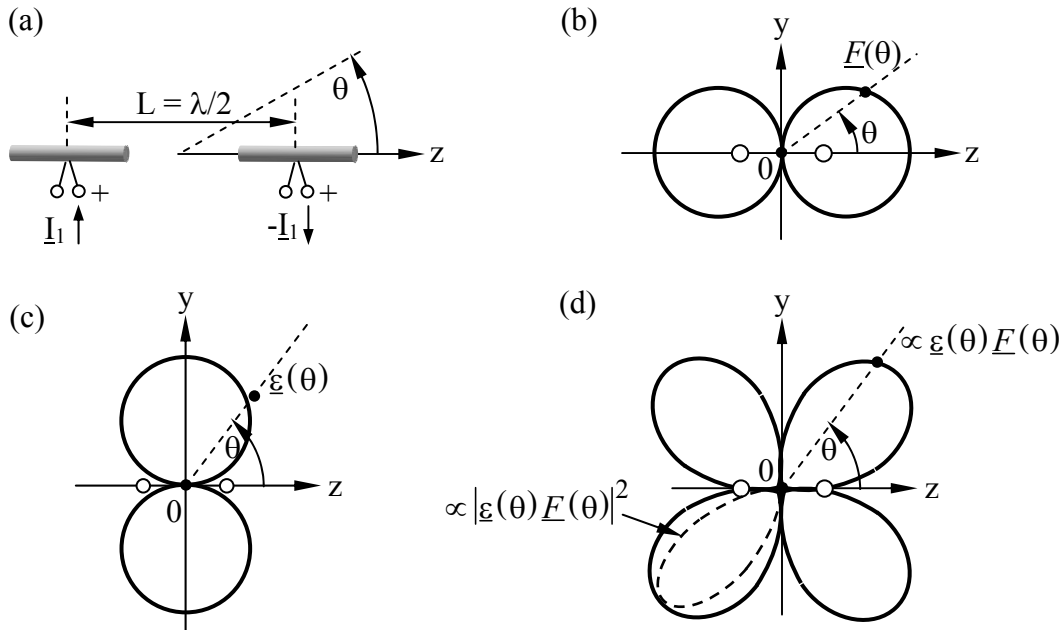
Figure 10.4.4 Normalized array and element factors for dipole arrays.

This element factor $\overline{\varepsilon}(\theta,\phi)$ is constant, independent of ϕ, and has a circular pattern. The total antenna pattern is $\left|\overline{\varepsilon}(\theta,\phi)\underline{F}(\theta,\phi)\right|^2$ where the array factor $\underline{F}(\theta,\phi)$ controls the array antenna pattern in the x-y plane for these two dipoles. The resulting antenna pattern $\left|\underline{F}(\theta,\phi)\right|^2$ in the x-y plane is plotted in Figure 10.4.4(a) and (b) for the special cases $L = \lambda/2$ and $L = \lambda$, respectively.

Both the array and element factors contribute to the pattern for this antenna in the x-z plane, narrowing its beamwidth (not illustrated).

Figure 10.4.4 illustrates a case where both the element and array factors are important; L = λ/2 here and the dipoles are fed 180$^\text{o}$ out of phase. In this case the out-of-phase signals from the two dipoles cancel everywhere in the x-y plane and add in phase along the z axis, corresponding to the array factor plotted in Figure 10.4.4(b) for the y-z plane. Note that when θ = 60$^\text{o}$ the two phasors are 45$^\text{o}$ out of phase and $\left|\underline{F}(\theta,\phi)\right|^2$ has half its peak value. The element factor in the y-z plane appears in Figure 10.4.4(c), and the dashed antenna pattern $\left|\underline{\varepsilon}(\theta)\underline{F}(\theta)\right|^2 \propto G(\theta)$ in Figure 10.4.4(d) shows the effects of both factors (only one of the four lobes is plotted). This antenna pattern is a figure of revolution about the z axis and resembles two wide rounded cones facing in opposite directions.

Example 10.4C

What are the element and array factors for the two-dipole array for the first part of Example 10.4A?

Solution: From (10.4.5) the element factor for such dipoles is $\hat{\theta}j\left(\eta_o d_{eff}/2\lambda r\right)\sin\theta$. The last factor of (10.4.5) is the array factor for such two-dipole arrays:

$$\underline{F}(\theta,\phi) = \underline{I}\left(e^{+0.5(j2\pi 2\lambda/\lambda)\sin\phi} + e^{-0.5(j2\pi 2\lambda/\lambda)\sin\phi}\right)$$

$$= \left(e^{2\pi j \sin\phi} + e^{-2\pi j \sin\phi}\right) = 2\underline{I}\cos(2\pi\sin\phi)$$

10.4.4   Uniform dipole arrays

Uniform dipole arrays consist of N identical dipole antennas equally spaced in a straight line. Their current excitation $\underline{I}_i$ has equal magnitudes for all i, and a phase angle that uniformly increases by ψ radians between adjacent dipoles. The fields radiated by the array can be determined using (10.4.5):

$$\overline{E}_{ff} \cong \left[\hat{\theta}j\left(\eta_o d_{eff}/2\lambda r\right)\sin\theta\right]\left[\sum_{i=1}^{N}\underline{I}_i e^{-jk r_i}\right] = \overline{\underline{\varepsilon}}(\theta,\phi)\underline{F}(\theta,\phi) \qquad (10.4.6)$$

The z axis is defined by the orientation of the dipoles, which are all parallel to it. The simplest arrangement of the dipoles is along that same z axis, as in Figure 10.4.5, although (10.4.6) applies equally well if the dipoles are spaced in any arbitrary direction. Figure 10.4.1(a) illustrates the alternate case where two dipoles are spaced along the y axis, and Figure 10.4.2 shows the effects on the patterns.
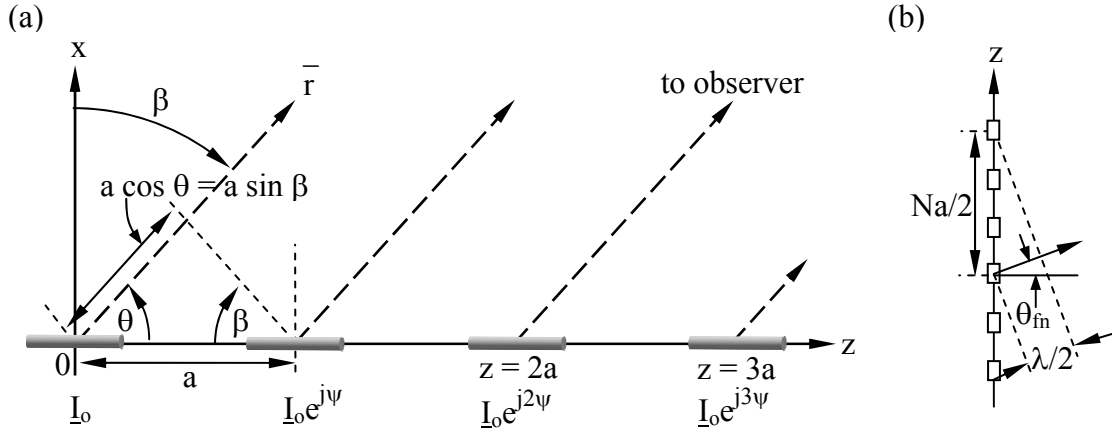
Figure 10.4.5   Uniform dipole array.

Consider the N-element array for Figure 10.4.5(a). The principal difference between these two-dipole cases and N-element uniform arrays lies in the array factor:

$$\underline{F}(\theta,\phi) = \sum_{i=1}^{N} \underline{I}_i e^{-jkr_i} = \underline{I}_o e^{-jkr} \sum_{i=0}^{N-1} e^{ji\psi} e^{jika\cos\theta} = \underline{I}_o e^{-jkr} \sum_{i=0}^{N-1} \left[e^{j(\psi + ka\cos\theta)} \right]^i \qquad (10.4.7)$$

The geometry illustrated in Figure 10.4.5(a) yields a phase difference of $(\psi + ka\cos\theta)$ between the contributions from adjacent dipoles.

Using the two identities:

$$\sum_{i=0}^{N-1} x^i = (1 - x^N)/(1 - x) \qquad (10.4.8)$$

$$1 - e^{jA} = e^{jA/2}\left(e^{-jA/2} - e^{+jA/2}\right) = -2je^{jA/2}\sin(A/2) \qquad (10.4.9)$$

(10.4.7) becomes:

$$\begin{aligned}
\underline{F}(\theta,\phi) &= \underline{I}_o e^{-jkr} \frac{1 - e^{jN(\psi + ka\cos\theta)}}{1 - e^{j(\psi + ka\cos\theta)}} \\
&= \underline{I}_o e^{-jkr} \times \frac{e^{jN(\psi + ka\cos\theta)/2}}{e^{j(\psi + ka\cos\theta)/2}} \frac{\sin\left[N(\psi + ka\cos\theta)/2\right]}{\sin\left[(\psi + ka\cos\theta)/2\right]}
\end{aligned} \qquad (10.4.10)$$

Since the element factor is independent of ϕ, the antenna gain has the form:

- 331 -

$$G(\theta) \propto \left|\underline{F}(\theta,\phi)\right|^2 \propto \frac{\sin^2\left[N(\psi + ka\cos\theta)/2\right]}{\sin^2\left[(\psi + ka\cos\theta)/2\right]} \qquad (10.4.11)$$

If the elements are excited in phase ($\psi = 0$), then the maximum gain is broadside with $\theta = 90^0$, because only in that direction do all N rays add in perfect phase. In this case the first nulls $\theta_{first\ null}$ bounding the main beam occur when the numerator of (10.4.11) is zero, which happens when:

$$\frac{N}{2} ka \cos\theta_{first\ null} = \pm\pi \qquad (10.4.12)$$

Note that the factor $ka = 2\pi a/\lambda$ is in units of radians, and therefore $\cos\theta_{first\ null} = \pm\lambda/Na$. If $\theta_{first\ null} \equiv \pi/2 \pm \theta_{fn}$, where θ_{fn} is the null angle measured from the x-y plane rather than from the z axis, then we have $\cos\theta_{first\ null} = \sin\theta_{fn} \cong \theta_{fn}$ and:

$$\theta_{fn} \cong \pm\lambda/Na \quad [\text{radians}] \qquad (10.4.13)$$

The following simple geometric argument yields the same answer. Figure 10.4.5(b) shows that the first null of this 6-dipole array occurs when the rays from the first and fourth dipole element cancel, for then the rays from the second and fifth, and the third and sixth will also cancel. This total cancellation occurs when the delay between the first and fourth ray is $\lambda/2$, which corresponds to the angle $\theta_{fn} = \pm\sin^{-1}[(\lambda/2)/(aN/2)] \cong \pm\lambda/aN$.
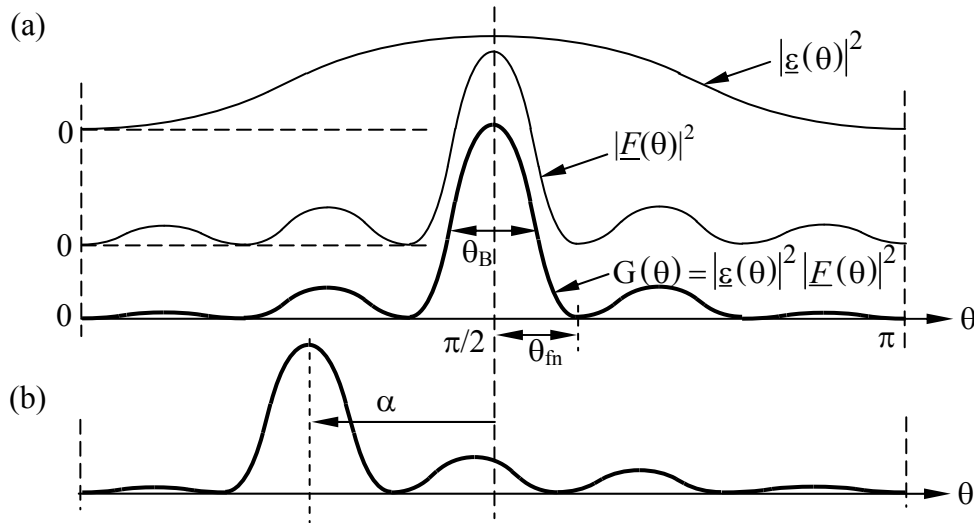
Figure 10.4.6 Antenna pattern for N-element linear dipole array.

The angle θ_{fn} between the beam axis and the first null is approximately the half-power beamwidth θ_B of an N-element antenna array. The antenna gain $G(\theta)$ associated with (10.4.11) for N = 6, $\psi = 0$, and $a = \lambda/2$ is sketched in Figure 10.4.6(a), together with the squares of the

array factor [from (10.4.10)] and element factor [from (10.4.6)]. In this case $\theta_{fn} = \pm\sin^{-1}(2/N) \cong \pm2/N$ radians $\cong \theta_B$.

If $\psi \neq 0$ so that the excitation phase varies linearly across the array, then the main beam and the rest of the pattern is "squinted" or "scanned" to one side by angle α. Since a phase delay of ψ is equivalent to a path delay of δ, where $\psi = k\delta$, and since the distance between adjacent dipoles is a, it follows that adjacent rays for a scanned beam will be in phase at angle $\theta = \pi/2 + \sin^{-1}(\delta/a) = \pi/2 + \alpha$, where:

$$\alpha = \sin^{-1}(\delta/a) = \sin^{-1}(\psi\lambda/2\pi a) \qquad\qquad \text{(scan angle)} \quad (10.4.14)$$

as sketched in Figure 10.4.6(b) for the case $\psi = 2$ radians, $a = \lambda/2$, and $\alpha \cong 40^o$.

Note that larger element separations a can produce multiple main lobes separated by smaller ones. Additional main lobes appear when the argument $(\psi + ka\cos\theta)/2$ in the denominator of (10.4.11) is an integral multiple of π so that the denominator is zero; the numerator is zero at the same angles, so the ratio is finite although large. To preclude multiple main lobes the spacing should be $a < \lambda$, or even $a < \lambda/2$ if the array is scanned.

Example 10.4D
A uniform row of 100 x-oriented dipole antennas lies along the z axis with inter-dipole spacing a $= 2\lambda$. At what angles θ in the y-z plane is the gain maximum? See Figure 10.4.5 for the geometry, but note that the dipoles for our problem are x-oriented rather than z-oriented. What is the angle Δ between the two nulls adjacent to $\theta \cong \pi/2$? What is the gain difference $\Delta G(dB)$ between the main lobe at $\theta = \pi/2$ and its immediately adjacent sidelobes? What difference in excitation phase ψ between adjacent dipoles is required to scan these main lobes 10° to one side?

Solution: The gain is maximum when the rays from adjacent dipoles add in phase, and therefore all rays add in phase. This occurs at $\theta = 0$, $\pm\pi/2$, and $\pm\sin^{-1}(\lambda/a) \cong 30°$ [see Figure 10.4.5(b) for the approximate geometry, where we want a phase lag of λ to achieve a gain maximum]. The nulls nearest $\theta = \pi/2$ occur at that θ_{fn} when the rays from the first and 51st dipoles first cancel [see text after (10.4.13)], or when

$$\theta_{fn} = \frac{\pi}{2} \pm \sin^{-1}\left(\frac{\lambda/2}{aN/2}\right) = \frac{\pi}{2} \pm \sin^{-1}\left(\frac{\lambda/2}{2\lambda N/2}\right) \cong \frac{\pi}{2} \pm \frac{1}{2N} ; \text{ thus } \Delta = 1/N \text{ radians} \cong 0.57^o.$$

The array factors for this problem and Figure 10.4.5(a) are the same, so (10.4.10) applies. Near $\theta \cong \pi/2$ the element factor is approximately constant and can therefore be ignored because we seek only gain ratios. We define $\beta \equiv \pi/2 - \theta$ so $\cos\theta$ becomes $\sin\beta$. Therefore (10.4.11) becomes $G_o(\theta) \propto |F(\theta,\phi)|^2 \propto \sin^2(Nk\lambda\sin\beta)/\sin^2(k\lambda\sin\beta)$ where $\psi = 0$. $\beta \ll 1$, so $\sin\beta \cong \beta$. Similarly, $k\lambda\beta \ll 1$ so $\sin(k\lambda\beta) \cong k\lambda\beta$. Thus $G_o(\beta = 0) \propto \sim (Nk\lambda\beta)^2/(k\lambda\beta)^2 = N^2$, and the first

adjacent peak in gain occurs when $Nk\lambda\sin\beta_{first\ peak} = 1$, so $G(\beta = \beta_{first\ peak}) \propto$

$\sim (k\lambda\beta_{first\ peak})^{-2}$. The numerator is unity when $Nk\lambda\beta_{first\ peak} \cong 3\pi/2$, or

$\beta_{first\ peak} \cong 3\pi/2Nk\lambda = 3/(4N)$. Therefore $G(\beta = \beta_{first\ peak}) \propto \sim (2\pi3/4N)^{-2} \cong 0.045N^2$, which is $10\log_{10}(0.045) = -13.5\ dB$ relative to the peak N^2. A 10° scan angle requires the rays from adjacent dipoles to be in phase at that angle, and therefore the physical lag δ meters between the two rays must satisfy $\sin\beta_{scan} = \delta/a = \delta/2\lambda$. The corresponding phase lag in the leading dipole is $\psi = k\delta = (2\pi/\lambda)(2\lambda\sin\beta_{scan}) = 4\pi\sin(10°)\ radians = 125°$.

10.4.5 Phasor addition in array antennas

Phasor addition can be a useful tool for analyzing antennas. Consider the linear dipole array of Figure 10.4.5, which consists of N identical z-oriented dipole antennas spaced at distance a equally along the z-axis. In direction θ from the z axis the array factor is the sum of the phasors emitted from each dipole. Figure 10.4.6(a) shows this sum $\underline{A}$ for the x-y plane ($\theta = 90°$) when the dipoles are all excited in phase and N = 8. This yields the maximum possible gain for this antenna. As θ departs from 90° (broadside radiation) the phasors each rotate differently and add to form a progressively smaller sum $\underline{B}$. When the total phasor $\underline{B}$ corresponds to $\Delta\phi$ = 5-degree lag for each successive contribution, then $\theta = \cos^{-1}(\frac{5}{360}\frac{\lambda}{a})$. Figure 10.4.6(b, c, and d) show the sum $\underline{B}$ when $\Delta\phi$ is 45°, 72°, and 90°, respectively. The antenna gain is proportional to $|\underline{B}|^2$. Figures (b) and (d) correspond to radiation angles θ that yield nulls in the pattern ($|\underline{B}| = 0$), while (c) is near a local maximum in the antenna pattern. Because $|\underline{C}|$ is ~0.2$|\underline{A}|$, the gain of this sidelobe is ~0.04 times the maximum gain ($|\underline{C}|^2 \cong 0.04|\underline{A}|^2$), or ~14 dB weaker.[54] The spatial angles θ corresponding to (a) - (d) depend on the inter-dipole distance 'a'.

If a = 2λ, then angles θ from the z axis that correspond to phasor $\underline{A}$ in Figure 10.4.7 (a) are 0°, 60°, 90°, 120°, and 180°; the peaks at 0 and 180° fall on the null of the element factor and can be ignored. The angle from the array axis is θ, and 'a' is the element spacing, as illustrated in Figure 10.4.5. The angles $\theta = 0°$ and 180° correspond to $\cos^{-1}(2\lambda/2\lambda)$, while $\theta = 60°$ and $\theta = 120°$ correspond to $\cos^{-1}(\lambda/2\lambda)$, and $\theta = 90°$ corresponds to $\cos^{-1}(0/2\lambda)$; the numerator in the argument of $\cos^{-1}$ is the lag distance in direction θ, and the denominator is the element spacing 'a'. Thus this antenna has three equal peaks in gain: $\theta = 60, 90,$ and $120°$, together with numerous smaller sidelobes between those peaks.

---

[54] dB $\equiv$ 10 $\log_{10}$N, so N = 0.04 corresponds to ~-14 dB.
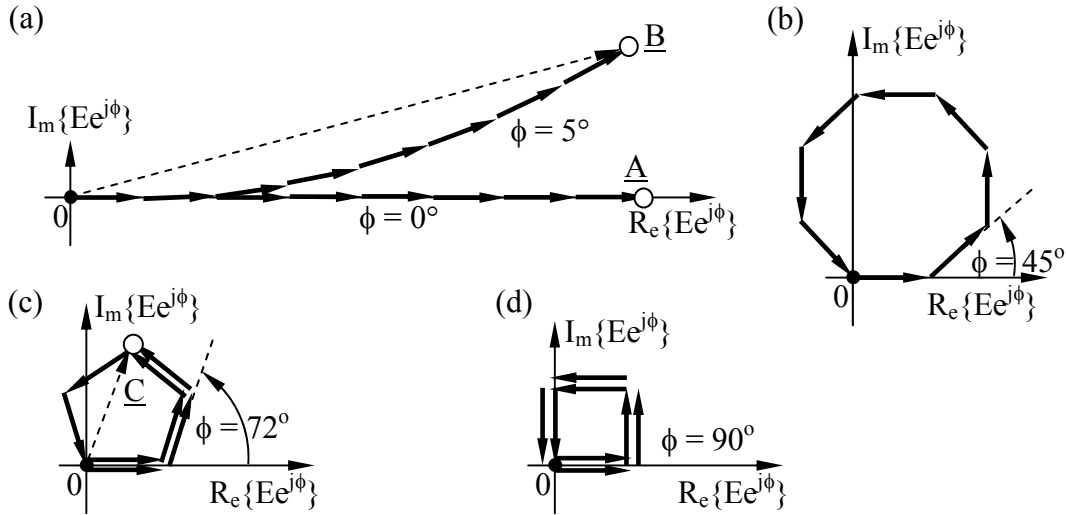
Figure 10.4.7 Phasor addition for an 8-element linear dipole array.

Four small sidelobes occur between the adjacent peaks at 60°, 90°, and 120°. The first sidelobe occurs in each case for $\phi \cong 70°$ as illustrated in Figure 10.4.2(c), i.e., approximately half-way between the nulls at $\phi = 45°$ [Figure 10.4.2(b)] and $\phi = 90°$ [Figure 10.4.2(d)], and the second sidelobe occurs for $\phi \cong 135°$, between the nulls for $\phi = 90°$ [Figure 10.4.2(d)] and $\phi = 180°$ (not illustrated). Consider, for example, the broadside main lobe at $\theta = 90°$; for this case $\phi = 0°$. As θ decreases from 90° toward zero, ϕ increases toward 45°, where the first null occurs as shown in (b); the corresponding $\theta_{null} = \cos^{-1}[(\phi\lambda/360)/2\lambda] = 86.4°$. The denominator 2λ in the argument is again the inter-element spacing. The first sidelobe occurs when $\phi \cong 72°$ as shown in (c), and $\theta \cong \cos^{-1}[(\phi\lambda/360)/2\lambda] = 84.4°$. The next null occurs at $\phi = 90°$ as shown in (d), and $\theta_{null} = \cos^{-1}[(\phi\lambda/360)/2\lambda] = 82.8°$. The second sidelobe occurs for $\phi \cong 135°$, followed by a null when $\phi = 180°$. The third and fourth sidelobes occur for $\phi \cong 225°$ and 290° as the phasor patterns repeat in reverse sequence: (d) is followed by (c) and then (b) and (a) as θ continues to decline toward the second main lobe at $\theta = 60°$. The entire gain pattern thus has three major peaks at 60, 90, and 120°, typically separated by four smaller sidelobes intervening between each major pair, and also grouped near $\theta = 0°$ and 180°.

Example 10.4E
What is the gain G_S of the first sidelobe of an n-element linear dipole array relative to the main lobe G_o as $n \rightarrow \infty$?

Solution: Referring to Figure 10.4.7(c), we see that as $n \rightarrow \infty$ the first sidelobe has an electric field E_{ffS} that is the diameter of the circle formed by the n phasors when $\sum_{i=1}^{n}|E_i| = E_{ffo}$ is ~1.5 times the circumference of that circle, or ~$1.5 \times \pi E_{ffS}$. The ratio of the gains is therefore $G_S/G_o = |E_{ffS}/E_{ffo}|^2 = (1/1.5\pi)^2 = 0.045$, or -13.5 dB.

10.4.6   Multi-beam antenna arrays

Some antenna arrays are connected so as to produce several independent beams oriented in different directions simultaneously; phased array radar antennas and cellular telephone base stations are common examples. When multiple antennas are used for reception, each can be filtered and amplified before they are added in as many different ways as desired. Sometimes these combinations are predetermined and fixed, and sometimes they are adjusted in real time to place nulls on sources of interference or to place maxima on transmitters of interest, or to do both.

The following cellular telephone example illustrates some of the design issues. The driving issue here is the serious limit to network capacity imposed by the limited bandwidth available at frequencies suitable for urban environments. The much broader spectrum available in the centimeter and millimeter-wave bands propagates primarily line-of-sight and is not very useful for mobile applications; lower frequencies that diffract well are used instead, although the available bandwidth is less. The solution is to reuse the same low frequencies multiple times, even within the same small geographic area. This is accomplished using array antennas that can have multiple inputs and outputs.

A typical face of a cellular base station antenna has 3 or 4 elements that radiate only into the forward half-space. They might also have a combining circuit that forms two or more desired beams. An alternate way to use these arrays based on switching is described later. Three such faces, such as those illustrated in Figure 10.4.8(a) with four elements spaced at 3λ, might be arranged in a triangle and produce two sets of antenna lobes, for example, the $\phi = 0$ set and the $\phi = \pi$ set indicated in (b) by filled and dashed lines, respectively.
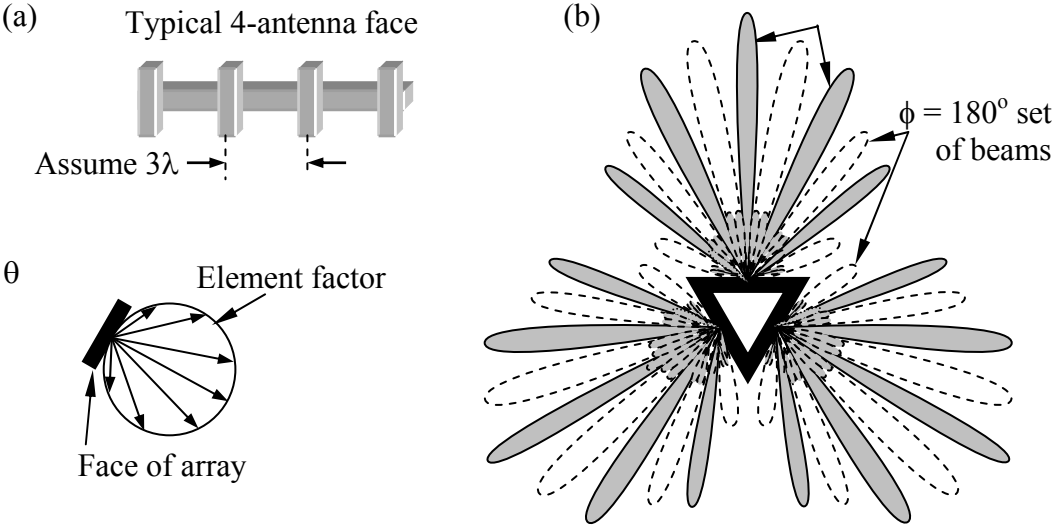
Figure 10.4.8   Cellular base station antenna patterns with frequency reuse.

As before, ϕ is the phase angle difference introduced between adjacent antenna elements. Inter-antenna separations of 3λ result in only 5 main lobes per face, because the two peaks in the plane

of each face are approximately zero for typical element factors. Between each pair of peaks there are two small sidelobes, approximately 14 dB weaker as shown above.

These two sets ($\phi = 0$, $\phi = \pi$) can share the same frequencies because digital communication techniques can tolerate overlapping signals if one is more than ~10-dB weaker. Since each face of the antenna can be connected simultaneously to two independent receivers and two independent transmitters, as many as six calls could simultaneously use the same frequency band, two per face. A single face would not normally simultaneously transmit and receive the same frequency, however. The lobe positions can also be scanned in angle by varying ϕ so as to fill any nulls. Designing such antennas to maximize frequency reuse requires care and should be tailored to the distribution of users within the local environment. In unobstructed environments there is no strong limit to the number of elements and independent beams that can be used per face, or to the degree of frequency reuse. Moreover, half the beams could be polarized one way, say right-circular or horizontal, and the other half could be polarized with the orthogonal polarization, thereby doubling again the number of possible users of the same frequencies. Polarization diversity works poorly for cellular phones, however, because users orient their dipole antennas as they wish.

In practice, most urban cellular towers do not currently phase their antennas as shown above because many environments suffer from severe *multipath* effects where reflected versions of the same signals arrive at the receiving tower from many angles with varying delays. The result is that at each antenna element the phasors arriving from different directions with different phases and amplitudes will add to produce a net signal amplitude that can be large or small. As a result one of the elements facing a particular direction may have a signal-to-interference ratio that is more than 10 dB stronger than another for this reason alone, even though the antenna elements are only a few wavelengths away in an obstacle-free local environment. Signals have different differential delays at different frequencies and therefore their peak summed values at each antenna element are frequency dependent. The antenna-use strategy in this case is to assign users to frequencies and single elements that are observed to be strong for that user, so that another user could be overlaid on the same frequency while using a different antenna element pointed in the same direction. The same frequency-reuse strategy also works when transmitting because of reciprocity.

That signal strengths are frequency dependent in multipath environments is easily seen by considering an antenna receiving both the direct line-of-sight signal with delay t_1 and a reflected second signal with comparable strength and delay t_2. If the differential lag $c(t_2 - t_1) = n\lambda = D$ for integer n, then the two signals will add in phase and reinforce each other. If the lag $D = (2n + 1)\lambda/2$, then they will partially or completely cancel. If $D = 10\lambda$ and the frequency f increases by 10 percent, then the lag measured in wavelengths will also change 10 percent as the sum makes a full peak-to-peak cycle with a null between. Thus the gap between frequency nulls is $\sim\Delta f = f(\lambda/D) = c/D$ Hz. The depth of the null depends on the relative magnitudes of the two rays that interfere. As the number of rays increases the frequency structure becomes more complex. This phenomenon of signals fading in frequency and time as paths and frequencies change is called *multipath fading*.

Chapter 11: Common Antennas and Applications

11.1   Aperture antennas and diffraction

11.1.1   Introduction

Antennas couple circuits to radiation, and vice versa, at wavelengths that can extend into the infrared region and beyond. The output of an antenna is a voltage or field proportional to the input field strength $\overline{E}(t)$ and at the same frequency. By this definition, devices that merely amplify, detect, or mix signals are not antennas because they do not preserve phase and frequency, although they generally are connected to the outputs of antennas. For example, some sensors merely sense the increased temperature and heating caused by incoming waves. Chapter 10 introduced short-dipole and small-loop antennas, and arrays thereof. Chapter 11 continues with an introductory discussion of aperture antennas and diffraction in Section 11.1, and of wire antennas in 11.2. Applications are then discussed in Section 11.4 after surveying the basics of wave propagation and thermal emission in Section 11.3.   These applications include communications, radar and lidar, radio astronomy, and remote sensing.   Most optical applications are deferred to Chapter 12.

11.1.2   Diffraction by apertures

Plane waves passing through finite openings emerge propagating in all directions by a process called *diffraction*. Antennas that radiate or receive plane waves within finite apertures are *aperture antennas*. Examples include the parabolic reflector antennas used for radio astronomy, radar, and receiving satellite television signals, as well as the lenses and finite apertures employed in cameras, microscopes, telescopes, and many optical communications systems. As in the case of dipole antennas; we assume reciprocity and knowledge of the source fields or equivalent currents.

Since we have already derived expressions for fields radiated by arbitrary current distributions, one approach to finding aperture-radiated fields is to determine current distributions equivalent to the given aperture fields.   Then these equivalent currents can be replaced by a continuous array of Hertzian dipoles for which we know the radiated far fields.

Consider a uniform current sheet $\overline{J}$ [A m^{-1}] occupying the x-z plane, as illustrated in Figure 11.1.1. Maxwell's equations are then satisfied by:

$$\overline{E} = \hat{z}\underline{E}_o e^{-jky} \qquad\qquad \overline{H} = \hat{x}\left(\underline{E}_o/\eta_o\right)e^{-jky} \qquad\qquad \text{(for y > 0)} \quad (11.1.1)$$

$$\overline{E} = \hat{z}\underline{E}_o e^{+jky} \qquad\qquad \overline{H} = -\hat{x}\left(\underline{E}_o/\eta_o\right)e^{+jky} \qquad\qquad \text{(for y < 0)} \quad (11.1.2)$$
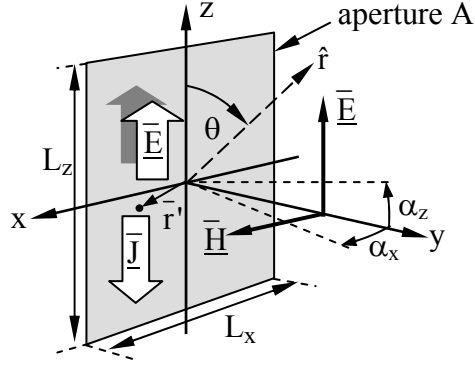
Figure 11.1.1   Aperture radiation from an equivalent current sheet.

The electric field $\hat{z}\underline{E}_o \equiv \overline{\underline{E}}_o$ must satisfy the boundary condition (2.6.11) that:

$$\overline{J}_s = \hat{y} \times \left[\overline{H}(y = 0_+) - \overline{H}(y = 0_-) \right] = \hat{y} \times \left[\hat{x}\frac{E_o}{\eta_o} + \hat{x}\frac{E_o}{\eta_o} \right] \qquad (11.1.3)$$

$$\overline{J}_s = -\hat{z}2\frac{E_o}{\underline{\eta_o}} = -2\frac{\overline{E}_o}{\eta_o} \left[Am^{-1} \right] \qquad (11.1.4)$$

Therefore we can consider any plane wave emerging from an aperture as emanating from an equivalent current sheet $\overline{J}_s$ given by (11.1.4) provided we neglect radiation from the charges and currents induced at the aperture edges. They can generally be neglected if the aperture is large compared to a wavelength and if we remain close to the y axis in the direction of propagation, because then the aperture area dominates the observable radiating area. This approximation (11.1.4) for a finite aperture is valid even if the strength of the plane wave varies across the aperture slowly relative to a wavelength.

The equivalent current sheet (11.1.4) radiates according to (11.1.5) [from (10.2.8)], where we represent the current sheet by an equivalent array of Hertzian dipoles of length dz and current $\overline{I} = \overline{J}_s dx$:

$$\overline{E} = \hat{\theta}\frac{jk\underline{I}d\eta_o}{4\pi r}e^{-jkr}\sin\theta \qquad \text{(far-field radiation)} \quad (11.1.5)$$

The far fields radiated by the z-polarized current sheet $\overline{J}_s(x,z)$ in the aperture A are then:

$$\begin{aligned}
\overline{\underline{E}}_{ff}(\theta,\phi) &\cong \hat{\theta}\frac{j\eta_o}{2\lambda r}\sin\theta\int_A \underline{J}_z(x,z)e^{-jkr(x,z)}dx\,dz \\
&\cong -\hat{\theta}\frac{j}{\lambda r}\sin\theta\int_A \underline{E}_o(x,y)e^{-jkr(x,z)}dx\,dz
\end{aligned} \qquad (11.1.6)$$

To simplify the integral we can assume all rays are parallel by using the Fraunhofer far-field approximation:

$$e^{-jkr(x,z)} \cong e^{-jkr_o} e^{+jk\hat{r}\cdot\overline{r}'} \qquad (\textit{Fraunhofer approximation}) \qquad (11.1.7)$$

where we define position within the aperture $\overline{r}' \equiv x\hat{x} + z\hat{z}$, and the distance $r_o = (x^2 + y^2 + z^2)^{0.5}$. The Fraunhofer approximation is generally used when $r_o > 2L^2/\lambda$. Then:

$$\overline{E}_{ff}(\theta,\phi) \cong -\hat{\theta}\frac{j}{\lambda r}e^{-jkr_o}\sin\theta \int_A \underline{E}_{oz}(x,z)e^{+jk\hat{r}\cdot\overline{}'}dx\ dz \qquad (11.1.8)$$

Those points in space too close to the aperture for the Fraunhofer approximation to apply lie in the *Fresnel region* where $r < \sim 2L^2/\lambda$, as shown in (11.1.4). If we restrict ourselves to angles close to the y axis we can define the angles α_x and α_z from the y axis in the x and z directions, respectively, as illustrated in Figure 11.1.1, so that:

$$\hat{r}\cdot\overline{r}' \cong x\sin\alpha_x + z\sin\alpha_z \cong x\alpha_x + z\alpha_z \qquad (11.1.9)$$

Therefore, close to the y axis (11.1.8) can be approximated[55] as:

$$\overline{E}_{ff}(\alpha_x,\alpha_z) \cong -\hat{\theta}\frac{j}{\lambda r}e^{-jkr_o}\int_A \underline{E}_{oz}(x,z)e^{+j2\pi(x\alpha_x + z\alpha_z)/\lambda}dx\ dz \qquad (11.1.10)$$

which is the Fourier transform of the aperture field distribution $\underline{E}_{oz}(x,z)$, times a factor that depends on distance r and wavelength λ. Unlike the usual Fourier transform for converting signals between the time and frequency domains, this reversible transform in (11.1.10) is between the aperture spatial domain and the far-field angular domain.

For reference, the *Fourier transform* for signals is:

$$\underline{S}(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}dt \qquad (11.1.11)$$

$$s(t) = \int_{-\infty}^{\infty} \underline{S}(f)e^{j2\pi ft}df \qquad (11.1.12)$$

The Fourier transform (11.1.11) has exactly the same form as the integral of (11.1.10) if we replace the aperture coordinates x and z with their wavelength-normalized equivalents x/λ and z/λ, analogous to time t; α is analogous to frequency f.

Assume the aperture of Figure 11.1.1 is z-polarized, has dimensions $L_x \times L_z$, and is uniformly illuminated with amplitude $\underline{E}_o$. Then its far fields can be computed using (11.1.10):

---

[55] In the *Huygen's approximation* a factor of $(1 + \cos\alpha)/2$ is added to improve the accuracy, but this has little impact near the y axis. In this expression α is the angle from the direction of propagation (y axis) in any direction.

$$\overline{E}_{ff}\left(\alpha_x,\alpha_z\right) \cong -\hat{\theta}\frac{j}{\lambda r}e^{-jkr_o}\int_{-L_z/2}^{+L_z/2}e^{+j2\pi\alpha_z z/\lambda}\int_{-L_x/2}^{+L_x/2}\underline{E}_o\left(x,z\right)e^{+j2\pi\alpha_x x/\lambda}dx\ dz \quad (11.1.13)$$

The inner integral yields:

$$\int_{-L_x/2}^{+L_x/2}\underline{E}_o\left(x,z\right)e^{+j2\pi\alpha_x x/\lambda}dx = \underline{E}_o\frac{\lambda}{j2\pi\alpha_x}\left[e^{+j\pi\alpha_x L_x/\lambda} - e^{-j\pi\alpha_x L_x/\lambda}\right]$$
$$= \underline{E}_o\frac{\sin\left(\pi\alpha_x L_x/\lambda\right)}{\pi\alpha_x/\lambda} \quad (11.1.14)$$

The outer integral yields a similar result, so the far field is:

$$\overline{E}_{ff}\left(\theta,\phi\right) \cong -\hat{\theta}\frac{j}{\lambda r}\underline{E}_{oz}e^{-jkr_o}\bullet L_x L_z\frac{\sin\left(\pi\alpha_x L_x/\lambda\right)}{\pi\alpha_x L_x/\lambda}\frac{\sin\left(\pi\alpha_z L_z/\lambda\right)}{\pi\alpha_z L_z/\lambda} \quad (11.1.15)$$

The total power P_t radiated through the aperture is simply $A\left|\underline{E}_o\right|^2\big/2\eta_o$, where $A = L_x L_z$, so the antenna gain $G(\alpha_x,\alpha_z)$ given by (10.3.1) is:

$$G\left(\alpha_x,\alpha_z\right) \cong \frac{\left|E_{ff}\left(\alpha_x,\alpha_z\right)\right|^2\big/2\eta_o}{P_t\big/4\pi r^2} \quad (11.1.16)$$

$$\cong A\frac{4\pi}{\lambda^2}\left(\frac{\sin^2\left(\pi\alpha_x L_x/\lambda\right)}{\left(\pi\alpha_x L_x/\lambda\right)^2}\right)\left(\frac{\sin^2\left(\pi\alpha_z L_z/\lambda\right)}{\left(\pi\alpha_z L_z/\lambda\right)^2}\right) \quad (11.1.17)$$

The function (sin x)/x appears so often in electrical engineering that it has its own symbol 'sinc(x)'. Note that sinc(0) = 1 since $\sin(x) \cong x - (x^3/6)$ for x << 1. This gain pattern is plotted in Figure 11.1.2. The first nulls occur when $\pi\alpha_i L_i/\lambda = \pi$ (i = x or z), and therefore $\alpha_{null} = \lambda/L$, where a narrower beamwidth α corresponds to a wider aperture L. The on-axis gain is:

$$G\left(0,0\right) = \frac{4\pi}{\lambda^2}A \qquad \text{(gain of uniformly illuminated aperture area A)} \quad (11.1.18)$$

Equation (11.1.18) applies to any uniformly illuminated aperture antenna, and such antennas have on-axis effective areas $A(\theta,\phi)$ that approach their physical areas A, and have peak gains $G_o = 4\pi A/\lambda^2$. The antenna pattern of Figure 11.1.2 vaguely resembles that of circular apertures as well, and the same nominal angle to first null, λ/L, roughly applies to all. Such diffraction patterns largely explain the limiting angular resolution of telescopes, cameras, animal eyes, and photolithographic equipment used for fabricating integrated circuits.
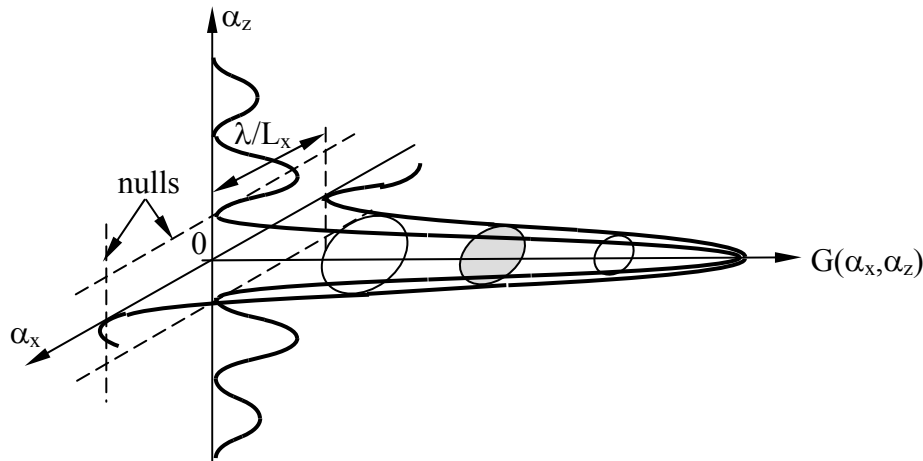
Figure 11.1.2   Antenna gain for uniformly illuminated rectangular aperture.

The coupling between two facing aperture antennas having effective areas A_1 and A_2 is:

$$P_{r_2} = \frac{P_{t_1}}{4\pi r^2} G_1 A_2 = \frac{P_{t_1}}{\lambda^2 r^2} A_1 A_2 = P_{t_1} \left(\frac{\lambda}{4\pi r}\right)^2 G_1 G_2 \qquad (11.1.19)$$

where P_{r_2} and P_{t_1} are the power received by antenna 2 and the total power transmitted by antenna 1, respectively. For (11.1.19) to be valid, $r^2\lambda^2 \gg A_1 A_2$; if $A_1 = A_2 = D^2$, then we require $r \gg D^2/\lambda$ for validity. Otherwise (11.1.19) could predict that more power would be received than was transmitted.

Example 11.1A
What is the angle between the first nulls of the diffraction pattern for a visible laser ($\lambda = 0.5$ microns) illuminating a 1-mm square aperture (about the size of a human iris)? What is the approximate diffraction-limited angular resolution of the human visual system? How does this compare to the maximum angular diameters of Venus, Jupiter, and the moon (~1, ~1, and ~30 arc minutes in diameter, respectively)?

Solution: The first null occurs at $\phi = \sin^{-1}(\lambda/L) \cong 5\times10^{-7}/10^{-3} = 5\times10^{-4}$ radians $= 0.029° \cong 1.7$ arc minutes. This is 70 percent larger than the planets Venus and Jupiter at their points of closest approach to Earth, and ~6 percent of the lunar diameter. Cleverly designed neuronal connections in the human visual system improve on this for linear features, as can a dark-adapted iris, which has a larger diameter.

Example 11.1B
A cell-phone dipole antenna radiates one watt toward a uniformly illuminated square aperture antenna of area A = one square meter. If $P_r = 10^{-9}$ watts are required by the receiver for satisfactory link performance, how far apart r can these two terminals be? Does this depend on the shape of the aperture antenna if A remains constant?

Solution: $P_r = AP_tG_t/4\pi r^2 \Rightarrow r = (AP_tG_t/4\pi P_r)^{0.5} = (1\times1\times1.5/4\pi10^{-9})^{0.5} \cong 10.9$ km The on-axis gain G_t of a uniformly illuminated constant-phase aperture antenna is given by (11.1.18). The denominator depends only on the power transmitted through the aperture A, not on its shape. The numerator depends only on the on-axis far field E_{ff}, given by (11.1.10), which again is independent of shape because the phase term in the integral is unity over the entire aperture. Since the on-axis gain is independent of aperture shape, so is the effective area A since $A(\theta,\phi) = G(\theta,\phi)\lambda^2/4\pi$. The on-axis effective area of a uniformly illuminated aperture approximates its physical area.

11.1.3 Common aperture antennas

Section 11.1.2 derived the basic equation (11.1.10) that characterizes the far fields radiated by aperture antennas excited with z-polarized electric fields $\overline{E}_o(x,z) = \hat{z}\underline{E}_{oz}(x,z)$ in the x-z aperture plane:

$$\overline{E}_{ff}(\theta,\phi) \cong \hat{\theta}\frac{j}{\lambda r}\sin\theta\, e^{-jkr} \oiint_A \underline{E}_{oz}(x,z)e^{jk\hat{r}\bullet\overline{r}'}dx\, dz \qquad (11.1.20)$$

The unit vector $\hat{r}$ points from the antenna toward the receiver and $\overline{r}'$ is a vector that locates $\overline{E}_o(r')$ within the aperture. This expression assumes the receiver is sufficiently far from the aperture that a single unit vector $\hat{r}$ suffices for the entire aperture and that the receiver is therefore in the *Fraunhofer region*. The alternative is the near-field Fresnel region where $r < 2D^2/\lambda$, as discussed in Section 11.1.4; D is the aperture diameter. It also assumes the observer is close to the axis perpendicular to the aperture, say within ~40°. The Huygen's approximation extends this angle further by replacing $\sin\theta$ with $(1 + \cos\beta)/2$, where θ is measured from the polarization axis and β is measured from the y axis:

$$\overline{E}_{ff}(\theta,\phi) \cong \hat{\theta}\frac{j}{2\lambda r}(1+\cos\beta)e^{-jkr} \oiint_A \underline{E}_{oz}(x,z)e^{-jk\hat{r}\bullet\overline{r}'}dx\, dz \quad (\textit{Huygen's approximation}) \quad (11.1.21)$$

Evaluating the on-axis gain of a uniformly excited aperture of physical area A having $\underline{E}_{oz}(x,z) = E_{oz}$ is straightforward when using (11.1.21) because the exponential factor in the integral is unity within the entire aperture. The gain follows from (11.1.16). The results are:

$$\overline{E}_{ff}(0,0) \cong \hat{\theta}\frac{j}{\lambda r}e^{-jkr} \oiint_A \underline{E}_{oz}\, dx\, dz \qquad \text{(on-axis field)} \quad (11.1.22)$$

$$G(0,0) = \frac{\left|\overline{E}_{ff}(0,0)\right|^2/2\eta_o}{P_T/4\pi r^2} \qquad \text{(on-axis gain)} \quad (11.1.23)$$

But the total power P_T transmitted through the aperture area A can be evaluated more easily than the alternative of integrating the radiated intensity $I(\theta,\phi)$ over all angles. The intensity I within the aperture is $|\overline{E}_{oz}|^2/2\eta_o$, therefore:

$$P_T = \frac{|\overline{E}_{oz}|^2}{2\eta_o}A \qquad (11.1.24)$$

Then substitution of (11.1.22) and (11.1.24) into (11.1.23) yields the gain of a uniformly illuminated lossless aperture of physical area A:

$$G(0,0) = \frac{(\lambda r)^{-2}(E_{oz}A)^2/2\eta_o}{(E_{oz}^2/2\eta_o)A/4\pi r^2} = \frac{\lambda^2}{4\pi}A \qquad \text{(gain of uniform aperture)} \quad (11.1.25)$$

The off-axis gain of a uniformly illuminated aperture depends on its shape, although the on-axis gain does not.

Perhaps the most familiar radio aperture antennas are parabolic dishes having a point feed that radiates energy toward a *parabolic mirror* so as to produce a planar wave front for transmission, as suggested in Figure 11.1.3(a). Conversely, incoming radiation is focused by the mirror on the *antenna feed*, which intercepts and couples it to a transmission line connected to the receiver. Typical focal lengths (labeled "f" in the figure) are ~half the diameter D for radio systems, and are often much longer for optical mirrors that produce images.
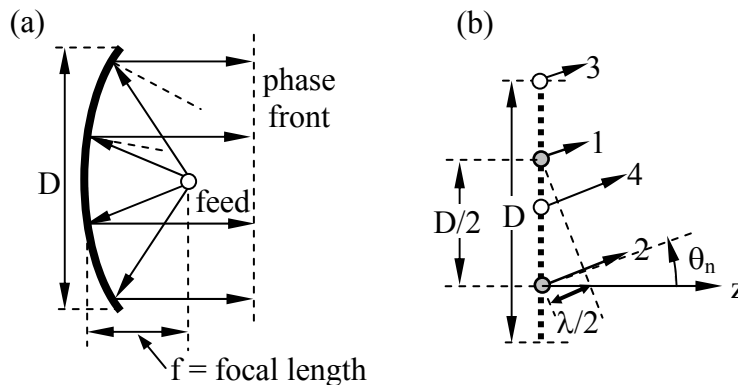
Figure 11.1.3   Aperture antennas and angle of first null.

Figure 11.1.3(b) suggests the angle θ_n at which the first null of a uniformly illuminated rectangular aperture of width D occurs; it is the angle at which all the phasors emanating from each point on the aperture integrate in (11.1.10) to zero. In this case it is easy to pair the phasors originating D/2 apart so each pair cancels at θ_n. For example, radiation from aperture element 2 has to travel $\lambda/2$ farther than radiation from element 1 and therefore they cancel each other. Similarly radiation from elements 3 and 4 cancel, and the sum of all such pairs cancel at the null angle:

$$\theta_n = \sin^{-1}(\lambda/D) \text{ [radians]} \tag{11.1.26}$$

where $\theta_n \cong \lambda/D$ for $\lambda/D \ll 1$.

Approximately the same null angle results for uniformly illuminated circular apertures, for which integration yields $\theta_n \cong 1.2\lambda/D$. Consider the human eye, which has a pupil that normally is ~2 mm in diameter, but can dilate to ~1 cm in the dark. For a wavelength of 5×10^{-7} meters, we find the normal diffraction-limited angular resolution of the eye is $\sim\lambda/D = 5\times10^{-7}/(2\times10^{-3}) = 2.5\times10^{-4}$ radians or ~0.014 degrees, or ~0.9 arc minutes. For comparison, the planets Venus and Jupiter are approximately 1 arc minute in diameter at closest approach, and the moon and sun are approximately 30 arc minutes in diameter.

A large astronomical telescope like the 200-inch system at Palomar has a nominal diffraction limit of $\lambda/D \cong 5\times10^{-7}/5.08 \cong 10^{-7}$ radians or ~0.02 arc seconds, where there are 60 arc seconds in an arc minute, and 60 arc minutes in a degree. This is adequate to resolve an automobile on the moon. Unfortunately mirror surface imperfections, focus misplacement, and atmospheric turbulence limit the actual angular resolution of Palomar to ~1 arc second on the very best nights; normal daytime turbulence is far worse.

Practical issues generally shape the design of parabolic radio antennas. First, mechanical (gravity and wind) and thermal issues (temperature gradients) usually limit their angular resolution to ~1 arc minute; most antennas are too small relative to λ to achieve this resolution, however. Second, the antenna feed that illuminates the parabola tends to spray its radiation in a broad pattern that extends past the edge of the reflector creating backlobes. Third, the finite extent of the aperture results in an antenna pattern with sidelobes and unwanted responsiveness to directions beyond the main lobe.

Equation (11.1.10) showed how the angular dependence of the far-fields of an aperture was proportional to the Fourier transform of the aperture excitation function. For example, (11.1.17) and Figure 11.1.2 showed the radiation pattern of a uniformly illuminated aperture measuring L_x by L_z. Significant energy was radiated beyond the first nulls at $\alpha_x = \lambda/L_x$ and $\alpha_z = \lambda/L_z$. A finite aperture necessarily radiates something at all angles, just as a finite voltage pulse in a circuit has at least some energy at all frequencies; the sharper the pulse edges, the more high-frequency content they have. Therefore, reducing the sharp discontinuities in field strength at the aperture edge, a strategy called tapering, can reduce diffraction sidelobes. Antenna feeds are typically designed to reduce field strengths by factors of 2-4 at the mirror edges for this reason, but the resulting effective reduction in aperture diameter produces a slightly broader main lobe, just as the Fourier transform of a narrower pulse produces a broader spectral band.

A final consideration is sometimes important when designing aperture antennas, and that involves aperture blockage, which results when transmitted radiation reflected from the mirror is blocked or scattered by the antenna feed at the focus of the parabola. Not only does the scattered radiation contribute to side or back lobes, but it also is lost to the main beam. Example 11.1C illustrates these issues.

Example 11.1C
A uniformly illuminated square aperture is 1000 wavelengths long on each side. What is its antenna gain $G(\alpha_x, \alpha_z)$ for $\alpha \ll 1$? What is the gain G_o' if the center of this fully illuminated aperture is blocked by a square absorber 100 wavelengths on a side? What is the extent and approximate magnitude of the sidelobes introduced by the blockage?

Solution: The on-axis gain $G_o = A4\pi/\lambda^2 = 1000^2 \times 4\pi$. The angular dependence is proportional to the square of the far-field, $|E(\alpha_x, \alpha_z)|^2$, where the far field is the Fourier transform of the aperture field distribution. The full solution for $G(\alpha_x, \alpha_z)$ is developed in Equations (11.1.13–17). If the blocked portion of the aperture is illuminated so the energy there is absorbed, then the total transmitted power P_t in the expression (11.1.16) for gain is unchanged, while the area over which $\overline{E}_{ff}'$ is integrated in (11.1.13) is reduced by the 1 percent blockage (100^2 is 1 percent of 1000^2). Therefore $|\overline{E}_{ff}'(0,0)|^2$, the numerator of (11.1.16), and $G_o'(0,0)$ are all reduced by a factor of $0.99^2 \cong 0.98$. Thus $G_o' \cong 0.98\ G_o$. If the blocked portion of the aperture were not illuminated so as to avoid the one percent absorption, then $G_o'(0,0)$ would be reduced by only 1 percent: $G_o' = A'4\pi/\lambda^2$. The sidelobes for the blocked aperture follow from the Fourier transform (11.1.13), where the aperture excitation $E_o(x,z)$ is the sum of a positive square "boxcar" function 1000λ on a side, and a negative square boxcar 100λ on a side. Since this transform is linear, $\overline{E}_{ff}(\alpha_x, \alpha_z)$ is the sum of the transforms of the positive and negative boxcar functions, and the antenna sidelobes therefore have contributions from each. Most important is the main lobe of the diffraction pattern of the smaller "blockage" boxcar, which has magnitude $\sim0.01^2$ that of G_o', and a half-power beamwidth θ_{BB} that is 10 times greater than the main lobe of the larger boxcar: $\theta_{BB} \cong \lambda/D_B = \lambda/100\lambda$. The total antenna pattern is the square of the summed transforms and more complicated; the innermost few sidelobes are approximately those of the original antenna, while the blockage-induced sidelobes are more important at greater angles.

11.1.4   Near-field diffraction and Fresnel zones

Often receivers are sufficiently close to the source that the Fraunhofer parallel-ray approximation of (11.1.7) is invalid. Then the *Huygen's approximation* (11.1.21) can be used:

$$\overline{E}_{ff} \cong \hat{\theta}\frac{j}{2\lambda r}(1+\cos\beta)\iint_A E_{oz}(x,z)e^{-jk\,r(x,y)}dx\ dz \qquad \text{(Huygen's approximation)} \quad (11.1.27)$$

for which the distance between the receiver and the point x,y in the aperture is defined as r(x,y). This region close to a source or obstacle where the Fraunhofer approximation is invalid is called the *Fresnel region*.

If the phase of $\underline{E}_{oz}$ in the source aperture is constant everywhere, then contributions to $\underline{E}_{ff}(0,0)$ from some parts of the aperture will tend to cancel contributions from other parts because they are out of phase. For example, contributions from the central circular zone where $r(x,y)$ ranges from r_o to $r_o + \lambda/2$ will largely cancel the contributions from the surrounding ring where $r(x,y)$ ranges from $r_o + \lambda/2$ to $r_o + \lambda$; it is easily shown that these two rings have approximately the same area, as do all such rings over which the delay varies by $\lambda/2$.[56] Such rings are illustrated in Figure 11.1.4(a).
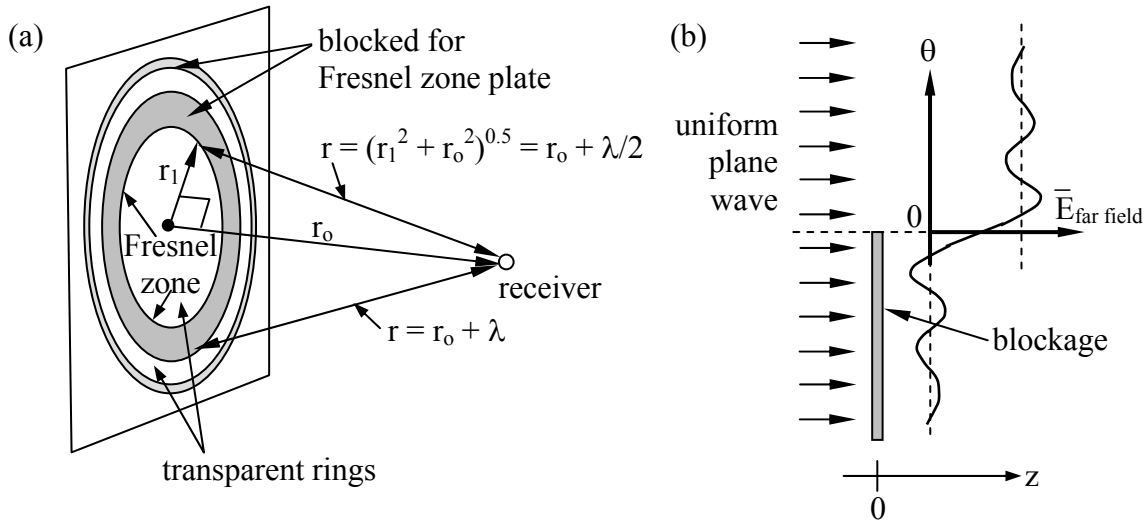
Figure 11.1.4   Fresnel zone plate.

One technique for maximizing diffraction toward an observer is therefore simply to physically block radiation from those alternate zones contributing negative fields, as suggested in Figure 11.1.4(a). Such a blocking device is called a *Fresnel zone plate*. The central ring having positive phase is called the *Fresnel zone*. Note that if only the central zone is permitted to pass, the received intensity is maximum, and if the first two zones pass, the received intensity is nearly zero because they have approximately the same area. The second zone is weaker, however, because r and θ are larger. Three zones can yield nearly the same intensity as the first zone alone because two of the three zones nearly cancel, and so on. By blocking alternate zones the received intensity can be many times greater than if there were no blockage at all. Thus a multi-ring zone plate acts as a lens by focusing energy received over a much larger area than would be intercepted by the receiver alone. This type of lense is particularly valuable for focusing very short-wave radiation such as x-rays which are difficult to reflect or diffract using traditional mirrors or lenses.

Another advantage of zone plate lenses is that they can be manufactured lithographically, and their critical dimensions are usually many times larger than the wavelengths involved. For example, an x-ray zone plate designed to operate at $\lambda = 10^{-8}$ [m] at a distance r_o of one centimeter

---

[56] The area of the inner circle (radius a) is $\pi a^2 = \pi[(r_o + \lambda/2)^2 - r_o^2] \cong \pi r_o\lambda$ if $\lambda \ll 2r_o$. The area of the immediately surrounding Fresnel ring (radius b) is $\pi(b^2 - a^2) = \pi[(r_o + \lambda)^2 - r_o^2] - \pi[(r_o + \lambda/2)^2 - r_o^2] \cong \pi r_o\lambda$, subject to the same approximation. Similarly, all other Fresnel rings can be shown to have approximately the same area if $\lambda \ll 2r_o$.

will have a central zone of diameter $2[(r_o + \lambda/2)^2 - r_o^2]^{0.5} \cong 2(r_o\lambda)^{0.5} = 2\times10^{-5}$ meters, a dimension easily fabricated using modern semiconductor lithographic techniques.

Another example of diffraction is wireless communications in urban environments, which often involves line-of-sight reception of waves past linear obstacles slightly to one side or slightly obscuring the source. Again Huygen's equation (11.1.27) can be used to determine the result. Referring to Figure 11.1.4(b), if there is no blockage, traditional equations can be used to compute the received intensity. If exactly half the path is blocked by a wall obscuring the bottom half of the illuminated aperture, for example, then the integral in (11.1.27) will yield exactly half the previous value of E_{ff}, and the power (proportional to E_{ff}^2) will be reduced by a factor of four, or ~6dB. If the observer moves up or down less than ~half the radius of the Fresnel zone, then the received power will vary only modestly. For example, an FM radio (say 10^8 Hz) about 100 meters beyond a tall wide metal wall can have a line of sight that passes through the wall a distance of $\sim(r\lambda)^{0.5}/2 = 17$ meters below its top without suffering great loss; $(r\lambda)^{0.5}$ is the radius of the Fresnel zone. Conversely, a line-of-sight that passes less than ~17 meters above the top of the wall will also experience modest diffractive effects.

The Fresnel region approximately begins when the central ray arrives at distance r_o, more than $\lambda/16$ ahead of rays from the perimeter of an aperture of diameter D. That is:

$$\sqrt{\left(\frac{D}{2}\right)^2 + r_o^2} - r_o \overset{\sim}{>} \frac{\lambda}{16} \tag{11.1.28}$$

For D << R this becomes:

$$r_o\left(\sqrt{\left(\frac{D}{2r_o}\right)^2 + 1} - 1\right) \cong \frac{D^2}{8R} \overset{\sim}{>} \frac{\lambda}{16} \tag{11.1.29}$$

Therefore the *Fresnel region* is:

$$r_o \overset{\sim}{<} \frac{2D^2}{\lambda} \qquad\qquad\qquad \text{(Fresnel region)} \quad (11.1.30)$$

11.2   Wire antennas

11.2.1   Introduction to wire antennas

Exact solution of Maxwell's equations for antennas is difficult because antennas typically have complex shapes for which it is difficult to match boundary conditions. Often complex wave expansions with many degrees of freedom are required, and even modern software tools can be challenged. Fortunately, most common *wire antennas* permit their current distributions to be guessed accurately relative to the given terminal current, as explained in Section 11.2.2. Once the current distribution is known everywhere, the radiated fields, radiation and dissipative

resistance, antenna gain, and antenna effective area can be calculated. If the antenna is used at a frequency far from resonance, the reactance can also be estimated.

If the antenna is small compared to a wavelength λ then its current distribution $\underline{I}$ and the open-circuit voltage $\underline{V}_{Th}$ can be determined using the quasistatic approximation. If the current distribution is known, then the radiated far-fields $\overline{\underline{E}}_{ff}$ can be computed using (10.2.8) by integrating the contributions $\Delta\overline{\underline{E}}_{ff}$ from each short current element $\underline{I}d$ (d is the element length and is replaced by ds in the integral), where:

$$\Delta\overline{\underline{E}}_{ff} = \hat{\theta}\frac{jk\underline{I}d\eta_o}{4\pi r}e^{-jkr}\sin\theta \tag{11.2.1}$$

$$\overline{\underline{E}}_{ff} \cong \frac{jk\eta_o}{4\pi r}\int_S \hat{\theta}\underline{I}(s)e^{-jkr}\sin\theta\,ds \tag{11.2.2}$$

For antennas small compared to λ the factor before the integral of (11.2.2) is nearly constant over the integrated length S, so average values suffice. If the wires run in more than one direction, the definition of $\hat{\theta}$ and θ must change accordingly; θ is defined by the local angle between $\overline{\underline{I}}$ and $\hat{r}$, where $\hat{r}$ is the unit vector pointing from the antenna to the observer, as suggested in Figure 10.2.3. Equation (11.2.2), not surprisingly, reduces to (11.2.1) for a short straight wire carrying constant current $\underline{I}$ over a distance $d \ll \lambda$.

Once the radiated fields are known for a given antenna input current $\underline{I}$, the radiated intensity can be integrated over a sphere surrounding the antenna to yield the total power radiated P_R and the radiation resistance R_r, which usually dominates the resistive component of the antenna impedance and corresponds to power lost through radiation (10.3.16). The radiation resistance is simply related to P_R:

$$R_r = \frac{2P_R}{|\underline{I}|^2}\ [ohms] \qquad\qquad \text{(radiation resistance)} \tag{11.2.3}$$

The open-circuit voltage can also be easily estimated for wire antennas small compared to λ. For example, the open-circuit voltage induced across a short dipole antenna shown in Figure 10.3.1 is simply the projection of the incident electric field $\overline{E}$ on the electrical centers of the two metallic structures comprising the dipole, and Example 10.3D showed how the open-circuit voltage across a loop antenna was proportional to the time derivative of the magnetic flux through it. In both cases the open-circuit voltage reveals the directional properties of the antenna. Computation of the *radiation resistance* requires knowledge of the radiated fields and integration of the radiated power over all angles, however. Equation (10.3.16) showed that the radiation resistance of a short dipole antenna of length d is $(2\pi\eta_o/3)(d/\lambda)^2$ ohms. Slightly more complicated integrals over angles yield the radiation resistance for half-wave dipoles of length d and N-turn loop antennas of diameter $d \ll \lambda$: ~73 ohms and ~$1.9\times10^4N^2(d/\lambda)^4$ ohms, respectively. The higher radiation resistance of loop antennas often makes them the antenna of

choice when space is limited relative to wavelength, particularly when they are wound on a ferrite core ($\mu \gg \mu_o$) that increases their magnetic dipole moment.

Most wire antennas are not small compared to a wavelength, however, and the methods of the next section are then often used.

11.2.2   Current distribution on wires

The current distribution on wires is governed by Maxwell's equations, which are most easily solved for simple geometries such as that of a coaxial cable, as discussed in Example 7.1B. The fields for a TEM wave in a coaxial cable are cylindrically symmetric and a function of radius r:

$$\overline{E}(r,z) = \hat{r}\underline{E}_o(z)/r \left[V\ m^{-1} \right]$$ (coaxial cable electric field)   (11.2.4)

$$\overline{H}(r,z) = \hat{\theta}\underline{H}_o(z)/r \left[A\ m^{-1} \right]$$ (coaxial cable magnetic field)[57]   (11.2.5)

The energy density and Poynting's vector are proportional to field strength squared, so they decay as r^{-2}. Therefore the electromagnetic behavior of the line is dominated by the geometry near the central conductor where most of the electromagnetic energy is located, and the outer conductor can be deformed substantially before the fields near the center are significantly perturbed. For example, two-thirds of the power propagates within 10 cm of a 1-mm wire centered within a 1-meter outer cylinder, even though this represents only one-percent of the volume. This is easily shown by integrating the energy density from radius a to radius b, $\int_a^b E_o^2 r^{-2} 2\pi r\ dr = 2\pi E_o^2 \ln(b/a)$, and comparing the results for different sub-volumes.

Therefore the fields near the axis of the coaxial cable illustrated in Figure 11.2.1(a) are altered but little if the outer conductor is replaced by a ground plane as illustrated in Figure 11.2.1(b), or even by a second wire, as shown in Figure 11.2.1(c). The significance of Figure 11.2.1 is therefore that current distributions on thin wire antennas closely resemble those on equivalent TEM lines, provided the lines are not so many wavelengths long that the energy is lost before it reaches the end, or so tightly bent that the segments induce strong voltages on their neighbors. This TEM approximation is valid for understanding the examples of this section.

A widely used antenna is the half-wave dipole, illustrated in Figure 11.2.1(d), which exhibits essentially no reactive impedance because the electric and magnetic energy storages approximately balance. The radiation resistance for any half-wave dipole in free space is ~73 ohms. Section 7.4.2 discusses how these energies balance in any TEM structure of length D = $n\lambda/2$ where n is an integer. Typical bandwidths $\Delta\omega$ of a half-wave dipole are $\Delta\omega/\omega_o = 1/Q \cong 0.1$, where $Q = \omega_o w_T/P_d$, as discussed in Section 7.4.3 and (7.4.4).

---

[57] The magnetic field around a central wire, $H = I/2\pi r$, was given in (1.4.3).
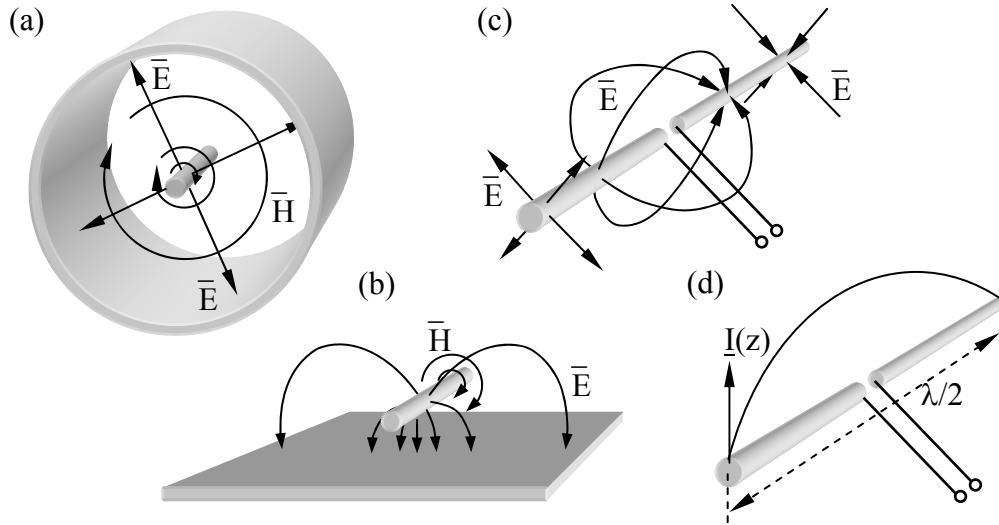
Figure 11.2.1   Fields near wire antennas resemble fields in TEM coaxial cables.

Figure 11.2.2 illustrates nominal current distributions on several antenna structures; these currents are consistent with those on comparable TEM lines propagating signals at the speed of light. The current distributions in the figure represent instantaneous distributions at the moment of current maximum.
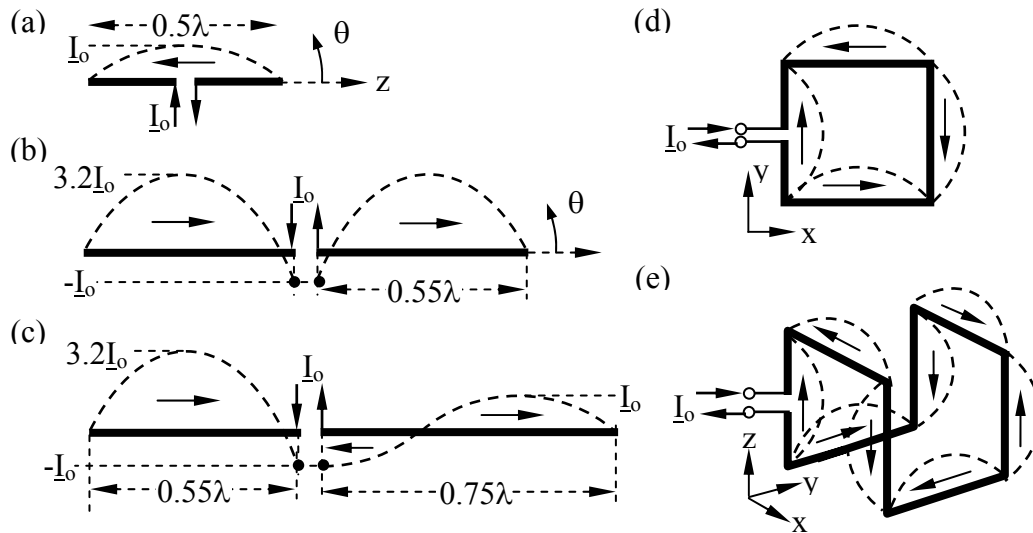
Figure 11.2.2   Current distributions on wire antenna structures.

In these idealized cases the currents everywhere on the antenna approach zero one-quarter cycle later as the energy all converts from magnetic to electric. The voltage distributions when the currents are zero resemble those on the equivalent TEM lines, and are offset spatially by $\lambda/4$; at

resonance the voltage peaks coincide with the current nulls. For example, the voltages and currents for Figure 11.2.2(a) resemble those of the open-circuited TEM resonator of Figure 7.4.1(a). The actual current and voltage distributions are slightly different from those pictured because radiation tends to weaken the currents farther from the antenna terminals, and because such free-standing or bent wires are not true TEM lines.

Figure 11.2.2(b) illustrates how terminal currents (I_o) can be made less than one-third the peak currents (3.2 I_o) flowing on the antenna simply by lengthening the two arms so they are each slightly longer than $\lambda/2$ so the current is close to a null at the terminals. Because smaller terminal currents thus correspond to larger antenna currents and radiated power, the effective radiation resistance of this antenna is increased well above the nominal 73 ohms of the half-wave dipole of (a). The reactance is slightly capacitive, however, and should be canceled with an inductor. Figure 11.2.2(c) illustrates how the peak currents can be made different in the two arms; note that the currents fed to the two arms must be equal and opposite, and this fact forces the two peak currents in the arms to differ. Figures (d) and (e) show more elaborate configurations, demonstrating that wire antennas do not have to lie in a straight line. The patterns for these antennas are discussed in the next section.

11.2.3   Antenna patterns

Once the current distributions on wire antennas are known, the antenna patterns can be computed using (11.2.2). Consider first the dipole antenna of Figure 11.2.2(a) and let its length be d, its terminal current be $I_o{'}$, and its maximum current be I_o. Then (11.2.2) becomes:

$$\overline{E}_{ff} \cong \frac{jk\eta_o}{4\pi r} \int_{-d/2}^{d/2} \hat{\theta}\underline{I}(s)e^{-jkr}\sin\theta\,ds \qquad (11.2.6)$$

$$\overline{E}_{ff} \cong \hat{\theta}\frac{j\eta_o I_o e^{-jkr}}{2\pi r\sin\theta}\left[\cos\left(\tfrac{kd}{2}\cos\theta\right) - \cos\left(\tfrac{kd}{2}\right)\right] \qquad (11.2.7)$$

This expression, which requires some effort to derive, applies to symmetric dipole antennas of any modest length d; I_o is the maximum current, which is not necessarily the terminal current. The common half-wave dipole has $d = \lambda/2$, so (11.2.7) reduces to:

$$\overline{E}_{ff} \cong \hat{\theta}\left(j\eta_o I_o e^{-jkr}\big/2\pi r\sin\theta\right)\cos\left[(\pi/2)\cos\theta\right] \qquad \text{(half-wave dipole)} \quad (11.2.8)$$

The antenna of Figure 11.2.2(b) can be considered to be a two-element antenna array (see Section 10.4.1) for which the two radiated phasors add in some directions and cancel in others, depending on the differential phase lag between the two rays. Antenna (b) has its peak gain at $\theta = \pi/2$, but its beamwidth is less than for (a) because rays from the two arms of the dipole are increasingly out of phase for propagation directions closer to the z axis, even more than for the half-wave dipole; thus the gain of (b) modestly exceeds that of (a). Whether one determines patterns numerically or by using the more intuitive phasor addition approach of Sections 10.4.1

and 10.4.5 is a matter of choice. Antenna (c) has very modest nulls for θ close to the ±z axis. The nulls are weak because the electric field due to $3.2I_o$ is only slightly reduced by the contributions from the phase-reversed segment carrying I_o.

Simple inspection of the current distribution for the antenna of Figure 11.2.2(d) and use of the methods of Section 10.4.1 reveal that its pattern has peaks in gain along the ±x and ±y axes, and a null along the ±z axes. Extending simple superposition and phase cancellation arguments to other angular directions makes it possible to guess the form of the complete antenna pattern $G(θ,ϕ)$, and therefore to check the accuracy of any integration using (11.2.6) for all antenna arms. Similar simple phase addition/cancellation analysis reveals that the more complicated antenna (e) has gain peaks along the ±x and ±y axes, and nulls along the ±z axes, although the polarization of each peak is somewhat different, as discussed in an example. Exact determination of pattern (e) is confused by the fact that these wires are sufficiently close to each other to interact, so the current distribution may be modified relative to the nominal TEM assumption sketched in the figure.

Example 11.2A

Determine the relative gains and polarizations along the ±x, ±y, and ±z axes for the antenna illustrated in Figure 11.2.2(e).

Solution: The two x-oriented wires do not radiate in the ±x direction. The four z-oriented wires emit radiation that cancels in that direction (one pair cancels the other), while the two y-oriented wires radiate in-phase y-polarized radiation in the ±x direction with relative total electric field strength $E_y = 2$. We assume each $λ/2$ segment radiates a relative electric field of unity. Similarly, the two y-oriented wires do not radiate in the ±y direction. The four z-oriented wires emit radiation that cancels in that direction (one pair cancels the other), while the two x-oriented wires radiate in-phase x-polarized radiation in the ±y direction with relative total electric field strength $E_x = 2$. The four z-oriented wires do not radiate in the ±z direction, and the two out-of-phase pairs of currents in the x and y directions also cancel in that direction, yielding a perfect null. Thus the gains are equal in the x and y directions (but with x polarization along the y axis, and y-polarization along the x axis), and the gain is zero on the z axis.

11.3   *Propagation of radio waves and thermal emission*

11.3.1   Multipath propagation

Electromagnetic waves can be absorbed, refracted, and scattered as they propagate through linear media. One result of this is that beams from the same transmitter can arrive at a receiver from multiple directions simultaneously with differing delays, strengths, polarizations, and Doppler shifts. These separate phasors add constructively or destructively to yield an enhanced or diminished total response that is generally frequency dependent. Since cellular telephones are

mobile and seldom have a completely unobstructed propagation path, they often exhibit strong fading and multipath effects.

Consider first the simple case where a single beam arrives via a direct path and a reflected beam with one-quarter the power of the first arrives along a reflected path that is 100λ longer. If the powers of these two beams are constant, then the total received power will fluctuate with frequency. If the voltage received for the direct beam is $\underline{V}$ and that of the second beam is $\underline{V}/2$ corresponding to quarter power, then when they are in phase the total received power is $1.5^2|\underline{V}|^2/2R$, where R is the circuit impedance. When they are 180° out of phase the power is $0.5^2|\underline{V}|^2/2R$, or one-ninth the maximum. This shift between maximum and minimum occurs each time the relative delay between the two paths changes by λ/2. Because the differential delay D is ~100λ, this represents a frequency change Δf of only one part in 200; Δf/f = λ/2D. Note that reflections can enhance or diminish the main signal, and clever antenna arrays can always compensate for the differential delays experienced from different directions so as to enhance the result.

Since cellular phones can have path differences of ~1 km at wavelengths of ~10 cm, their two-beam frequency maxima can be separated by as little as $10^{-4}f$, where f can be ~10^9 Hz. Fortunately this separation of 10^5 Hz is large compared to typical voice bandwidths. Alternatively, cellular phone signals can be coded to cover bandwidths large compared to fading bandwidths so the received signal strength is averaged over multiple frequency nulls and peaks and is therefore more stable.

Multipath also produces nulls in space if the rays arrive from different directions. For example, if two rays A and B of wavelength λ and arrive from angles separated by a small angle γ, then the distance D between intensity maxima and minima along a line roughly perpendicular to the direction of arrival will be ~λ/(2sinγ). The geometry is sketched in Figure 11.3.1. Three or more beams can be analyzed by similar phasor addition methods. Sometimes one of the beams is reflected from a moving surface, or the transmitter or receiver are moving, so these maxima and minima can vary rapidly with time.
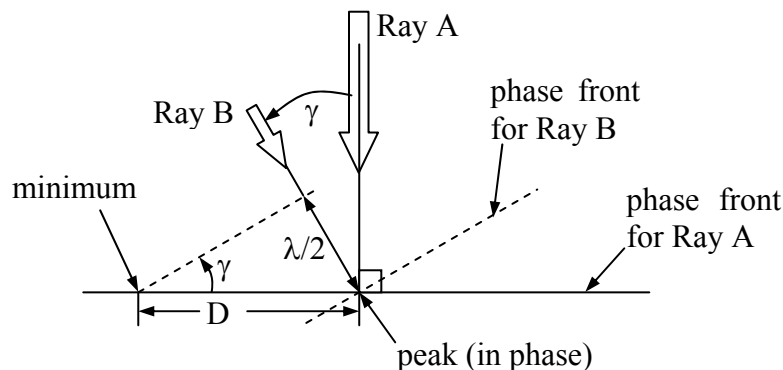
Figure 11.3.1   Maxima and minima created by multipath.

Example 11.3A

Normal broadcast NTSC television signals have 6-MHz bandwidth. If a metal building reflects perfectly a signal that travels a distance L further than the direct beam before the two equal-strength beams sum at the receiving antenna, how large can L be and still ensure that there are not two nulls in the 6-MHz passband between 100 and 106 MHz?

Solution: The differential path L is L/λ wavelengths long. If this number of wavelengths increases by one, then $L/\lambda' = L/\lambda + 1$ as λ decreases to λ'; this implies $\lambda/\lambda' = 1 + \lambda/L = f'/f = 1.06$. When the direct and reflected signals sum, the 2π phase change over this frequency band will produce one null, or almost two nulls if they fall at the band edges. Note that only the differential path length is important here. Therefore $L = \lambda/(1.06 - 1) \cong 16.7\lambda = 16.7c/f \cong 16.7\times3\times10^8/10^8 = 50.0$ meters.

11.3.2   Absorption, scattering, and diffraction

The terrestrial *atmosphere* can absorb, scatter, and refract electromagnetic radiation. The dominant gaseous absorbers at radio and microwave wavelengths are water vapor and oxygen. At submillimeter and infrared wavelengths, numerous trace gases such as ozone, NO, CO, OH, and others also become important. At wavelengths longer than 3 mm only the oxygen absorption band ~50 - 70 GHz is reasonably opaque. Horizontal attenuation at some frequencies 57-63 GHz exceeds 10 dB/km, and vertical attenuation can exceed 100 dB. The water vapor band 20-24 GHz absorbs less than 25 percent of radiation transmitted toward zenith or along a ~2-km horizontal path.

More important at low frequencies is the *ionosphere*, which reflects all radiation below its plasma frequency f_o, as discussed in Section 9.5.3. Radio waves transmitted vertically upward at frequency f are reflected directly back if any ionospheric layer has a plasma frequency $f_p <$ f, where f_p is given by (9.5.25) and is usually below 15 MHz. The ionosphere generally extends from ~70 to ~700 km altitude, with electron densities peaking ~300 km and exhibiting significant drops below ~200 km at night when solar radiation no longer ionizes the atmosphere fast enough to overcome recombination.

Above the plasma frequency f_p radio waves are also perfectly reflected at an angle of reflection θ_r equal to the angle of incidence θ_i if θ_i exceeds the critical angle $\theta_c(f)$ (9.2.30) for any ionospheric layer. The critical angle $\theta_c(f) = \sin^{-1}(\varepsilon_{ion}/\varepsilon_o)$, where the permittivity of the ionosphere $\varepsilon_{ion}(f) = \varepsilon_o[1 - (f_p/f)^2]$. Since $\varepsilon_o > \varepsilon_{ion}$ at any finite frequency, there exists a grazing angle of incidence θ_i where waves are perfectly reflected from the ionosphere at frequencies well above f_p. The curvature of the earth precludes grazing incidence with $\theta_i \rightarrow 90°$ unless the bottom surface of the ionosphere is substantially tilted. Therefore the maximum frequency at which radio waves can bounce around the world between the ionosphere and the surface of the earth is limited to ~$2f_p$, depending on the height of the ionosphere for the frequency of interest.

The most important non-gaseous atmospheric absorbers are clouds and rain, where the latter can attenuate signals 30 dB or more. Rain is a major absorber for centimeter-wavelength satellite dishes, partly in the atmosphere and partly as the rain accumulates on the antennas. At

longer wavelengths most systems have enough sensitivity to tolerate such attenuation. In comparison, clouds are usually not a problem except for through-the-air optical communication systems.

Atmospheric refraction is dominated by water vapor at radio wavelengths and by atmospheric density at optical wavelengths. These effects are not trivial. The radio sun can appear to set almost one solar diameter later on a very humid summer day (the sun emits strong radio waves too), and weak scattering from inhomogeneities in atmospheric humidity was once used as a major long-distance radio communications technique that avoided reliance on signals reflected from the ionosphere, as well as providing bandwidths of several GHz. Refraction by the ionosphere is even more extreme, and the angles of refraction can be computed using the properties of plasmas noted in Section 9.5.3 and Snell's law (9.2.26).

It is often convenient to model urban multipath and diffractive communications links by a power law other than r^{-2}. One common model is $r^{-3.8}$, which approximates the random weaking of signals by sequences of urban obstacles as signals 1.5-5 GHz propagate further. In any study of wireless communications systems propagation effects such as these must always be considered.

11.3.3   Thermal emission

A final effect impacting wireless communications systems is thermal noise arising from the environment, plus other forms of interference. Usually the thermal noise is considered interference too, but in radio astronomy and remote sensing it is the signal of interest. *Thermal noise* arises from electromagnetic radiation emitted by electrons colliding randomly with other particles in thermal equilibrium at temperature T. These collisions cause electrons to accelerate in random directions and therefore radiate. Thus every material object or medium radiates thermal noise provided that object or medium is coupled to the radiation field to any degree at all. Decoupled media perfectly reflect or transmit electromagnetic radiation without loss and are rare.

Thermal radiation propagating in a single-mode transmission line has intensity:

$$I[W/Hz] = \frac{hf}{e^{hf/kT} - 1} \cong kT \quad \text{for hf} \ll kT \quad [58] \qquad \text{(thermal intensity)} \quad (11.3.1)$$

Because there is a one-to-one relationship between intensity I and the corresponding brightness temperature T, the *brightness temperature* $T[K] = I/k$ often replaces I because of its more natural physical significance. T is the temperature of a matched load ($R = Z_o$) that would naturally radiate the same intensity $I = kT$ Watts/Hz for hf << kT. This *Rayleigh-Jeans approximation* for I is valid at temperatures T above 50K for all frequencies f below ~100 GHz.

Thus the Thevenin equivalent circuit of a resistor at temperature T includes a voltage source producing a generally observable gaussian white voltage $v_{Th}(t)$ called *Johnson noise*. This

---

[58] $e^{\delta} = 1 + \delta + \delta^2/2! + \ldots$ for $\delta \ll 1$.

source voltage $v_{Th}(t)$ radiates kTB [W] down a matched transmission line within the bandwidth B [Hz]. This Johnson noise voltage $v_{Th}(t)$ also divides across the Thevenin resistance R and its matched load $Z_o = R$ to produce the propagating line voltage $v_+(t, z=0) = v_{Th}/2$. But the radiated power is:

$$P_+ = \frac{\langle v_+^2 \rangle}{Z_o} = \frac{\langle (v_{Th}/2)^2 \rangle}{Z_o} = kTB \ [W] \qquad \text{(thermal noise power)} \qquad (11.3.2)$$

Therefore within bandwidth B the root-mean-square open-circuit thermal voltage v_{Thrms} across a resistor R at temperature T is:

$$v_{Thrms} = \sqrt{4kTBR} \ [V] \qquad \text{(Johnson noise)} \qquad (11.3.3)$$

A TEM line of impedance Z_o does not add any Johnson noise to that of the resistor if the line is lossless and therefore decoupled from the radiation.

Any antenna matched to its TEM transmission line therefore receives thermal noise power kT_AB [W] from the environment, where T_A is defined as the *antenna temperature*. T_A is the gain-weighted average of the brightness temperature T_B of the environment over 4π steradians:

$$T_A = \frac{1}{4\pi} \int_{4\pi} T_B(\theta,\phi) G(\theta,\phi) \ d\Omega \qquad \text{(antenna temperature)} \qquad (11.3.4)$$

If the entire field of view has brightness temperature $T_B = T_o$, and if the antenna is lossless so that $G(\theta,\phi) = D(\theta,\phi)$, then $T_A = T_o$ since $\int_{4\pi} D(\theta,\phi) d\Omega = 4\pi$ (10.3.3).

11.3.4 Radio astronomy and remote sensing

An antenna looking down at the earth sees a brightness temperature T_B, which is the sum of thermal radiation emitted by the earth plus downward propagating power that is then reflected from the same surface: $T_B = \xi T + |\Gamma|^2 T_B'$, where the emissivity of the earth $\xi = 1 - |\Gamma|^2$, Γ is the wave reflection coefficient of the earth, and T_B' is the brightness temperature of the radiation reflected from the earth into the antenna beam. The radiation from space at microwave frequencies has a brightness temperature near 2.7K arising from the "big bang" that occurred at the birth of the universe, and reaches temperatures over 7000K in the direction of the sun and certain astronomical objects, depending on frequency. The science of *radio astronomy* involves the study of such celestial radio waves.

The emissivity $1 - |\Gamma|^2$ of the terrestrial surface is typically 0.85-0.98 over land and $\gtrsim 0.3$ over ocean. Since most communications antennas point horizontally, about half their beam intercepts the earth (~260K) and half intercepts space (~4K at microwave frequencies), so the thermal noise from the environment typically adds ~132K to the antenna temperature and total system noise.

The study of natural radio, infrared, and visible emission from the earth is called *remote sensing*, although one can also remotely sense biological, manufacturing, and other systems. Today many satellites in polar and geostationary orbits routinely observe the earth at tens to thousands of wavelengths across the radio and optical spectrum for meteorological and other geophysical purposes. For example, a satellite observing in the opaque 53-67 GHz oxygen resonance band can not see much lower than 70 km altitude at the very centers of the strongest spectral lines, and therefore those channels observe the temperature of the atmosphere at those high altitudes. At nearby frequencies where the atmosphere is more transparent these sensors see the air temperatures at lower altitudes. Combinations of such observations yield the temperature profile of the atmosphere all over the globe, enabling better numerical weather predictions. Channels near the centers of water vapor, ozone, and other spectral lines can similarly measure their abundance and altitude profiles for similar purposes. Channels in the more transparent bands see closer to the terrestrial surface and permit estimates to be made of rain rate, surface winds, soil moisture, and other parameters.

Communications, radioastronomy, and remote sensing systems all receive non-thermal *radio interference* as well. Man-made interference comes from other transmitters in the same or nearby bands, automobiles, microwave ovens, motors, power supplies, corona around power lines, and other electrical devices. Each unshielded wire in any electrical device is a small antenna that radiates. For example, computers can emit highly structured signals that reveal the state of the computation and, in special cases, even the contents of registers. Poorly shielded power supplies often radiate at very high harmonics of their fundamental operating frequencies. Fortunately, regulations increasingly restrict radio emissions from modern electrical and electronic systems. Natural non-thermal emission arises from lightning, solar bursts, the planet Jupiter, and other sources.

11.4   Applications

11.4.1   Wireless communications systems

Section 11.4.1 introduces simple communications systems without using Maxwell's equations and Section 11.4.2 then discusses radar and lidar systems used for surveillance and research. Optical communications is deferred to Chapter 12, while the design, transformation, and switching of the communications signals themselves are issues left to other texts.

Wireless communications systems have a long history, beginning with wireless telegraph systems installed several years after Hertz's laboratory demonstrations of wireless links late in the nineteenth century. These systems typically used line-of-sight propagation paths, and sometimes inter-continental ionospheric reflections. Telephone, radio, and television systems followed. In the mid-twentieth century, the longer interstate and international wireless links were almost entirely replaced by more capable and reliable coaxial cables and multi-hop microwave links. These were soon supplemented by satellite links typically operating at frequencies up to ~14 GHz; today frequencies up to ~100 GHz are used. At century's end, these longer microwave links were then largely replaced again, this time by optical fibers with bandwidths of Terahertz. At the same time many of the shorter links are being replaced or supplemented by wireless cellular technology, which was made practical by the development of

inexpensive r.f. integrated circuits. Each technical advance markedly boosted capacity and market penetration, and generally increased performance and user mobility while reducing costs.

Most U.S. homes and offices are currently served by twisted pairs of telephone wires, each capable of conveying ~50 kbs - 1.5 Mbps, although coaxial cables, satellite links, and wireless services are making significant inroads. The most common wireless services currently include cell phones, wireless phones (within a home or office), wireless internet connections, wireless intra-home and intra-office connections, walkie-talkies (dedicated mobile links), satellite links, microwave tower links, and many specialized variations designed for private or military use. In addition, optical or microwave line-of-sight links between buildings offer instant broadband connectivity for the "last mile" to some users; the last mile accounts for a significant fraction of all installed plant cost. Weather generally restricts optical links to very short hops or to weather-independent optical fibers. Specialized wireless medical devices, such as RF links to video cameras inside swallowed pills, are also being developed.

Broadcast services now include AM radio near 1 MHz, FM radio near 100 MHz and higher frequencies, TV in several bands between 50 and 600 MHz for local over-the-air service, and TV and radio delivered by satellite at ~4, ~12, and ~20 GHz. Shortwave radio below ~30 MHz also offers global international broadcasts dependent upon ionospheric conditions, and is widely used by radio hams for long-distance communications.

Wireless services are so widespread today that we may take them for granted, forgetting that a few generations ago the very concept of communicating by invisible silent radio waves was considered magic. Despite the wide range of services already in use, it is reasonable to assume that over the next few decades numerous other wireless technologies and services will be developed by today's engineering students.

Communications systems convey information between two or more nodes, usually via wires, wireless means, or optical fibers. After a brief discussion relating signaling rates (bits per second) to the signal power required at the wireless receiver, this section discusses in general terms the launching, propagation, and reception of electromagnetic signals and messages in wired and wireless systems.

Information is typically measured in bits. One *bit of information* is the information content of a single yes-no decision, where each outcome is equally likely. A string of M binary digits (equiprobable 0's or 1's) conveys M bits of information. An analog signal measured with an accuracy of one part in 2^M also conveys M bits because a unique M-bit binary number corresponds to each discernable analog value. Thus both analog and digital signals can be characterized in terms of the bits of information they convey. All wireless receivers require that the energy received per bit exceed a rough minimum of $w_o \cong 10^{-20}$ Joules/bit, although most practical systems are orders of magnitude less sensitive.[59]

---

[59] Most good communications systems can operate with acceptable probabilities of error if $E_b/N_o >\sim 10$, where E_b is the energy per bit and $N_o = kT$ is the noise power density [W Hz^{-1}] = [J]. Boltzmann's constant $k \cong 1.38 \times 10^{-23}$ [J ^{o}K^{-1}], and T is the system noise temperature, which might approximate 100K in a good system at RF frequencies. Thus the nominal minimum energy E_b required to detect each bit of information is $\sim 10N_o \cong 10^{-20}$ [J].

To convey N bits per second [b s^{-1}] of information therefore requires that at least ~Nw_o watts [W] be intercepted by the receiver, and that substantially more power be transmitted. Note that [W] = [J s^{-1}] = [J b^{-1}][b s^{-1}]. Wireless communications is practical because so little power P_r is actually required at the receiver. For example, to communicate 100 megabits per second (Mbps) requires as little as one picowatt (10^{-12}W) at the receiver if $w_o = 10^{-20}$; that is, we require $P_r > Nw_o \cong 10^8 \times 10^{-20} = 10^{-12}$ [W].

It is fortunate that radio receivers are so sensitive, because only a tiny fraction of the transmitted power usually reaches them. In most cases the path loss between transmitter and receiver is primarily geometric; the radiation travels in straight lines away from the transmitting antenna with an intensity I [W m^{-2}] that grows weaker with distance r as r^{-2}. For example, if the transmitter is isotropic and radiates its power P_t equally in all 4π directions, then $I(\theta,\phi,r) = P_t/4\pi r^2$ [W m^{-2}]. The power P_r intercepted by the receiving antenna is proportional to the incident wave intensity $I(\theta,\phi)$ and the receiving *antenna effective area* $A(\theta,\phi)$ [m^2], or "capture cross-section", where the power P_r received from a plane wave incident from direction θ,ϕ is:

$$P_r = I(\theta,\phi,r) A(\theta,\phi) \ [W] \qquad\qquad \text{(antenna gain)} \quad (11.4.1)$$

The power received from an isotropic transmitting antenna is therefore $P_r = (P_t/4\pi r^2)A(\theta,\phi)$, so in this special case the line-of-sight path loss between transmitter and receiver is $P_r/P_t = A(\theta,\phi)/4\pi r^2$, or that fractional area of a sphere of radius r represented by the receiving antenna cross-section A. Sometimes additional propagation losses due to rain, gaseous absorption, or scattering must be recognized too, as discussed in Section 11.3.2.

In general, however, the transmitting antenna is not isotropic, but is designed to radiate power preferentially in the direction of the receivers. We define *antenna gain* $G(\theta,\phi)$, often called "gain over isotropic", as the ratio of the intensity $I(\theta,\phi,r)$ [W m^{-2}] of waves transmitted in the direction θ,ϕ (spherical coordinates) at distance r, to the intensity that would be transmitted by an isotropic antenna. That is:

$$G(\theta,\phi) \equiv \frac{I(\theta,\phi,r)}{P_t \big/ 4\pi r^2} \qquad\qquad \text{(antenna gain)} \quad (11.4.2)$$

If the radiated power is conserved, then the integral of wave intensity over a spherical surface enclosing the antenna is independent of the sphere's radius r. Therefore the angular distribution of power and $G(\theta,\phi)$ plotted in spherical coordinates behave much like a balloon that must push out somewhere when it is pushed inward somewhere else, as suggested in Figure 11.4.1. The maximum gain G_o often defines the z axis and is called the on-axis gain. The angular width θ_B of the main beam at the half-power points where $G(\theta,\phi) \cong G_o/2$ is called the *antenna beamwidth* or "half-power beamwidth". Other local peaks in gain are called *sidelobes,* and those sidelobes behind the antenna are often called backlobes. Angles at which the gain is nearly zero are called nulls.
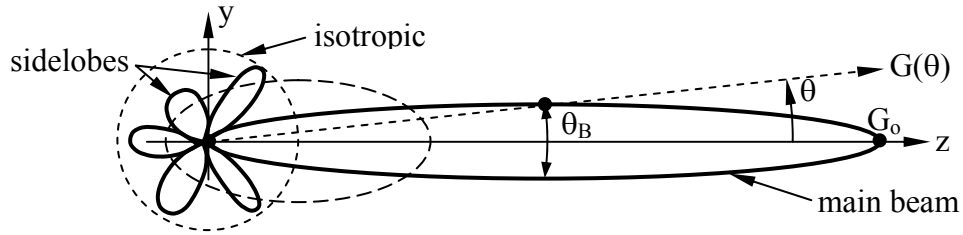
Figure 11.4.1 Isotropic and directive antenna gain patterns.

Antennas with $G(\theta,\phi) > 1$ generally focus their radiated energy by using lenses, mirrors, or multiple radiators phased so their radiated contributions add in phase in the desired direction, and largely cancel otherwise. Typical gains for most wire antennas range from ~1.5 to ~100, and large aperture antennas such as parabolic dishes or optical systems can have gains of 10^8 or more. The directionality or gain of a mirror or any antenna system is generally the same whether it is transmitting or receiving.[60] The fundamentals of transmission and reception are presented in more detail in Section 10.3.1.

Consider the following typical example. A television station transmits 100 kW at ~100 MHz toward the horizon with an antenna gain of ~10. Because the gain is much greater than unity in the desired horizontal direction, it is therefore less than unity for most other downward and upward directions where users are either nearby or absent. The intensity I [W m^{-2}] sensed by users on the horizon at 100-km range follows from (11.4.2):

$$I \cong G\frac{P_t}{4\pi r^2} = 10 \times \frac{10^5}{4\pi\left(10^5\right)^2} \cong 10^{-5} \left[W/m^2\right]$$ (11.4.3)

Whether this intensity is sufficient depends on the properties of the receiving antenna and receiver. For the example of Equation (11.4.3), a typical TV antenna with an effective area $A \cong 2$ [m^2] would capture $IA \cong 10^{-5}$[W/m^2] $\times$ 2[m^2] = 2×10^{-5} [W]. If the received power is $\langle v^2(t)\rangle/R \cong 2 \times 10^{-5}$ [W], and the receiver has an input impedance R of 100 ohms, then the root-mean-square (rms) voltage $v_{rms} \equiv \langle v^2(t)\rangle^{0.5}$ would be $(0.002)^{0.5} \cong 14$ mv, much larger than typical noise levels in TV receivers (~10 μv).[61]

---

[60] The degree of focus is the same whether the waves are transmitted or received. That is, if we reverse the direction of time for a valid electromagnetic wave solution to Maxwell's equations, the result is also a valid solution if the system is lossless and reciprocal. Reciprocity requires that the complex matrices characterizing $\underline{\varepsilon}$, $\underline{\mu}$, and $\underline{\sigma}$ near the antenna equal their own transposes; this excludes magnetized plasmas such as the ionosphere, and magnetized ferrites, as discussed further in Section 10.3.4.

[61] Typical TV receivers might have a superimposed noise voltage of power N = kTB [W], where the system noise temperature T might be ~10^4 K (much is interference), Boltzmann's constant k = 1.38×10^{-23}, and B is bandwidth [Hz]. B $\cong$ 6 MHz for over-the-air television. Therefore N $\cong 1.38 \times 10^{-23} \times 10^4 \times 6 \times 10^6 \cong 8 \times 10^{-13}$ watts, and a good TV signal-to-noise ratio S/N of ~10^4 requires only ~ 8×10^{-9} watts of signal S. Since N $\cong n_{rms}^2/R$, the rms noise voltage $\cong$ $(NR)^{0.5}$, or ~10 μv if the receiver input impedance R = 100 ohms.

Because most antennas are equally focused whether they are receiving or transmitting, their effective area $A(\theta,\phi)$ and gain $G(\theta,\phi)$ are closely related:

$$G(\theta,\phi) = \frac{4\pi}{\lambda^2} A(\theta,\phi)$$

(11.4.4)

Therefore the on-axis gain $G_o = 4\pi A_o/\lambda^2$. This relation (11.4.4) was proven for a short dipole antenna in Section 10.3.3 and proven for other types of antenna in Section 10.3.4, although the proof is not necessary here. This relation is often useful in estimating the peak gain of aperture antennas like parabolic mirrors or lenses because their peak effective area A_o often approaches their physical cross-section A_p within a factor of two; typically $A_o \cong 0.6\, A_p$. This approximation does not apply to wire antennas, however. Thus we can easily estimate the on-axis gain of such aperture antennas:

$$G_o = 0.6 \times \frac{4\pi}{\lambda^2} A_o$$

(11.4.5)

Combining (11.4.1) and (11.4.3) yields the *link expression* for received power:

$$P_r = G_t \frac{P_t}{4\pi r^2} A_r \, [\text{W}]$$

(link expression) (11.4.6)

where G_t is the gain of the transmitting antenna and A_r is the effective area of the receiving antenna. The data rate R associated with this received power is : $R = P_r/E_b$ [bits s^{-1}].

A second example illustrates how a communications system might work. Consider a *geosynchronous communications satellite*[62] transmitting 12-GHz high-definition television (HDTV) signals at 20 Mbps to homes with 1-meter dishes, and assume the satellite antenna spreads its power P_t roughly equally over the eastern United States, say 3×10^6 km^2. Then the intensity of the waves falling on the U.S. is: $I \cong P_t/(3 \times 10^{12})$ [W m^{-2}], and the power P_r received by an antenna with effective area $A_o \cong 0.6$ [m^2] is:

$$P_r = A_o I = 0.6 \frac{P_t}{3 \times 10^{12}} = 2 \times 10^{-13} \, P_t \, [\text{W}]$$

(11.4.7)

If $E_b = 10^{-20}$ Joules per bit suffices, then an R = 20-Mbps HDTV signal requires:

$$P_r = E_b R = 10^{-20} \times (2 \times 10^7) = 2 \times 10^{-13} \, [\text{W}]$$

(11.4.8)

The equality of the right-hand parts of (11.4.7) and (11.4.8) reveals that one watt of transmitter power P_t in this satellite could send a digital HDTV signal to all the homes and businesses in the

---

[62] A satellite approximately 35,000 km above the equator circles the earth in 24 hours at the same rate at which the earth rotates, and therefore can remain effectively stationary in the sky as a communications terminal serving continental areas. Such satellites are called "geostationary" or "geosynchronous".

eastern U.S. Since a 20-dB margin[63] for rain attenuation, noisy receivers, smaller or poorly pointed home antennas, etc. is desirable, 100-watt transmitters might be used in practice.

We can also estimate the physical area A_p of the aperture antenna on the satellite. If we know P_t and I at the earth, then we can determine the satellite gain G using $I = GP_t/4\pi r^2$ (11.4.2), where $r \cong 40,000$ km in the northern U.S; here we have $I \cong 3.3\times10^{-12}$ when $P_t = 1$ watt. The wavelength λ at 12 GHz is 2.5 cm ($\lambda = c/f$). But $A_p \cong 1.5A_o$, where A_o is related to G by (11.4.4). Therefore we obtain the reasonable result that a 2.5-meter diameter parabolic dish on the satellite should suffice:

$$A_p \cong 1.5A_o = \left(1.5\lambda^2/4\pi\right)G = \left(1.5\lambda^2/4\pi\right)(4\pi r^2 I/P_t) \cong 5\left[m^2\right] \qquad (11.4.9)$$

The same result could have been obtained by determining the angular extent of the U.S. coverage area as seen from the satellite and then, as discussed in Section 11.1.2, determining what diameter antenna would have a diffraction pattern with that same beamwidth.

Thus we can design digital communications systems for a *data rate* R [b s^{-1}] if we know the range r, wavelength λ, and receiver sensitivity (Joules required per bit). For analog systems we also need to know the desired signal-to-noise ratio (SNR) at the receiver and the noise power N. Table 11.4.1 lists typical data rates R for various applications, and Table 11.4.2 lists typical SNR values required for various types of analog signal.

Table 11.4.1   Digital data rates for typical applications and source coding techniques[64].

Applications	Data rate R after source coding	R before coding
Intelligible voice	>~1200 bps	~64 kbps
Good voice	>~4.8 – 9.6 kbps	~128 kbps
Excellent voice	>~16 kbps	~256 kbps
CD-quality music	2×128 kbps	~1.4 Mbps
Talking head, lip read	>~64 kbps	~1.4 Mbps
Good video conference	>~128-384 kbps	~12 Mbps
VHS video	>~1.5 Mbps	~30 Mbps
NTSC studio video	>~6 Mbps	~256 Mbps
HDTV video	>~18 Mbps	~1 Gbps

---

[63] Decibels (dB) are defined for a ratio R such that dB = $10 \log_{10}R$ and $R = 10^{(dB)/10}$; thus 20 dB $\rightarrow$ R = 100.

[64] Source coding reduces the number of bits to be communicated by removing redundancies and information not needed by the user. The table lists typical data rates before and after coding.

Table 11.4.2   Signal-to-noise ratios[65] for typical wireless applications.

Application	Desired SNR $\geq\sim$
Digital communications at ~ 1 bps/Hz	10 dB E_b/N_o
Digital communications at $>\sim$4 bps/Hz	20 dB E_b/N_o
Amplitude modulated (AM) signals (20 kHz typical)	30 dB S/N
Frequency modulated (FM) signals (100 kHz typical)	20 dB S/N
NTSC broadcast television (6 MHz typical)	35 dB S/N
CD-quality music (55-dB SNR + 40-dB dynamic range)	95 dB S/N

Example 11.4A

A parabolic reflector antenna of 2-meter diameter transmits $P_t = 10$ watts at 3 GHz from beyond the edge of the solar system ($R \cong 10^{10}$ km) to a similar antenna on earth of 50-m diameter at a maximum data rate N bits/sec. What is N if the receiver requires 10^{-20} Joules bit^{-1}?

Solution: Recall that the on-axis effective area A of a circular aperture antenna equals ~0.6 times its physical area (πr^2), and it has gain $G = 4\pi A/\lambda^2 = (2\pi r/\lambda)^2$.   The received power is $P_{rec} = P_t G_t A_r / 4\pi R^2$ (11.4.6); therefore:

$$R \cong P_{rec}/E_b = \left[P_t(0.6)^2 \left(2\pi r_t/\lambda\right)^2 \pi r_r^2 \right] / \left[4\pi R^2 E_b \right]$$

$$= \left[10 \times (0.6)^2 \left(2\pi \times 1/0.1\right)^2 \pi 25^2 \right] / \left[4\pi 10^{26} \times 10^{-20} \right] \cong 2.2 \text{ bps}$$

11.4.2   Radar and lidar

Radar (RAdio Direction and Range finding) and *lidar* (LIght Direction and Range finding) systems transmit signals toward targets of interest and receive echoes. They typically determine: 1) target distance using the round-trip propagation delay, 2) target direction using echo strength relative to antenna orientation, 3) target radial velocity using the observed Doppler shift, and 4) target size or scattering properties using the maximum echo strength. Figure 11.4.2 illustrates the most common radar configuration.
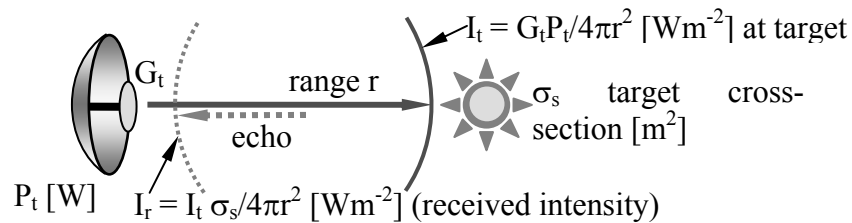
Figure 11.4.2   Radar signals reflected from a target.

---

[65] For digital signals the dimensionless signal-to-noise ratio (SNR) given here is the energy-per-bit E_b divided by the noise power density N_o [W Hz^{-1}], where $N_o = kT$ and T is the noise temperature, say 100-10^{4}K typically. For analog signals, S and N are the total signal and noise powers, respectively, where $N = kTB$ and B is signal bandwidth [Hz].

To compute the received power, we first compute the intensity I_t of radiation at the target at range r for a transmitter power and antenna gain of P_t and G, respectively:

$$I_t = G\frac{P_t}{4\pi r^2} \left[\text{W/m}^2\right] \qquad \text{(intensity at target)} \quad (11.4.10)$$

The target then scatters this radiation in some pattern and absorbs the rest. Some of this scattered radiation reaches the receiver with intensity I_r, where:

$$I_r = I_t\frac{\sigma_s}{4\pi r^2} \left[\text{W/m}^2\right] \qquad \text{(intensity at radar)} \quad (11.4.11)$$

where σ_s is the *scattering cross-section* of the target and is defined by (11.4.11). That is, σ_s is the capture cross-section [m^2] at the target that would produce I_r if the target scattered incident radiation isotropically. Thus targets that preferentially scatter radiation toward the transmitter can have scattering cross-sections substantially larger than their physical cross-sections.

The received power P_r is then simply $I_r A_r$ [W], where A_r is the effective area of the receiving antenna. That is:

$$P_r = I_r A_r = \frac{I_t \sigma_s}{4\pi r^2}A_r = G\frac{P_t \sigma_s}{\left(4\pi r^2\right)^2}A_r \qquad (11.4.12)$$

$$P_r = P_t\frac{\sigma_s}{4\pi}\left(\frac{G\lambda}{4\pi r^2}\right)^2 [\text{W}] \qquad \text{(radar equation)} \quad (11.4.13)$$

where we used $A_r = G\lambda^2/4\pi$, and where (11.4.13) is often called the *radar equation*. The dependence of received power on the fourth power of range and the square of antenna gain often control radar system design.

Atmospheric attenuation is often included in the radar equation by means of a round-trip attenuation factor $e^{-2\alpha r}$, where α is the average atmospheric attenuation coefficient (m^{-1}) and r is range. Atmospheric attenuation is discussed in Section 11.3.2 and below 200 GHz is due principally to oxygen, water vapor, and rain; it is usually not important below ~3 GHz. Oxygen absorption occurs primarily in the lowest 10 km of the atmosphere ~50-70 GHz and near 118 GHz, water vapor absorption occurs primarily in the lowest 3 km of the atmosphere above ~10 GHz, and rain absorption occurs up to ~15 km in the largest rain cells above ~3 GHz.

Lidar systems also obey the radar equation, but aerosol scattering by clouds, haze, or smoke becomes more of a concern. Also the phase fronts of optical beams are more easily disturbed by refractive inhomogeneities in the atmosphere that can modulate received echoes on time scales of milliseconds with random fading of ten dB or more.

A simple example illustrates use of the radar equation (11.4.13). Suppose we wish to know the range r at which we can detect dangerous asteroids having diameters over ~300m that are

approaching the earth. Assume the receiver has additive noise characterized by the system noise temperature T_s, and that the radar bandwidth is one Hertz because the received sinusoid will be averaged for approximately one second. Detectable radar echos must have $P_r > kT_sB$ [W], where k is Boltzmann's constant (k = 1.38×10^{-23}) and B is the system bandwidth (~1 Hz); this implies $P_r \cong 1.4 \times 10^{-23} T_s$ watts. We can estimate σ_s for a 300-meter asteroid by assuming it reflects roughly as well as the earth, say fifteen percent, and that the scattering is roughly isotropic; then $\sigma_s \cong 10^4$ [m^2]. If we further assume our radar is using near state-of-the-art components, then we might have $P_t \cong 1$ Mw, $G_t \cong 10^8$, $\lambda = 0.1$ m, and $T_s \cong 10$K. The radar equation then yields:

$$ r \cong \left[P_t\sigma_s(G_t\lambda)^2 \big/ (4\pi)^3 P_r \right]^{0.25} \cong 5 \times 10^7 \text{ km} \tag{11.4.14} $$

This range is about one-third of the distance to the sun and would provide about 2-3 weeks warning.

Optical systems with a large aperture area A might perform this task better because their antenna gain $G = A4\pi/\lambda^2$, and λ for lidar is typically 10^{-5} that of a common radar. For antennas of the same physical aperture and transmitter power, 1-micron lidar has an advantage over 10-cm radar of $\sim 10^{10}$ in P_r/P_t.

Radar suffers because of its dependence on the fourth power of range for targets smaller than the antenna beamwidth. If the radar can place all of its transmitted energy on target, then it suffers only the range-squared loss of the return path. The ability of lidar systems to strongly focus their transmitting beam totally onto a small target often enables their operation in the highly advantageous r^{-2} regime rather than r^{-4}.

Equations (11.4.13) and (11.4.14) can easily be revised for the case where all the radar energy intercepts the target. The radar equation then becomes:

$$ P_r = P_t RG(\lambda/4\pi r)^2 \text{ [W]} \tag{11.4.15} $$

where the target retro-reflectivity R is defined by (11.4.15) and is the dimensionless ratio of back-scattered radiation intensity at the radar to what would be back scattered if the radiation were scattered isotropically by the target. For the same assumptions used before, asteroids could be detected at a range r of $\sim 3 \times 10^{12}$ km if $R \cong 0.2$, a typical value for icy rock. The implied detection distance is now dramatically farther than before, and reaches outside our solar system. However, the requirement that the entire radar beam hit the asteroid would be essentially impossible even for the very best optical systems, so this approach to boosting detection range is usually not practical for probing small distant objects.

Radar systems often use phased arrays of antenna elements, as discussed in Section 10.4, to focus their energy on small spots or to look in more than one direction at once. In fact a single moving radar system, on an airplane for example, can coherently receive sequential reflected radar pulses and digitally reassemble the signal over some time period so as to synthesize the equivalent of a phased array antenna that is far larger than the physical antenna. That is, a small

receiving antenna can be moved over a much larger area A, and by combining its received signals from different locations in a phase-coherent way, can provide the superior angular resolution associated with area A. This is called *synthetic aperture radar* (SAR) and is not discussed further here.

Example 11.4B

A radar with 1-GHz bandwidth and 40-dB gain at 10 GHz views the sun, which has angular diameter 0.5 degrees and brightness temperature T_B = 10,000K. Roughly what is the antenna temperature T_A and the power received by the radar from the sun if we ignore any radar reflections?

Solution: The power received is the intensity at the antenna port I [W/Hz] times the bandwidth B [Hz], where $I \cong kT_A$ (11.3.1), and T_A is the antenna temperature given by the integral in (11.3.4). This integral is trivial if $G(\theta,\phi)$ is nearly constant over the solid angle Ω_S of the sun; then $T_A \cong G_oT_B\Omega_S/4\pi$. Constant gain across the sun requires the antenna beamwidth $\theta_B \gg 0.5$ degrees. We can roughly estimate θ_B by approximating the antenna gain as a constant G_o over solid angle Ω_B, and zero elsewhere; then (10.3.3) yields $\int_{4\pi} G(\theta,\phi)d\Omega = 4\pi = G_o\Omega_B = 10^4\Omega_B$. Therefore $\Omega_B \cong 4\pi/10^4 \cong \pi(\theta_B/2)^2$, and $\theta_B \cong 0.04$ radians $\cong 2.3$ degrees, which is marginally greater than the solar diameter required for use of the approximation $\theta_B \gg 0.5$ in a rough estimate. It follows that $T_A \cong G_oT_B\Omega_S/4\pi = 10^4 \times 10^4 \times \pi(\theta_S/2)^2/4\pi \cong$ 480 degrees Kelvin, somewhat larger than the noise temperature of good receivers. The power received is $kT_AB \cong 1.38\times10^{-23}\times480\times10^9 \cong 6.6\times10^{-12}$ watts. This is a slight overestimate because the gain is actually slightly less than G_o at the solar limb.

Example 11.4C

What is the scattering cross-section σ_s of a small distant flat plate of area F oriented so as to reflect incident radiation directly back toward the transmitter?

Solution: The radar equation (11.4.12) relates the transmitted power P_t to that received, P_r, in terms of σ_s. A similar relation can be derived by treating the power reflected from the flat plate as though it came from an aperture uniformly illuminated with intensity $P_tG_t/4\pi r^2$ [W m^{-2}]. The power P_r received by the radar is then the power radiated by the flat-plate aperture, $FP_tG_t/4\pi r^2$ [W], inserted as P_t into the link expression (11.4.6): $P_r = (FP_tG_t/4\pi r^2)G_fA_r/4\pi r^2$. The gain of the flat plate aperture is $G_f = F4\pi/\lambda^2$, and $A_r = G\lambda^2/4\pi$, so $G_fA_r = GF$. Equating P_r in this expression to that in the radar equation yields: $P_t\sigma_s(G\lambda/4\pi r^2)^2/4\pi = (FP_tG_t/4\pi r^2)GF/4\pi r^2$, so $\sigma_s = F(4\pi F/\lambda^2)$. Note $\sigma_s \gg F$ if $F \gg \lambda^2/4\pi$. *Corner reflectors* (three flat plates at right angles intersecting so as to form one corner of a cube) reflect plane waves directly back toward their source if the waves impact the concave portion of the reflector from any angle. Therefore the corner reflector becomes a very area-efficient radar target if its total projected area F is larger than λ^2.

Chapter 12: Optical Communications

12.1   Introduction to optical communication links

12.1.1   Introduction to optical communications and photonics

Optical communications is as ancient as signal fires and mirrors reflecting sunlight, but it is rapidly being modernized by *photonics* that integrate optics and electronics in single devices. Photonic systems are usually analyzed in terms of individual photons, although wave methods still characterize the guidance of waves through optical fibers, space, or other media. This chapter introduces optical communications and applications of photonics in Section 12.1. It then discusses simple optical waveguides in Section 12.2, lasers in Section 12.3, and representative components of optical communications systems in Sections 12.4, including photodetectors in 12.4.1-2, multiplexers in 12.4.3, interferometers in 12.4.4, and optical switches in 12.4.5.

12.1.2   Applications of photonics

Perhaps the single most important application of photonics today is to optical communications through low-loss glass fibers. Since 1980 this development has dramatically transformed global communications.   The advantage of an optical fiber for communications is that it has a bandwidth of approximately one terahertz, and can propagate signals over continental and even global distances when assisted by optical amplifiers. These amplifiers are currently separated more than ~80 km, and this separation is steadily increasing as technology improves. In contrast, coaxial cable, wire-pair, and wireless links at radio frequencies still dominate most communication paths of bandwidth < ~2 MHz, provided the length is less than ~1–50 km.

One broadband global wireless alternative to optics is microwave communications satellites in geosynchronous orbit[66] that can service ships at sea and provide moveable capacity addressing transient communications shortfalls or failures across the globe; the satellites simply point their antenna beams at the new users, who can be over 10,000 km apart. The greatest use of satellites, however, is for broadcast of entertainment over continental areas, either to end-users or to the head ends of cable distribution systems. In general, the limited terrestrial radio spectrum is more efficiently used for broadcast than for one-to-one communications unless there is re-use of spectrum as described in Section 10.4.6.   Optical techniques are disadvantaged for satellite-ground links or ground-to-ground links through air because of clouds and fog, which restrict such links to very short distances or to cases where spatial diversity[67] offers clear-air alternatives.

Optical links also have great potential for very broadband inter-satellite or diversity-protected satellite-earth communications because small telescopes easily provide highly focused antenna beams. For example, beamwidths of telescopes with 5-inch apertures are typically one

---

[66] Geosynchronous satellites at 22,753-mile altitude orbit Earth once every 24 hours and can therefore hover stationary in the sky if they are in an equatorial orbit.
[67] Spatial diversity involves use of spatially distinct communications links that suffer any losses independently; combining these signals in non-linear ways improves overall message reliability.

arc-second[68], corresponding to antenna gains of $\sim 4\pi \times (57 \times 3600)^2 \cong 5 \times 10^{11}$, approximately 5000 times greater than is achievable by all but the very best radio telescopes. Such optical links are discussed in Section 12.1.4.

Optical fibers are increasingly being used for much shorter links too, simply because their useable bandwidth can readily be expanded after installation and because they are cheaper for larger bandwidths. The distance between successive amplifiers can also be orders of magnitude greater (compare the fiber losses of Figure 12.2.6 with those of wires, as discussed in Section 7.1.4 and Section 8.3.1). The bandwidth per wire is generally less than ~0.1 GHz for distances between amplifiers of 1 km, whereas a single optical fiber can convey ~1 THz for 100 km or more. Extreme data rates are now also being conveyed optically between and within computers and even chips, although wires still have advantages of cost and simplicity for most ultra-short and high-power applications.

Optical communication is not the only application for photonics, however. Low-power lasers are used in everyday devices ranging from classroom pointers and carpenters' levels to bar-code readers, laser copiers and printers, surgical tools, medical and environmental instruments, and DVD players and recorders. Laser pulses lasting only 10^{-15} second (0.3 microns length) are used for biological and other research. High power lasers with tens of kilowatts of average power are used for cutting and other manufacturing purposes, and lasers that release their stored energy in sub-picosecond intervals can focus and compress their energy to achieve intensities of $\sim 10^{23}$ W/m^2 for research or, for example, to drive small thermonuclear reactions in compressed pellets. Moreover, new applications are constantly being developed with no end in sight.

12.1.3 Link equations

The link equations governing through-the-air optical communications are essentially the same as those governing radio, as described in Section 10.3. That is, the received power P_r is simply related to the transmitted power P_t by the gain and effective area of the transmitting and receiving antennas, G_t and A_e:

$$P_r = \left(G_t P_t / 4\pi r^2\right) A_e \ \ [W] \qquad \text{(optical link equation)} \quad (12.1.1)$$

The gain and effective area of single-mode optical antennas are related by the same equation governing radio waves, (10.3.23):

$$G = 4\pi A / \lambda^2 \qquad (12.1.2)$$

Some optical detectors intercept multiple independent waves or modes, and their powers add. In this case, the gain and effective area of any single mode are then less relevant, as discussed in Section 12.1.4.

---

[68] One arc-second is 1/60 arc-minutes, $1/60^2$ degrees, $1/(57.3 \times 3600)$ radians, or 1/60 of the largest apparent diameters of Venus or Jupiter in the night sky.

The maximum bit rate that can be communicated over an optical link is not governed by the $E_b > \sim 10^{-20}$ Joules-per-bit limit characteristic of radio systems, however, but rather by the number of photons the receiver requires per bit of information, perhaps ~10 for a typical good system. Each photon has energy E = hf Joules. Thus to receive R bits/second might require received power of:

$$P_r = E_b R \cong 10hfR \ \ [W] \qquad \qquad \text{(optical rate approximation)} \quad (12.1.3)$$

where h is Planck's constant (6.624×10^{-34}) and f is photon frequency [Hz]. Clever design can enable many bits to be communicated per photon, as discussed in the following section.

12.1.4   Examples of optical communications systems

Three examples illustrate several of the issues inherent in optical communications systems: a trans-oceanic optical fiber cable, an optical link to Mars, and an incoherent intra-office link carrying computer information.

First consider a trans-oceanic *optical fiber*. Section 12.2.2 discusses losses in optical fibers, which can be as low as ~0.2 dB/km near 1.5-micron wavelength ($f \cong 2 \times 10^{14}$ Hz). To ensure the signal (zeros and ones) remains unambiguous, each link of an R = 1-Gbps fiber link must deliver to its receiver or amplifier more than ~10hfR watts, or ~$10 \times 6 \times 10^{-34} \times 2 \times 10^{14} \times 10^9 \cong 1.2 \times 10^{-9}$ watts; a more typical design might deliver ~10^{-6} watts because errors accumulate and equipment can degrade. If one watt is transmitted and 10^{-6} watts is received, then the associated 60-dB loss corresponds to 300 km of fiber propagation between optical amplifiers, and perhaps ~20 amplifiers across the Atlantic Ocean per fiber. In practice, erbium-doped fiber amplifiers, discussed in Section 12.3.1, are now spaced approximately 80 km apart.

Next consider an *optical link* communicating between Earth and astronauts on Mars. Atmospheric diffraction or "seeing" limits the focusing ability of terrestrial telescopes larger than ~10 cm, but Mars has little atmosphere. Therefore a Martian optical link might employ the equivalent of a one-square-meter optical telescope on Mars and the equivalent of 10-cm-square optics on Earth. It might also employ a one-watt laser transmitter on Earth operating at 0.5-micron wavelength, in the visible region. The nominal link and rate equations, (12.1.1) and (12.1.3), yield the maximum data rate R possible at a range of ~10^{11} meters (approximate closest approach of Mars to Earth):

$$R = P_r/E_b \cong \left(G_t P_t/4\pi r^2\right)A_e\big/10hf \ \ [\text{bits s}^{-1}] \qquad \qquad (12.1.4)$$

The gain G_t of the transmitter given by (12.1.2) is $G_t \cong 4\pi A/\lambda^2 \cong 5 \times 10^{11}$, where $A \cong (0.1)^2$ and $\lambda \cong 5 \times 10^{-7}$ [m]. The frequency $f = c/\lambda = 3 \times 10^8/[5 \times 10^{-7}] = 6 \times 10^{14}$. Therefore (12.1.4) becomes:

$$R \cong \left\{[5 \times 10^{11} \times 1]\big/\left[4\pi(10^{11})^2\right]\right\} \left\{1\big/[10 \times 6.624 \times 10^{-34} \times 6 \times 10^{14}]\right\} \cong 1 \text{ Mbps} \quad (12.1.5)$$

Table 11.4.1 suggests that this data rate is adequate for full-motion video of modest quality. The delay of the signal each way is $\tau = r/c = 10^{11}/[3\times10^8]$ seconds $\cong 5.6$ minutes, impeding conversation. This delay becomes several times greater when Mars is on the far side of the sun from Earth, and the data rate R would then drop by more than a factor of ten.

This 1-Mbps result (12.1.5) assumed 10 photons were required per bit of information. However this can be reduced below one photon per bit by using *pulse-position modulation*. Suppose $\sim10^6$ 1-nsec 10-photon pulses were received per second, where each pulse could arrive in any of 1024 time slots because the ratio of pulse width to average inter-pulse spacing is 1024. This timing information conveys ten bits of information per pulse because $\log_2 1024 = 10$. Since each 10-photon pulse conveys 10 bits of information, the average is one bit per photon received. With more time slots still fewer photons per bit would be required. If a tunable laser can transmit each pulse at any of 1024 colors, for example, then another factor of 10 can be achieved. Use of both pulse position and pulse-frequency modulation can permit more than 10 bits to be communicated per photon on average.

The final example is that of a 1-mW laser diode transmitting digitally modulated light at $\lambda = 5\times10^{-7}$ [m] isotropically within a large office over ranges r up to 10 meters, where the light might travel directly to the isotropic receiver or bounce off walls and the ceiling first. Such optical communications systems might link computers, printers, personal digital assistants (pda's), and other devices within the room. In this case $G_t = 1$ and $A_e = G\lambda^2/4\pi = (5\times10^{-7})^2/4\pi \cong 2\times10^{-14}$ [m^2]. The maximum data rate R can again be found using (12.1.4):

$$R = P_r/E_b \cong \left(1\times10^{-3}/4\pi10^2\right)\left(2\times10^{-14}\right)/\left(10\times6.6\times10^{-34}\times6\times10^{14}\right) \cong 0.004 \;\left[\text{bits s}^{-1}\right] \quad (12.1.6)$$

The fact that we can send 10^6 bits per second to Mars with a one-watt transmitter, but only 4 millibits per second across a room with a milliwatt, may conflict with intuition.

The resolution of this seeming paradox lies in the assumption that the receiver in this example is a single mode device like that of typical radio receivers or the Martian optical receiver considered above. If this room-link receiver were isotropic and intercepted only a single mode, its effective area A_e given by (12.1.2) would be 2×10^{-14} [m^2]. The tiny effective area of such low-gain coherent optical antennas motivates use of incoherent photodetectors instead, which respond well to the total photon flux from all directions of arrival. For example, intra-room optical links of this type are commonly used for remote control of many consumer electronic devices, but with a much larger multimode antenna (photodiode) of area $A \cong 2\times10^{-6}$ [m^2] instead of 2×10^{-14}. This "antenna" is typically responsive to all photons impacting its area that arrive within roughly one steradian. That is, a photodetector generally intercepts all photons impacting it, even though those photons are incoherent with each other. Thus the solution (12.1.6) is increased by a factor of $10^{-6}/10^{-14}$ if a two-square-millimeter photodetector replaces the single-mode antenna, and R then becomes 0.4 Mbps, which is more capacity than normally required. In practice such inexpensive area detectors are noisier and require orders of mangitude more photons per bit. Better semiconductor detectors can achieve 10 photons per bit or better, however, particularly at visible wavelengths and if stray light at other wavelengths is filtered out.

12.2   Optical waveguides

12.2.1   Dielectric slab waveguides

Optical waveguides such as optical fibers typically trap and guide light within rectangular or cylindrical boundaries over useful distances. Rectangular shapes are easier to implement on integrated circuits, while cylindrical shapes are used for longer distances, up to 100 km or more. Exact wave solutions for such structures are beyond the scope of this text, but the same basic principles are evident in dielectric slab waveguides for which the derivations are simpler. *Dielectric slab waveguides* consist of an infinite flat dielectric slab of thickness 2d and permittivity ε imbedded in an infinite medium of lower permittivity ε_o, as suggested in Figure 12.2.1(a) for a slab of finite width in the y direction. For simplicity we here assume $\mu = \mu_o$ everywhere, which is usually the case in practice too.
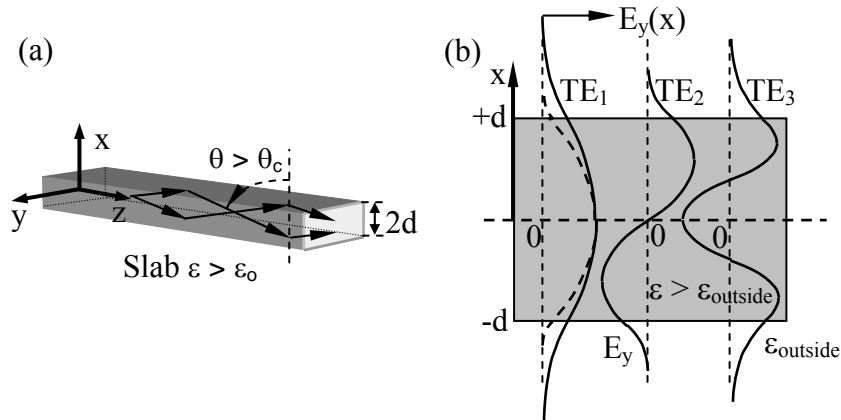
Figure 12.2.1   Dielectric slab waveguide and TE mode structure.

As discussed in Section 9.2.3, uniform plane waves within the dielectric are perfectly reflected at the slab boundary if they are incident beyond the critical angle $\theta_c = \sin^{-1}(c_\varepsilon/c_o)$, where c_ε and c_o are the velocities of light in the dielectric and outside, respectively.   Such a wave and its perfect reflection propagate together along the z axis and form a standing wave in the orthogonal x direction.   Outside the waveguide the waves are evanescent and decay exponentially away from the guide, as illustrated in Figure 12.2.2. This figure portrays the fields inside and outside the lower half of a dielectric slab having $\varepsilon > \varepsilon_o$; the lower boundary is at x = 0. The figure suggests two possible positions for the upper slab boundary that satisfy the boundary conditions for the TE_1 and TE_2 modes. Note that the TE_1 mode waveguide can be arbitrarily thin relative to λ and still satisfy the boundary conditions. The field configurations above the upper boundary mirror the fields below the lower boundary, but are not illustrated here.   These waveguide modes are designated TE_n because the electric field is only transverse to the direction of propagation, and there is part of n half-wavelengths within the slab. The orthogonal modes (not illustrated) are designated TM_n.
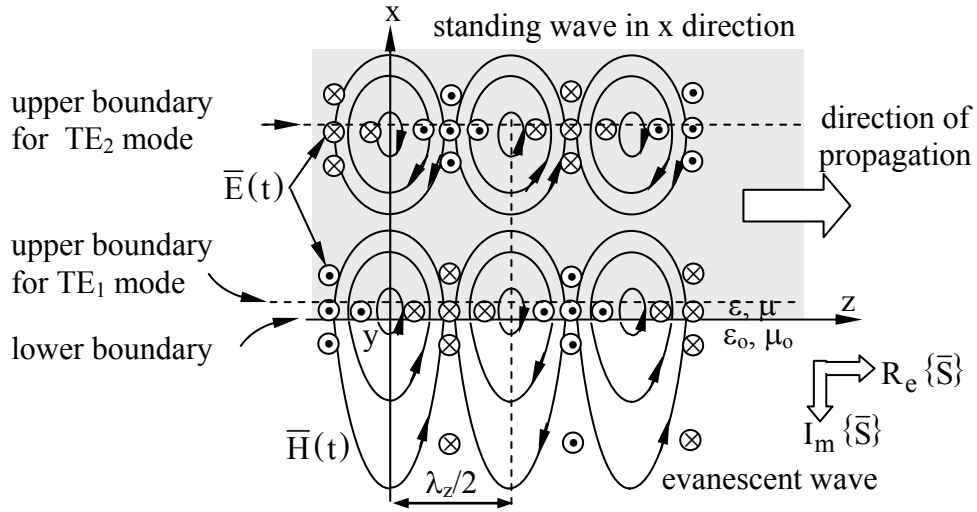
Figure 12.2.2 Fields in dielectric slab waveguides for TE_n modes.

The fields inside a dielectric slab waveguide have the same form as (9.3.6) and (9.3.7) inside parallel-plate waveguides, although the boundary positions are different; also see Figures 9.3.1 and 9.3.3. If we define $x = 0$ at the axis of symmetry, and the thickness of the guide to be 2d, then within the guide the electric field for TE modes is:

$$\overline{E} = \hat{y}\underline{E}_o \{\sin k_x x \ or \ \cos k_x x\} e^{-jk_z z} \qquad \qquad \text{for } |x| \le d \qquad (12.2.1)$$

The fields outside are the same as for TE waves incident upon dielectric interfaces beyond the critical angle, (9.2.33) and (9.2.34):

$$\overline{E} = \hat{y}\underline{E}_1 e^{-\alpha x - jk_z z} \qquad \qquad \text{for } x \ge d \qquad (12.2.2)$$

$$\overline{E} = \{\text{- } or \text{ +}\}\hat{y}\underline{E}_1 \ ^{+\alpha x - jk_z z} \qquad \qquad \text{for } x \le -d \qquad (12.2.3)$$

The first and second options in braces correspond to anti-symmetric and symmetric TE modes, respectively. Since the waves decay away from the slab, α is positive. Faraday's law in combination with (12.2.1), (12.2.2), and (12.2.3) yields the corresponding magnetic field inside and outside the slab:

$$\overline{H} = \left[\hat{x} \ k_z \{\sin k_x x \ or \ \cos k_x x\} \right.$$
$$\left. + \hat{z} \ jk_x \{\cos k_x x \ or \ \sin k_x x\} \right] (\underline{E}_o / \omega\mu_o) e^{-jk_z z} \qquad \text{for } |x| \le d \qquad (12.2.4)$$

$$\overline{H} = -(\hat{x} k_z + \hat{z} j\alpha)(\underline{E}_1 / \omega\mu_o) e^{-\alpha x - jk_z z} \qquad \qquad \text{for } x \ge d \qquad (12.2.5)$$

$$\overline{\underline{H}} = \{+ \text{ } or \text{ } -\}(\hat{x}k_z - \hat{z}j\alpha)(\underline{E}_1/\omega\mu_o)e^{\alpha x - jk_z z} \qquad \text{for } x \le -d \qquad (12.2.6)$$

The TE$_1$ mode has the interesting property that it approaches TEM behavior as $\omega \to 0$ and the decay length approaches infinity; most of the energy is then propagating outside the slab even though the mode is guided by it. Modes with $n \ge 2$ have non-zero cut-off frequencies. There is no TM mode that propagates for f→0 in dielectric slab waveguides, however.

Although Figure 12.2.1(a) portrays a slab with an insulating medium outside, the first option in brackets {•} for the field solutions above is also consistent for $x > 0$ with a slab located $0 < x < d$ and having a perfectly conducting wall at $x = 0$; all boundary conditions are matched; these are the anti-symmetric TE modes. This configuration corresponds, for example, to certain optical guiding structures overlaid on conductive semiconductors.

To complete the TE field solutions above we need additional relations between $\underline{E}_o$ and $\underline{E}_1$, and between k_x and α. Matching $\overline{\underline{E}}$ at $x = d$ for the symmetric solution [cos $k_x x$ in (12.2.1)] yields:

$$\hat{y}\underline{E}_o \cos(k_x d)e^{-jk_z z} = \hat{y}\underline{E}_1 e^{-\alpha d - jk_z z} \qquad (12.2.7)$$

Matching the parallel ($\hat{z}$) component of $\overline{\underline{H}}$ at $x = d$ yields:

$$-\hat{z}jk_x \sin(k_x d)(\underline{E}_o/\omega\mu_o)e^{-jk_z z} = -\hat{z}j\alpha(\underline{E}_1/\omega\mu_o)e^{-\alpha d - jk_z z} \qquad (12.2.8)$$

The *guidance condition* for the symmetric TE dielectric slab waveguide modes is given by the ratio of (12.2.8) to (12.2.7):

$$k_x d \tan(k_x d) = \alpha d \qquad \text{(slab guidance condition)} \qquad (12.2.9)$$

Combining the following two dispersion relations and eliminating k_z can provide the needed additional relation (12.2.12) between k_x and α:

$$k_z^2 + k_x^2 = \omega^2 \mu_o \varepsilon \qquad \text{(dispersion relation inside)} \qquad (12.2.10)$$

$$k_z^2 - \alpha^2 = \omega^2 \mu_o \varepsilon_o \qquad \text{(dispersion relation outside)} \qquad (12.2.11)$$

$$k_x^2 + \alpha^2 = \omega^2(\mu_o \varepsilon - \mu_o \varepsilon_o) > 0 \qquad \text{(slab dispersion relation)} \qquad (12.2.12)$$

By substituting into the guidance condition (12.2.9) the expression for α that follows from the slab dispersion relation (12.2.12) we obtain a transcendental guidance equation that can be solved numerically or graphically:

$$\tan k_x d = \left(\left[\omega^2\mu_o(\varepsilon-\varepsilon_o)d^2\big/k_x^2 d^2\right]-1\right)^{0.5} \qquad \text{(guidance equation)} \quad (12.2.13)$$

Figure 12.2.3 plots the left- and right-hand sides of (12.2.13) separately, so the modal solutions are those values of $k_x d$ for which the two families of curves intersect.
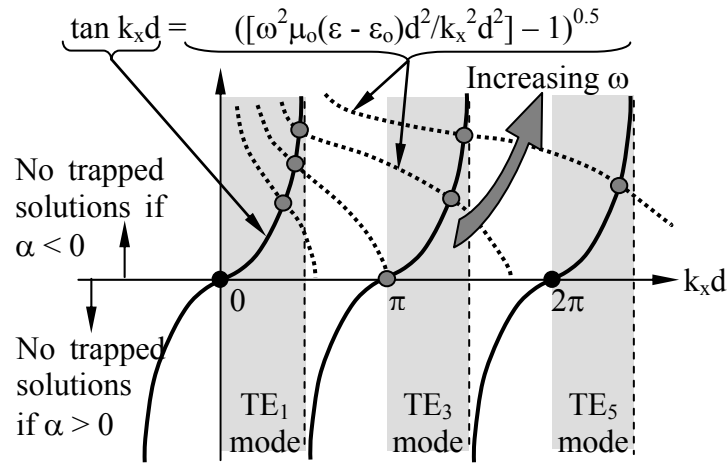
Figure 12.2.3   TE modes for a dielectric slab waveguide.

Note that the TE_1 mode can be trapped and propagate at all frequencies, from nearly zero to infinity. At low frequencies the waves guided by the slab have small values of α and decay very slowly away from the slab so that most of the energy is actually propagating in the z direction outside the slab rather than inside.   The value of α can be found from (12.2.12), and it approaches zero as both $k_x d$ and ω approach zero.

The TE_3 mode cannot propagate near zero frequency however. Its cutoff frequency ω_{TE3} occurs when $k_x d = \pi$, as suggested by Figure 12.2.3; ω_{TE3} can be determined by solving (12.2.12) for this case. This and all higher modes cannot be trapped at low frequencies because then the plane waves that comprise them impact the slab wall at angles beyond θ_c that permit escape. As ω increases, more modes can propagate. Figures 12.2.2 and 12.2.1(b) illustrate symmetric TE_1 and TE_3 modes, and the antisymmetric TE_2 mode. Similar figures could be constructed for TM modes.

These solutions for dielectric-slab waveguides are similar to the solutions for optical fibers, which instead take the form of Bessel functions because of their cylindrical geometry. In both cases we have lateral standing waves propagating inside and evanescent waves propagating outside.

12.2.2   Optical fibers

An *optical fiber* is generally a very long solid glass wire that traps lightwaves inside as do the dielectric slab waveguides described in Section 12.2.1. Fiber lengths can be tens of kilometers or

more. Because the fiber geometry is cylindrical, the electric and magnetic fields inside and outside the fiber are characterized by *Bessel functions*, which we do not address here. These propagating electromagnetic fields exhibit lateral standing waves inside the fiber and evanescence outside. To minimize loss the fiber core is usually overlaid with a low-permittivity glass cladding so that the evanescent decay also occurs within low-loss glass.

A typical glass optical fiber transmission line is perhaps 125 microns in diameter with a high-permittivity glass core having diameter ~6 microns. The core permittivity $\varepsilon + \Delta\varepsilon$ is typically ~2 percent greater than that of the cladding (ε). If the lightwaves within the core impact the cladding beyond the critical angle θ_c, where:

$$\theta_c = \sin^{-1}\left(\varepsilon/(\varepsilon + \Delta\varepsilon)\right)$$
(12.2.14)

then these waves are perfectly reflected and trapped. The evanescent waves inside the cladding decay approximately exponentially away from the core to negligible values at the outer cladding boundary, which is often encased in plastic about 0.1 mm thick that may be reinforced. Graded-index fibers have a graded transition in permittivity between the core and cladding. Some fibers propagate multiple modes that travel at different velocities so as to interfere at the output and limit information extraction (data rate). Multiple fibers are usually bundled inside a single cable. Figure 12.2.4 suggests the structure of a typical fiber.
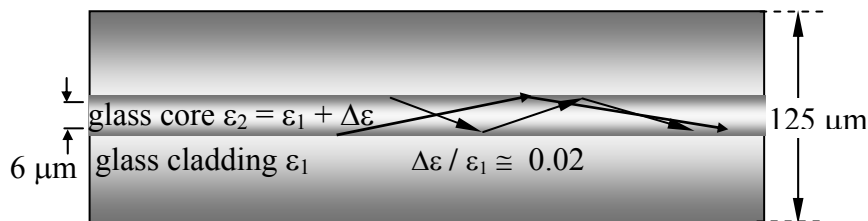
Figure 12.2.4   Typical clad optical fiber.

Figure 12.2.5 shows four common forms of optical fiber; many others exist. The multimode fiber is thicker and propagates several modes, while the single-mode fiber is so thin that only one mode can propagate. The diameter of the core determines the number of propagating modes. In all cylindrical structures, even single-mode fibers, both vertically and horizontally polarized waves can propagate independently and therefore may interfere with each other when detected at the output. If a single-mode fiber has an elliptical cross-section, one polarization can be made to escape so the signal becomes pure. That is, one polarization decays more slowly away from the core so that it sees more of the absorbing material that surrounds the cladding.
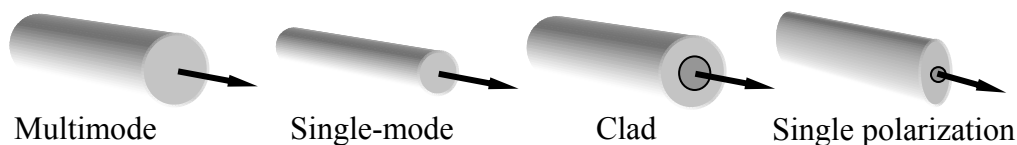
Figure 12.2.5   Types of optical fiber.

The initial issue faced in the 1970's by designers of optical fibers was propagation loss. Most serious was absorption due to residual levels of impurities in the glass, so much research and development involved purification. Water posed a particularly difficult problem because one of its harmonics fell in the region where attenuation in glass was otherwise minimum, as suggested in Figure 12.2.6.
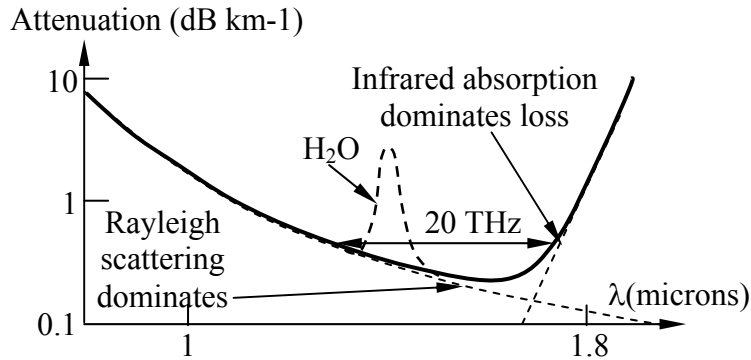
Figure 12.2.6 Loss mechanisms in optical fibers.

At wavelengths shorter than ~1.5 microns the losses are dominated by Rayleigh scattering of the waves from the random fluctuations in glass density on atomic scales. These scattered waves exit the fiber at angles less than the critical angle. *Rayleigh scattering* is proportional to f^4 and occurs when the inhomogeneities in ε are small compared to $\lambda/2\pi$. Inhomogeneities in glass fibers have near-atomic scales, say 1 nm, whereas the wavelength is more than 1000 times larger. Rayleigh scattering losses are reduced by minimizing unnecessary inhomogeneities through glass purification and careful mixing, and by decreasing the critical angle. Losses due to scattering by rough fiber walls are small because drawn glass fibers can be very smooth and little energy impacts the walls.

At wavelengths longer than ~1.5 microns the wings of *infrared absorption* lines at lower frequencies begin to dominate. This absorption is due principally to the vibration spectra of inter-atomic bonds, and is unavoidable. The resulting low-attenuation band centered near 1.5-microns between the Rayleigh and IR attenuating regions is about 20 THz wide, sufficient for a single fiber to provide each person in the U.S.A. with a bandwidth of $20\times10^{12}/2.5\times10^8 = 80$ kHz, or 15 private telephone channels! Most fibers used for local distribution do not operate anywhere close to this limit for lack of demand, although some undersea cables are pushing toward it.

The fibers are usually manufactured first as a preform, which is a glass rod that subsequently can be heated at one end and drawn into a fiber of the desired thickness. Preforms are either solid or hollow. The solid ones are usually made by vapor deposition of SiO_2 and GeO_2 on the outer surface of the initial core rod, which might be a millimeter thick. By varying the mixture of gases, usually $Si(Ge)Cl_4 + O_2 \Rightarrow Si(Ge)O_2 + 2Cl_2$, the permittivity of the deposited glass cladding can be reduced about 2 percent below that of the core. The boundary between core and cladding can be sharp or graded in a controlled way. Alternatively, the preform cladding is large and hollow, and the core is deposited from the inside by hot gases in

the same way; upon completion there is still a hole through the middle of the fiber. Since the core is small compared to the cladding, the preforms can be made more rapidly this way. When the preform is drawn into a fiber, any hollow core vanishes. Sometimes a hollow core is an advantage. For example, some newer types of fibers have cores with laterally-periodic lossless longitudinal hollows within which much of the energy can propagate.

Another major design issue involves the *fiber dispersion* associated with frequency-dependent phase and group velocities, where the *phase velocity* $v_p = \omega/k$. If the *group velocity* v_g, which is the velocity of the envelope of a narrowband sinusoid, varies over the optical bandwidth, then the signal waveform will increasingly distort as it propagates because the faster moving frequency components of the envelope will arrive early. For example, a digital pulse of light that lasts T seconds is produced by multiplying a *boxcar modulation* envelope (the T-second pulse shape) by the sinusoidal optical carrier, so the frequency spectrum is the convolution of the spectrum for the sinusoid (a spectral impulse) and the spectrum for a boxcar pulse ($\propto [\sin(2\pi t/T)]/[2\pi t/T]$). The outermost frequencies suffer from dispersion the most, and these are primarily associated with the sharp edges of the pulse.

The group velocity v_g derived in (9.5.20) is the slope of the dispersion relation at the optical frequency of interest:

$$v_g = \left(\partial k / \partial \omega\right)^{-1} \qquad\qquad\qquad (12.2.15)$$

Figure 12.2.7 illustrates the dispersion relation for three different modes; the higher order modes propagate information more slowly.
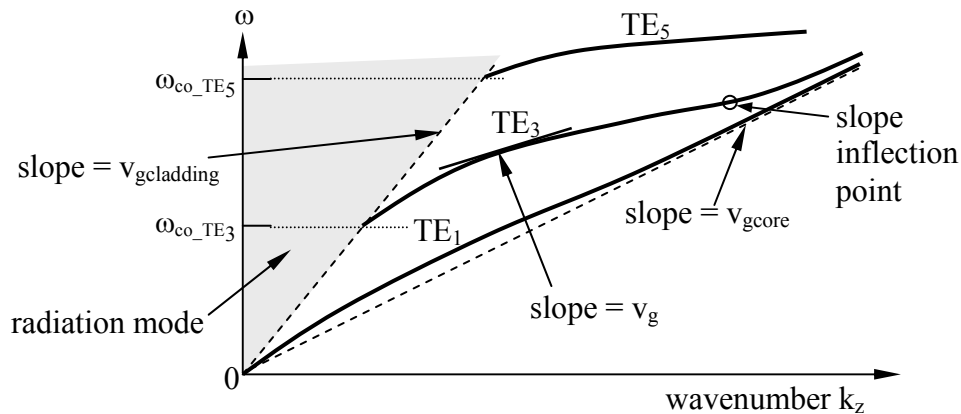
Figure 12.2.7   Group velocities for optical fiber modes.

The group velocity v_g is the slope of the $\omega(k)$ relation and is bounded by the slopes associated with the core (v_{gcore}) and with the cladding ($v_{gcladding}$), where the cladding is assumed to be infinite. The figure has greatly exaggerated the difference in the slope between the core and cladding for illustrative purposes.

A dispersive line eventually transforms a square optical pulse into a long "frequency chirped" pulse with the faster propagating frequencies in the front and the slower propagating frequencies in the back. This problem can be minimized by carefully choosing combinations of: 1) the dispersion $n(f)$ of the glass, 2) the permittivity contour $\varepsilon(r)$ in the fiber, and 3) the optical center frequency f_o. Otherwise we must reduce either the bandwidth of the signal or the length of the fiber. To increase the distance between amplifiers the dispersion can be compensated periodically by special fibers or other elements with opposite dispersion.

Pulses spread as they propagate over distance L because their outermost frequency components ω_1 and $\omega_2 = \omega_1 + \Delta\omega$ have arrival times at the output separated by:

$$\Delta t = L/v_{g1} - L/v_{g2} = L\left[d\left(v_g^{-1}\right)/d\omega\right]\Delta\omega = L\left(d^2k/d\omega^2\right)\Delta\omega \qquad (12.2.16)$$

where v_{gi} is the group velocity at ω_i (12.2.15). Typical pulses of duration T_p have a bandwidth $\Delta\omega \cong T_p^{-1}$, so brief pulses spread faster. The spread Δt is least at frequencies where $d^2k/d\omega^2 \cong 0$, which is near the representative slope inflection point illustrated in Figure 12.2.7.

This natural fiber dispersion can, however, help solve the problem of fiber nonlinearity. Since attenuation is always present in the fibers, the amplifiers operate at high powers, limited partly by their own nonlinearities and those in the fiber that arise because ε depends very slightly on the field strength E. The effects of non-linearities are more severe when the signals are in the form of isolated high-energy pulses. Deliberately dispersing and spreading the isolated pulses before amplifying and introducing them to the fiber reduces their peak amplitudes and the resulting nonlinear effects. This pre-dispersion is made opposite to that of the fiber so that the fiber dispersion gradually compensates for the pre-dispersion over the full length of the fiber. That is, if the fiber propagates high frequencies faster, then those high frequency components are delayed correspondingly before being introduced to the fiber. When the pulses reappear in their original sharp form at the far end of the fiber their peak amplitudes are so weak from natural attenuation that they no longer drive the fiber nonlinear.

Example 12.2A
If 10-ps pulses are used to transmit data at 20 Gbps, they would be spaced 5×10^{-10} sec apart and would therefore begin to interfere with each other after propagating a distance L_{max} sufficient to spread those pulses to widths of 50 ps. A standard single-mode optical fiber has dispersion $d^2k/d\omega^2$ of 20 ps^2/km at 1.5 μm wavelength. At what distance L_{max} will such 10-ps pulses have broadened to 50 ps?

Solution: Using (12.2.16) and $\Delta\omega \cong T_p^{-1}$ we find:

$$L_{max} = \Delta t/[\Delta\omega(d^2k/d\omega^2)] = 50 \text{ ps} \times 10 \text{ ps} /(20 \text{ ps}^2/\text{km}) = 25 \text{ km}$$

Thus we must slow this fiber to 10 Gbps if the amplifiers are 50 km apart.

12.3   Lasers

12.3.1   Physical principles of stimulated emission and laser amplification

Lasers (Light Amplification by Stimulated Emission of Radiation) amplify electromagnetic waves at wavelengths ranging from radio to ultraviolet and x-rays. They were originally called *masers* because the first units amplified only microwaves. Lasers can also oscillate when the amplified waves are reflected back into the device. The physical principles are similar at all wavelengths, though the details differ. Laser processes can occur in solids, liquids, or gases.

Lasers have a wide and growing array of applications. For example, optical fiber communications systems today commonly use *Erbium-doped fiber amplifiers* (EDFA's) that amplify ~1.5-micron wavelength signals having bandwidths up to ~4 THz. Semiconductor, gas, and glass fiber laser amplifiers are also used to communicate within single pieces of equipment and for local fiber or free-space communications. Lasers also generate coherent beams of light used for measuring distances and angles; recording and reading data from memory devices such as CD's and DVD's; and for cutting, welding, and shaping materials, including even the human eye. Laser pointers have been added to pocket pens while higher-power industrial units can cut steel plates several inches thick. Weapons and laser-driven nuclear fusion reactions require still higher-power lasers. Peak laser pulse powers can exceed 10^{15} watts, a thousand times the total U.S. electrical generating capacity of ~5×10^{11} watts. The electric field strengths within a focal spot of <100-micron diameter can strip electrons from atoms and accelerate them to highly relativistic velocities within a single cycle of the radiation. The roles of lasers in science, medicine, industry, consumer products, and other fields are still being defined.

Laser operation depends intimately upon the quantum nature of matter and the fact that charges trapped in atoms and molecules generally move at constant energy without radiating. Instead, transitions between atomic or molecular energy states occur abruptly, releasing or absorbing a photon.[69] This process and lasers can fortunately be understood semi-classically without reference to a full quantum description.

Electrons within atoms, molecules, and crystals occupy discrete *energy states*; the lower energy states are preferentially occupied. Energy states can also be vibrational, rotational, magnetic, chemical, nuclear, etc.[70] The number of possible states greatly exceeds those that are occupied.

---

[69] Alternatively, acoustic phonons with energy hf can be released or absorbed, or an additional molecular or atomic state transition can occur to conserve energy. Phonons are acoustic quanta associated with mechanical waves in materials. Optical transitions can also absorb or emit two photons with total energy equal to $E_2 - E_1$, although such *two-photon transitions* are much less likely.

[70] The distances between adjacent nuclei in molecules can oscillate sinusoidally with quantized amplitudes and frequencies characterisitic of each vibrational state. Isolated molecules can spin at specific frequencies corresponding to various rotational energy states. Electron spins and orbits together have magnetic dipole moments that align with or oppose an applied magnetic field to a quantized degree. Atoms bond to one another in quantized ways having specific chemical consequences. Nuclear magnetic moments can also align with other atomic or molecular magnetic moments in quantized ways corresponding to discrete energy states.

For example, as illustrated in Figure 12.3.1(a), an electron trapped in an atom, molecule, or crystal with energy E_1 can be excited into any vacant higher-energy state (E_2) by absorbing a photon of frequency f and energy ΔE where:

$$\Delta E = E_2 - E_1 = hf \ [J] \qquad\qquad (12.3.1)$$

The constant h is *Planck's constant* (6.625×10^{-34} [Js]), and the small circles in the figure represent electrons in specific energy states.
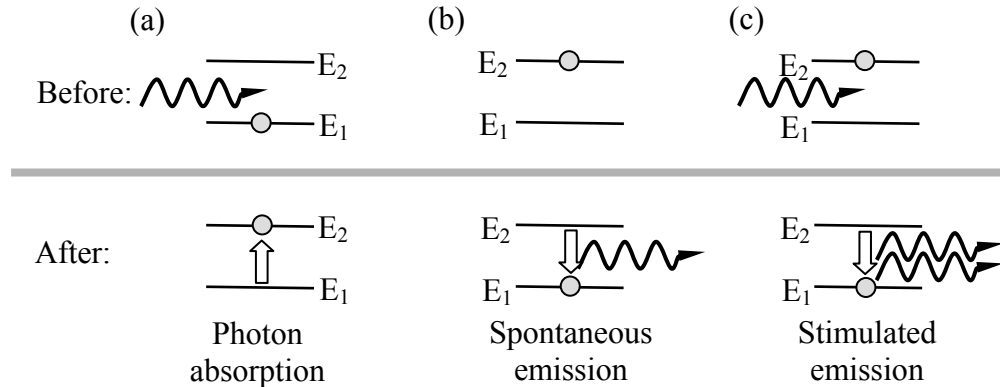
Figure 12.3.1 Photon absorption, spontaneous emission, and stimulated emission.

Figures 12.3.1(b) and (c) illustrate two additional basic photon processes: *spontaneous emission* and *stimulated emission*. *Photon absorption* (a) occurs with a probability that depends on the photon flux density [Wm^{-2}], frequency [Hz], and the cross-section for the energy transition of interest. Spontaneous emission of photons (b) occurs with a probability A that depends only on the transition, as discussed below. Stimulated emission (c) occurs when an incoming photon triggers emission of a second photon; the emitted photon is always exactly in phase with the first, and propagates in the same direction. Laser action depends entirely on this third process of stimulated emission, while the first two processes often weaken it.

The net effect of all three processes—absorption, spontaneous emission, and stimulated emission—is to alter the relative populations, N_1 and N_2, of the two energy levels of interest. An example exhibiting these processes is the Erbium-doped fiber amplifiers commonly used to amplify optical telecommunications signals near 1.4-micron wavelength on long lines. Figure 12.3.2 illustrates how an optical fiber with numerous atoms excited by an optical pump (discussed further below) can amplify input signals at the proper frequency. Since the number of excited atoms stimulated to emit is proportional to the input wave intensity, perhaps only one atom might be stimulated to emit initially (because the input signal is weak), producing two in-phase photons—the original plus the one stimulated. These two then propagate further stimulating two emissions so as to yield four in-phase photons.
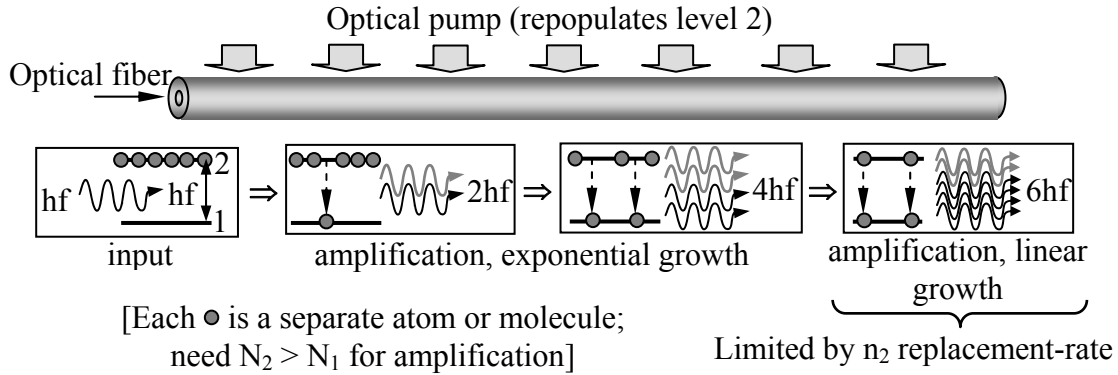
Figure 12.3.2 Optical fiber amplifier with exponential and linear growth.

This exponential growth continues until the pump can no longer empty E_1 and refill E_2 fast enough; as a result absorption $[m^{-1}]$ approaches emission $[m^{-1}]$ as N_1 approaches N_2 locally. In this limit the increase in the number of photons per unit length is limited by the number n_p of electrons pumped from E_1 to E_2 per unit length. Thereafter the signal strength then increases only linearly with distance rather than exponentially, as suggested in Figure 12.3.3; the power increase per unit length then approaches $n_p hf$ $[Wm^{-1}]$.
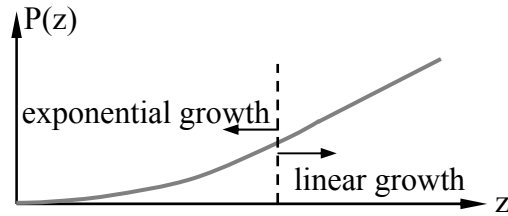
Figure 12.3.3 Exponential and linear growth regimes in optical fiber amplifiers.

Simple equations characterize this process quantitatively. If $E_1 < E_2$ were the only two levels in the system, then:

$$dN_2/dt = -A_{21}N_2 - I_{21}B_{21}(N_2 - N_1) \left[s^{-1} \right]$$

(12.3.2)

The probability of spontaneous emission from E_2 to E_1 is A_{21}, where $\tau_{21} = 1/A_{21}$ is the $1/e$ lifetime of state E_2. The intensity of the incident radiation at $f = (E_2 - E_1)/h$ [Hz] is:

$$I_{21} = F_{21}hf \left[Wm^{-2} \right]$$

(12.3.3)

where F_{21} is the photon flux [photons $m^{-2}s^{-1}$] at frequency f. The right-most term of (12.3.2) corresponds to the difference between the number of stimulated emissions ($\propto N_2$) and absorptions ($\propto N_1$), where the rate coefficients are:

$$B_{21} = A_{21}\left(\pi^2 c^2 / \hbar \omega^3 n^2\right)\left[m^2 J^{-1}\right] \tag{12.3.4}$$

$$A_{21} = 2\omega^3 D_{21}^2 / \hbar \varepsilon c^3 \left[s^{-1}\right] \tag{12.3.5}$$

In these equations n is the index of refraction of the fiber and D_{21} is the quantum mechanical electric or magnetic dipole moment specific to the state-pair 2,1. It is the sharply varying values of the *dipole moment* D_{ij} from one pair of levels to another that makes pumping practical, as explained below.

Laser amplification can occur only when N_2 exceeds N_1, but in a two-level system no pump excitation can accomplish this; even infinitely strong incident radiation I_{21} at the proper frequency can only equalize the two populations via (12.3.2).[71] Instead, three- or four-level lasers are generally used. The general principle is illustrated by the *three-level laser* of Figure 12.3.4(a), for which the optical *laser pump radiation* driving the 1,3 transition is so strong that it roughly equalizes N_1 and N_3. The key to this laser is that the spontaneous rate of emission A_{32} $\gg A_{21}$ so that all the active atoms quickly accumulate in the metastable long-lived level 2 in the absence of stimulation at f_{21}. This generally requires $D_{32} \gg D_{21}$, and finding materials with such properties for a desired laser frequency can be challenging.
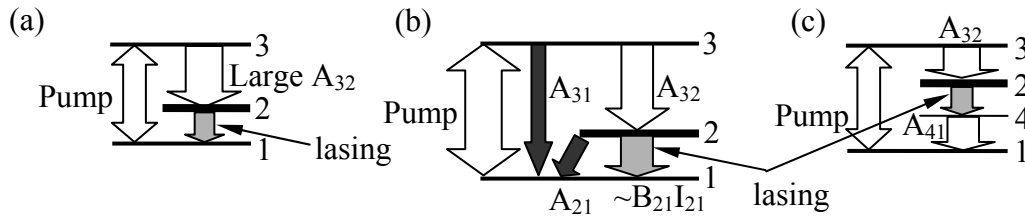
Figure 12.3.4   Energy diagrams for three- and four-level lasers.

Since it requires hf_{13} Joules to raise each atom to level 3, and only hf_{21} Joules emerges as amplified additional radiation, the power efficiency η (power out/power in) cannot exceed the intrinsic limit $\eta_I = f_{21}/f_{31}$. In fact the efficiency is lowered further by a factor of η_A corresponding to spontaneous emission from level 3 directly to level 1, bypassing level 2 as suggested in Figure 12.3.4(b), and to the spontaneous decay rate A_{21} which produces radiation that is not coherent with the incoming signal and radiates in all directions. Finally, only a fraction η_p of the pump photons are absorbed by the transition 1→2. Thus the maximum power efficiency for this laser in the absence of propagation losses is:

$$\eta = \eta_I \eta_A \eta_p \tag{12.3.6}$$

---

[71] Two-level lasers have been built, however, by physically separating the excited atoms or molecules from the unexcited ones. For example, excited ammonia molecules can be separated from unexcited ones by virtue of their difference in deflection when a beam of such atoms in vacuum passes through an electric field gradient.

Figure 12.3.4(c) suggests a typical design for a four-level laser, where both A_{32} and A_{41} are much greater than A_{24} or A_{21} so that energy level 2 is metastable and most atoms accumulate there in the absence of strong radiation at frequency f_{24} or f_{21}. The strong pump radiation can come from a laser, flash lamp, or other strong radiation source. Sunlight, chemical reactions, nuclear radiation, and electrical currents in gases pump some systems.

The ω^3 dependence of A_{21} (12.3.5) has a profound effect on maser and laser action. For example, any two-level maser or laser must excite enough atoms to level 2 to equal the sum of the stimulated and spontaneous decay rates. Since the spontaneous decay rate increases with ω^3, the pump power must also increase with ω^3 times the energy hf of each excited photon. Thus pump power requirements increase very roughly with ω^4, making construction of x-ray or gamma-ray lasers extremely difficult without exceptionally high pump powers; even ultraviolet lasers pose a challenge. Conversely, at radio wavelengths the spontaneous rates of decay are so extremely small that exceedingly low pump powers suffice, as they sometimes do in the vast darkness of interstellar space.

Many types of *astrophysical masers* exist in low-density interstellar gases containing H_2O, OH, CO, and other molecules. They are typically pumped by radiation from nearby stars or by collisions occurring in shock waves. Sometimes these lasers radiate radially from stars, amplifying starlight, and sometimes they spontaneously radiate tangentially along linear circumstellar paths that have minimal relative Doppler shifts. Laser or maser action can also occur in darkness far from stars as a result of molecular collisions. The detailed frequency, spatial, and time structures observed in astrophysical masers offer unique insights into a wide range of astrophysical phenomena.

Example 12.3A

What is the ratio of laser output power to pump power for a three-level laser like that shown in Figure 12.3.4(a) if: 1) all pump power is absorbed by the $1\rightarrow3$ transition, 2) $N_2 \gg N_1$, 3) $A_{21}/I_{21}B_{21} = 0.1$, 4) $A_{31} = 0.1A_{32}$, and 5) $f_{31} = 4f_{21}$?

Solution: The desired ratio is the efficiency η of (12.3.6) where the intrinsic efficiency is $\eta_I = f_{21}/f_{31} = 0.25$, and the pump absorption efficiency $\eta_p = 1$. The efficiency η_A is less than unity because of two small energy losses: the ratio $A_{31}/A_{32} = 0.1$, and the ratio $A_{21}/I_{21}B_{21} = 0.1$. Therefore $\eta_A = 0.9^2 = 0.81$, and $\eta = \eta_I\eta_A\eta_p = 0.25 \times 0.81 \cong 0.20$.

12.3.2 Laser oscillators

Laser amplifiers oscillate nearly monochromatically if an adequate fraction of the amplified signal is reflected back to be amplified further. For example, the *laser oscillator* pictured in Figure 12.3.5 has parallel mirrors at both ends of a laser amplifier, separated by L meters. One mirror is perfect and the other transmits a fraction T (say ~0.1) of the incident laser power. The roundtrip gain in the absence of loss is e^{2gL}. This system oscillates if the net roundtrip gain at any frequency exceeds unity, where round-trip absorption ($e^{-2\alpha L}$) and the partially transmitting mirror account for most loss.
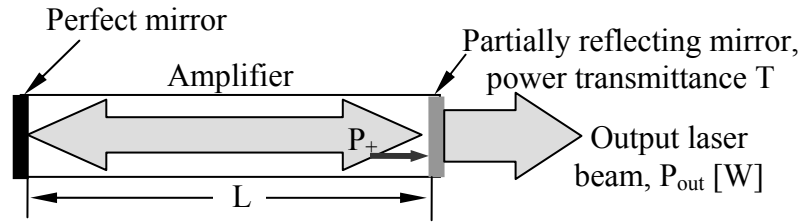
Figure 12.3.5   Laser oscillator.

Amplifiers at the threshold of oscillation are usually in their exponential region, so this net roundtrip gain exceeds unity when:

$$(1-T)e^{2(g-\alpha)L} > 1 \qquad\qquad\qquad (12.3.7)$$

Equation (12.3.7) implies $e^{2(g-\alpha)L} \geq (1-T)^{-1}$ for oscillation to occur. Generally the gain g per meter is designed to be as high as practical, and then L and T are chosen to be consistent with the desired output power. The pump power must be above the minimum threshold that yields $g > \alpha$.

The output power from such an oscillator is simply $P_{out} = TP_+$ watts, and depends on pump power P_{pump} and laser efficiency. Therefore:

$$P_+ = P_{out}/T = \eta\, P_{pump}/T \qquad\qquad\qquad (12.3.8)$$

Thus small values of T simply result in higher values of P_+, which can be limited by internet breakdown or failure.

One approach to obtaining extremely high laser pulse powers is to abruptly increase the Q (reverberation) of the laser resonator after the pump source has fully populated the upper energy level. To prevent lasing before that level is fully populated, strong absorption can be introduced in the round-trip laser path to prevent amplification of any stimulated emission. The instant the absorption ceases, i.e. after Q-switching, the average round-trip gain g of the laser per meter exceeds the average absorption α and oscillation commences. At high Q values lasing action is rapid and intense, so the entire upper population is encouraged to emit instantly, particularly if the lower level can be rapidly emptied. Such a device is called a *Q-switched laser*. Resonator Q is discussed further in Section 7.8.

The electronic states of glass fiber amplifiers are usually associated with quantized electron orbits around the added Erbium atoms, and state transitions simply involve electron transfers between two atomic orbits having different energies. In contrast, the most common lasers are *laser diodes*, which are transparent semiconductor p-n junctions for which the electron energy transitions occur between the conduction and valence bands, as suggested in Figure 12.3.6.

(a)

(b)
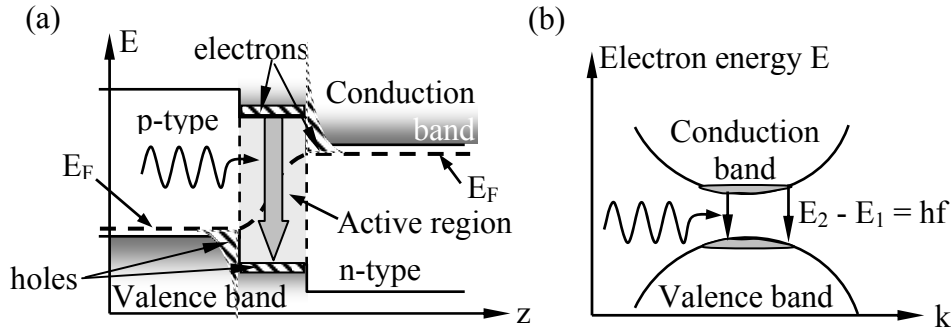
Figure 12.3.6   Laser diode – a forward-biased p-n junction
bounded by mirrors promoting oscillation.

Parallel mirrors at the sides of the *p-n junction* partially trap the laser energy, forming an oscillator that radiates perpendicular to the mirrors; one of the mirrors is semi-transparent. Strong emission does not occur in any other direction because without the mirrors there is no feedback.   Such lasers are pumped by forward-biasing the diode so that electrons thermally excited into the n-type conduction band diffuse into the active region where photons can stimulate emission, yielding amplification and oscillation within the ~0.2-μm thick p-n junction. Vacancies in the valence band are provided by the holes that diffuse into the active region from the p-type region.   Voltage-modulated laser diodes can produce digital pulse streams at rates above 100 Mbps.

The vertical axis E of Figure 12.3.6(a) is electron energy and the horizontal axis is position z through the diode from the p to n sides of the junction.   The exponentials suggest the Boltzmann energy distributions of the holes and electrons in the valence and conduction bands, respectively. Below the *Fermi level*, E_F, energy states have a high probability of being occupied by electrons; $E_F(z)$ tilts up toward the right because of the voltage drop from the p-side to the n-side.   Figure 12.3.6(b) plots electron energy E versus the magnitude of the k vector for electrons (quantum approaches treat electrons as waves characterized by their wavenumber k), and suggests why diode lasers can have broad bandwidths: the energy band curvature with k broadens the *laser linewidth* Δf.   Incoming photons can stimulate any electron in the conduction band to decay to any empty level (hole) in the valence band, and both of these bands have significant energy spreads ΔE, where the linewidth $\Delta f \cong \Delta E/h$ [Hz].

The resonant frequencies of laser diode oscillators are determined by $E_2 - E_1$, the linewidth of that transition, and by the resonant frequencies of the TEM mirror cavity resonator.   The width $\Delta\omega$ of each resonance is discussed further later.   If the mirrors are perfect conductors that force $\overline{E}_{//} = 0$, then there must be an integral number m of half wavelengths within the cavity length L so that $m\lambda_m = 2L$.   The wavelength λ_m' is typically shorter than the free-space wavelength λ_m due to the index of refraction n of the laser material.   Therefore $\lambda_m = 2Ln/m = c/f_m$, and:

$$f_m = cm/2Ln \qquad\qquad (12.3.9)$$

For typical laser diodes L and n might be 0.5 mm and 3, respectively, yielding a spacing between cavity resonances of: $c/2Ln = 3\times10^8/(2\times10^{-3}\times1.5) = 100$ GHz, as suggested in Figure 12.3.7(a). The figure suggests how the natural (atomic) laser line width might accommodate multiple cavity resonances, or possibly only one.
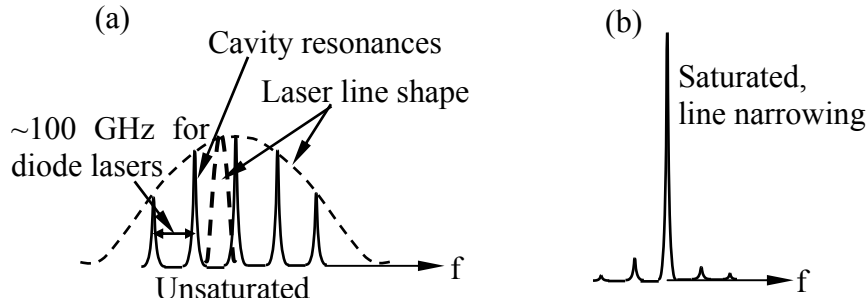
Figure 12.3.7   Line widths and frequencies of the resonances of a cavity laser.

If the amplifier line shape is narrow compared to the spacing between cavity resonances, then the cavity length L might require adjustment in order to place one of the cavity resonances on the line center before oscillations occur. The line width of a laser depends on the widths of the associated energy levels E_i and E_j. These can be quite broad, as suggested by the laser diode energy bands illustrated in Figure 12.3.6(b), or quite narrow. Similarly, the atoms in an EDFA are each subject to slightly different local electrical fields due to the random nature of the glassy structure in which they are imbedded. This results in each atom having slightly different values for E_i so that EFDA's amplify over bandwidths much larger than the bandwidth of any single atom.

Lasers for which each atom has its own slightly displaced resonant frequency due to local fields are said to exhibit *inhomogeneous line broadening*. In contrast, many lasers have no such frequency spread induced by local factors, so that all excited atoms exhibit the same line center and width; these are said to exhibit *homogeneous line broadening*. The significance of this difference is that when laser amplifiers are saturated and operate in their linear growth region, homogeneously broadened lasers permit the strongest cavity resonance within the natural line width to capture most of the energy available from the laser pump, suppressing the rest of the emission and narrowing the line, as suggested in Figure 12.3.7(b). This suppression of weak resonances is reduced in inhomogeneously broadened lasers because all atoms are pumped equally and have their own frequency sub-bands where they amplify independently within the natural line width.

In gases the width of any spectral line is also controlled by the frequency of molecular collisions. Figure 12.3.8(b) illustrates how an atom or molecule with sinusoidal time variations in its dipole moment might be interrupted by collisions that randomly reset the phase. An electromagnetic wave interacting with this atom or molecule would then see a less pure sinusoid. This new spectral characteristic would no longer be a spectral impulse, i.e., the Fourier transform of a pure sinusoid, but rather the transform of a randomly interrupted sinusoid, which has the *Lorentz line shape* illustrated in Figure 12.3.8(a).   Its half-power width is Δf, which is approximately the collision frequency divided by 2π. The limited lifetime of an atom or

molecule in any state due to the probability A of spontaneous emission results in similar broadening, where $\Delta f \cong A/2\pi$; this is called the intrinsic *linewidth* of that transition.
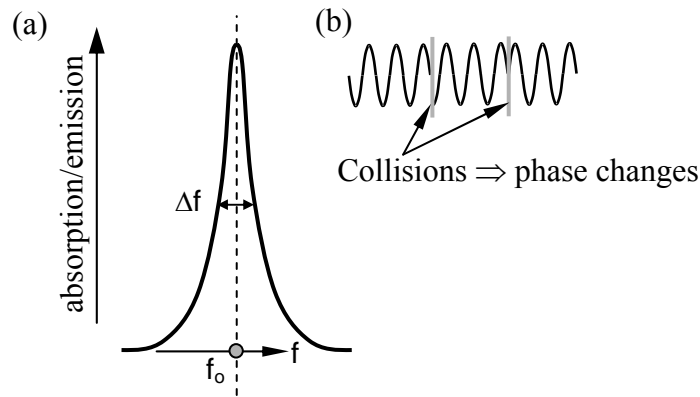
Figure 12.3.8   Lorentzian line shape and origins of intrinsic line width.

Example 12.3B

A Q-switched 1-micron wavelength laser of length L = 1 mm is doped with 10^{18} active atoms all pumped to their upper state. When the Q switches instantly to 100, approximately what is the maximum laser power output P [W]? Assume $\varepsilon = 4\varepsilon_o$.

Solution: The total energy released when the Q switches is $10^{18}hf \cong 10^{18} \times 6.6\times10^{-34} \times 3\times10^{14} =$ 0.20 Joules. If the laser gain is sufficiently high, then a triggering photon originating near the output could be fully amplified by the time the beam reaches the rear of the laser, so that all atoms would be excited as that reflected pulse emerges from the front of the laser. A triggering photon at the rear of the laser would leave some atoms unexcited. Thus the minimum time for full energy release lies between one and two transit times τ of the laser, depending on its gain; $\tau = L/c' = 2L/c = 6.7\times10^{-12}$. Lower laser gains may require many transit times before all atoms are stimulated to emit. Therefore $P < \sim 0.2/(6.7\times10^{-12}) \cong 30$ GW.

12.4   *Optical detectors, multiplexers, interferometers, and switches*

12.4.1   Phototubes

Sensitive radio-frequency detectors typically require at least 10^{-20} Joules per bit of information, which roughly corresponds to thousands of photons of energy hf, where Planck's constant h = 6.625×10^{-34} Joules Hz^{-1}. This number of photons is sufficiently high that we can ignore most quantum effects and treat the arriving radio signals as traditional waves. In contrast, many optical detectors can detect single photons, although more than five photons are typically used to distinguish each pulse from interference; this requires more energy per bit than is needed at radio wavelengths. The advantage of long-range optical links lies instead in the extremely low losses of optical fibers or, alternatively, in the ability of relatively small mirrors or telescopes to focus

energy in extremely small beams so as to achieve much higher gains than can practical radio antennas.

Typical photon detectors include phototubes and semiconductors. A *phototube* detects photons having energies hf > Φ using the *photoelectric effect*, where Φ is the *work function* [J] of the metal surface (*cathode*) that intercepts the photons. Photons with energies above this threshold eject an electron from the cathode with typical probabilities η (called the *quantum efficiency*) of ~10-30 percent. These ejected electrons are then pulled in vacuum toward a positively charged *anode* and contribute to the current I through the load resistor R, as illustrated in Figure 12.4.1(a). Although early phototubes ejected electrons from the illuminated surface, it is now common for the metal to be sufficiently thin and transparent that the electrons are emitted from the backside of the metal into vacuum; the metal is evaporated in a thin layer onto the interior surface of the tube's evacuated glass envelope.
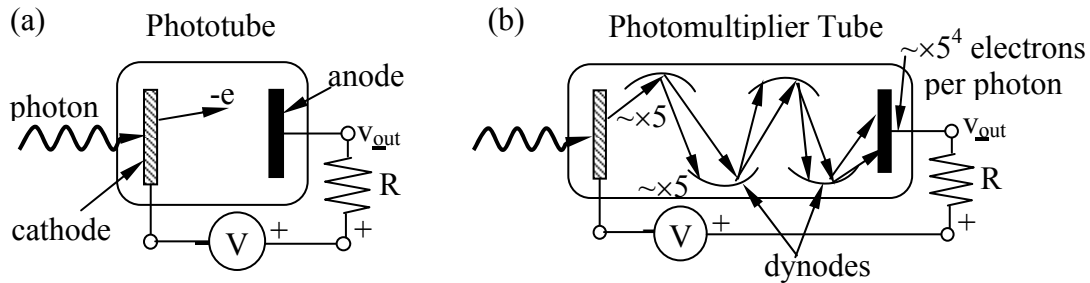
Figure 12.4.1   Phototube and photomultiplier tube detectors.

The current I is proportional to the number N of photons incident per second with energies above Φ:

$$I = -\eta N e \ [A] \tag{12.4.1}$$

The work functions of most metals are ~2-6 electron volts, where the energy of one *electron volt* (e.v.) is $-eV = 1.602 \times 10^{-19}$ Joules[72]. Therefore phototubes do not work well for infrared or longer wavelengths because their energy hf is too small; 2 e.v. corresponds to a wavelength of 0.62 microns and the color red.

Because the charge on an electron is small, the currents I are often too small to induce voltages across R (see Figure 12.4.1) that exceed the thermal noise (Johnson noise) of the resistor unless the illumination is bright. *Photomultiplier tubes* release perhaps 10^4 electrons per detected photon so as to overcome this noise and permit each detected photon to be unambiguously counted. The structure of a typical photomultiplier tube is illustrated in Figure 12.4.1(b). Each photoelectron emitted by the cathode is accelerated toward the first *dynode* at ~50-100 volts, and gains energy sufficient to eject perhaps five or more low energy electrons from the dynode that are then accelerated toward the second dynode to be multiplied again. The

---

[72] Note that the energy associated with charge Q moving through potential V is QV Joules, so QV = 1 e.v. = e×1 = 1.602×10^{-19} Joules.

illustrated tube has four dynodes that, when appropriately charged, each multiply the incident electrons by ~5 to yield ~$5^4 \cong 625$ electrons at the output for each photon detected at the input. Typical tubes have more dynodes and gains of ~10^4-10^7. Such large current pulses generally overwhelm the thermal noise in R, so random electron emissions induced by cosmic rays or thermal effects at the cathode dominate the detector noise. The collecting areas of such tubes can be enhanced with lenses or mirrors.

12.4.2   Photodiodes

Phototubes are generally large (several cubic inches), expensive, and fragile, and therefore semiconductor *photodiodes* are more commonly used. Photodiodes also respond better to visible and infrared wavelengths and operate at much lower voltages. Figure 12.4.2(a) illustrates the *energy diagram* for a typical short-circuited *p-n junction* between p-type and n-type semiconductors, where the vertical axis is electron energy E and the horizontal axis is distance z perpendicular to the planar junction.
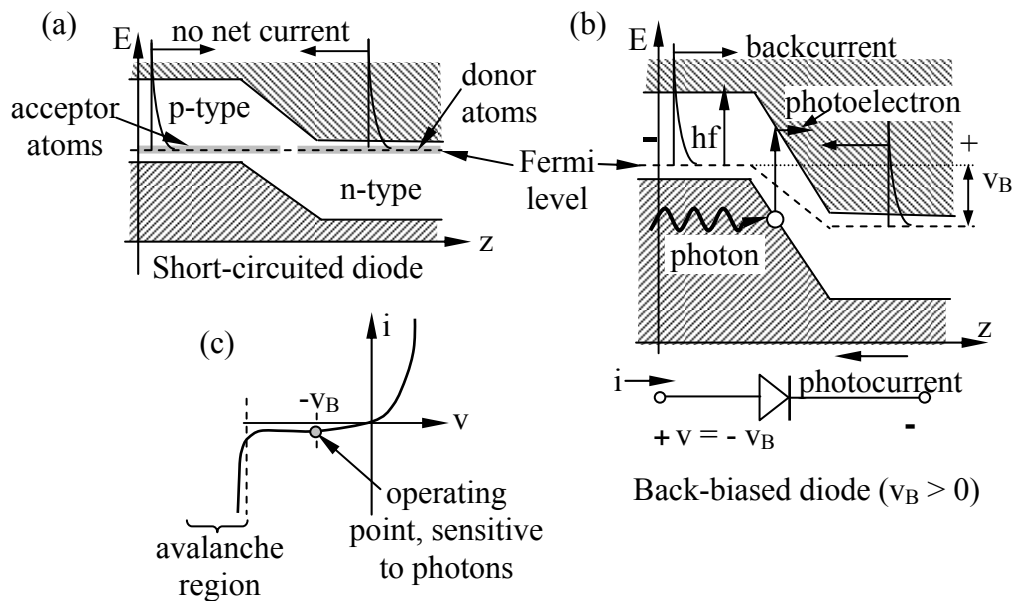
Figure 12.4.2   Semiconductor photodiodes.

The lower cross-hatched area is the *valence band* within which electrons are lightly bound to ions, and the upper area is the *conduction band* within which electrons are free to move in response to electric fields. The *band gap* between these regions is ~1.12 electron volts (e.v.) for silicon, and ranges from 0.16 e.v. for InSb (indium antimonide to ~7.5 e.v. for BN (boron nitride). In metals there is no such gap and some electrons always reside in the conduction band and are mobile. Additional discussion of p-n junctions appears in Section 8.2.4.

Electrons move freely in the conduction band, but not if they remain in the valence band. Most photons entering the junction region with energy greater than the bandgap between the Fermi level and the lower edge of the conduction band can excite electrons into the conduction

band to enhance device conductivity. In semiconductors the *Fermi level* is that level corresponding to the nominal maximum energy of electrons available for excitation into the conduction band. The local Fermi level is determined by impurities in the semiconductors that create electron donor or acceptor sites; these sites easily release or capture, respectively, a free electron. The Fermi level sits just below the conduction band for n-type semiconductors because *donor atoms* easily release one of their electrons into the conduction band, as illustrated in Figure 12.4.2(a) and (b). The Fermi level sits just above the valence band for p-type semiconductors because *acceptor atoms* easily capture an extra electron from bound states in nearby atoms.

If the p-n junction is short-circuited externally, the Fermi level is the same on both halves, as shown in Figure 12.4.2(a). Random *thermal excitation* produces an exponential *Boltzmann distribution* in electron energy, as suggested in the figure, the upper tails of which lie in the conduction band on both halves of the junction. When the device is short-circuited these current flows from thermal excitations in the p and n halves of the junction balance, and the external current is zero. If, however, the diode is back-biased by V_B volts as illustrated in Figure 12.4.2(b), then the two exponential tails do not balance and a net back-current current flows, as suggested by the I-V characteristic for a p-n junction illustrated in Figure 12.4.2(c). The back current for an un-illuminated photodiode approaches an asymptote determined by V_B and the number of thermal electrons excited per second into the conduction band for the p-type semiconductor. When an un-illuminated junction is forward biased, the current increases roughly exponentially.

When a p-n junction is operated as a photodiode, it is back-biased so that every detected photon contributes current flow to the circuit, nearly one electron per photon received. By cooling the photodiode the thermal contribution to diode current can be reduced markedly so that the diode becomes more sensitive to dim light. Cooling is particularly important for photodiodes with the small bandgaps needed for detecting infrared radiation; otherwise the detected infrared signals must be bright so they exceed the detector noise.

If photodiodes are sufficiently back-biased, they can enter the avalanche region illustrated in Figure 12.4.2(c), where an excited electron is accelerated sufficiently as it moves through the semiconductor that it can impact and excite another electron into the conduction band; both electrons can now accelerate and excite even more electrons, exponentially, until they all exit the high-field zone so that further excitations are not possible. In response to a single detected photon such *avalanche photodiodes* (APD's) can produce an output pulse of $\sim 10^4$ electrons that stands out sufficiently above the thermal noise that photons can again be counted individually. The number of photons detected per second is proportional to input power, and therefore to the square of the incident electric field strength.

12.4.3   Frequency-multiplexing devices and filters

The major components in fiber-optic communications systems are the fibers themselves and the optoelectronic devices that manipulate the optical signals, such as detectors (discussed in Sections 12.4.1–2), amplifiers and sources (Section 12.3), multiplexers and filters (this section), modulators, mixers, switches, and others (Section 12.4.4). These are assembled to create useful communications, computing, or other systems.

A typical wave-division multiplexed (WDM) amplifier is pictured in Figure 12.4.3; narrowband optical signals of different colors are aggregated at a point of departure and merged onto a single long fiber by a *frequency multiplexer* (MUX). Along this fiber extremely broadband optical amplifiers (OAMPs) are spaced perhaps 80 km apart to sustain the signal strength. OAMPs today are typically erbium-doped fiber amplifiers (EFDA's). At the far end the signal is de-multiplexed into its spectral components, which are then directed appropriately along separate optical fibers. Before broadband amplifiers were available, each narrow band had to have separate amplifiers and often separate fibers.
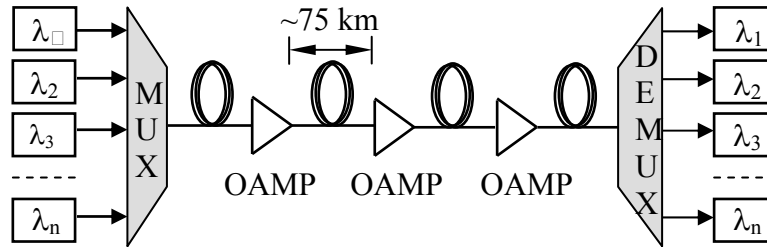
Figure 12.4.3   Wave-division multiplexed amplifier.

Such multiplexers can be made using prisms or diffraction gratings that refract or diffract different colors at different angles, as suggested in Figure 12.4.4(a) and (b); by reciprocity the same devices can be used either for multiplexing (superimposing multiple frequency bands on one beam) or demultiplexing (separation of a single beam into multiple bands), depending on which end of the device receives the input.

The diffraction grating of Figure 12.4.4(b) is typically illuminated by normally incident uniform plane waves, and consists of closely spaced ruled straight lines where equal-width stripes typically alternate between transmission and reflection or absorption. Alternate stripes sometimes differ only in their phase. Each stripe must be more than $\lambda/2$ wide, and λ is more typical. In this case the rays from each transparent stripe 2λ apart will add in phase straight forward ($\theta = 0$) and at $\theta = \sin^{-1}(\lambda/2\lambda) = 30^o$, exactly analogous to the grating lobes of dipole array antennas (Section 10.4). Since the stripe separation (2λ here) is fixed, as the frequency $f = c/\lambda$ varies, so does θ, thus directing different frequencies toward different angles of propagation, much like the prism.

Another useful optical device is the *Fabry-Perot resonator*, which is the optical version of the TEM resonator illustrated in Figure 7.4.3(a) and explained in Section 7.4.3. For example, an optical TEM resonator can be fabricated using parallel mirrors with uniform plane waves trapped between them; the allowed resonator modes have an integral number n of half-wavelengths in the distance L between the parallel conductors:

$$n\lambda_n/2 = L \tag{12.4.2}$$
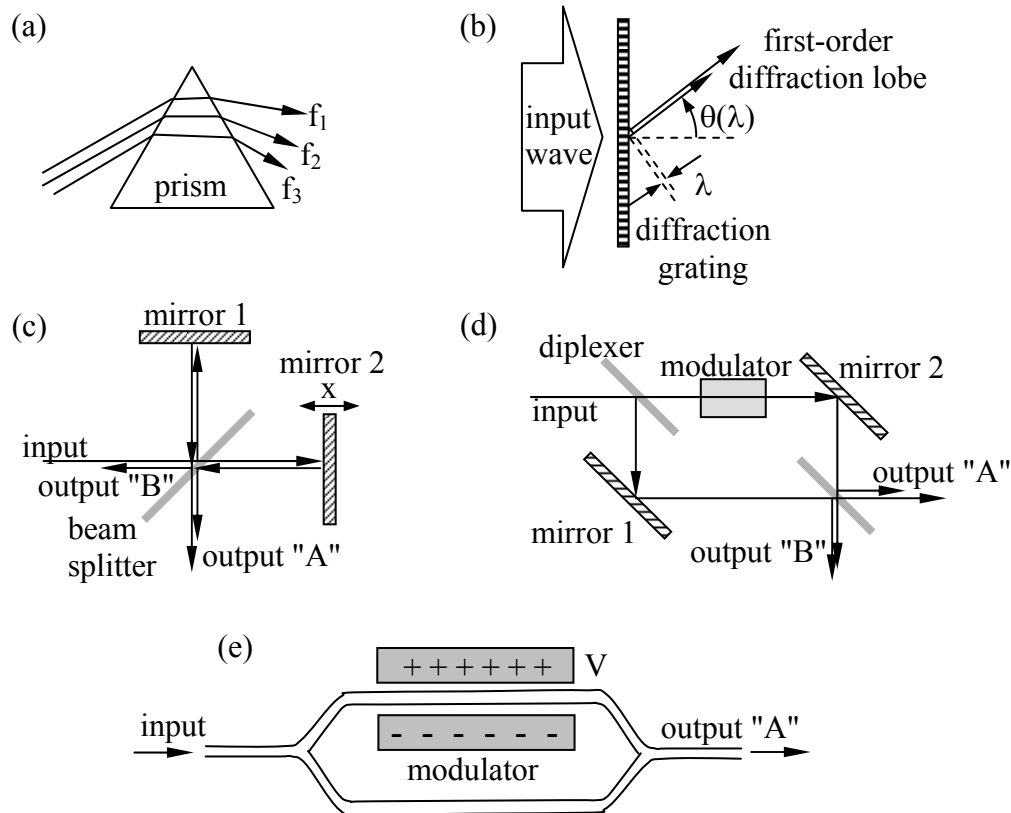
Figure 12.4.4 Optical frequency multiplexers, interferometers, and modulators:
(a) prism, (b) diffraction grating, (c) Michelson interferometer, (d) Mach-Zehnder interferometer,
(e) waveguide Mach-Zehnder interferometer.

Thus the frequency f_n of the nth TEM resonance is:

$$f_n = c/\lambda_n = nc/2L \quad [Hz] \tag{12.4.3}$$

and the separation between adjacent resonances is c/2L [Hz]. For example, if L is 1.5 mm, then the resonances are separated $3 \times 10^8 / 0.003 = 100$ GHz.

If the input and output mirrors transmit the same small fraction of the power incident upon them, then the "internal" and external Q's of this resonator are the same, where the "internal Q" here (Q_I) is associated with power escaping through the output mirror and the "external Q" (Q_E) is associated with power escaping through the input mirror. As suggested by the equivalent circuits in Figures 7.4.4–5, there is perfect power transmission through the resonator at resonance when the internal and external Q's are the same, provided there are no dissipative losses within the resonator itself. The width of the resonance is then found from (7.4.45–6):

$$\Delta\omega = \omega_o/Q_L = 2\omega_o/Q_E = 2P_E/w_T \tag{12.4.4}$$

- 394 -

The loaded $Q_L = Q_E/2$ when $Q_E = Q_I$, and P_E [W] is the power escaping through the input mirror when the total energy stored in the resonator is w_T [J]. With high reflectivity mirrors and low residual losses the bandwidth of such a resonator can be made almost arbitrarily narrow. At optical frequencies the ratio of cavity length L to wavelength λ is also very large. This increases the ratio of w_T to P_E proportionally and leads to very high Q_L and narrow linewidth.

If the medium in a Fabry-Perot interferometer is dispersive, then it can be shown that the spacing between resonances is $v_g/2L$ [Hz], or the reciprocal of the round-trip time for a pulse. Thus such a resonator filled with an active medium could amplify a single pulse that rattles back and forth in the resonator producing sharp output pulses with a period $2L/v_g$. The Fourier transform of this pulse train is a train of impulses in the frequency domain with spacing $v_g/2L$, i.e., representing the set of resonant frequencies for this resonator. The resonant modes of such a *mode-locked laser* are synchronized, so they can usefully generate pulse trains for subsequent modulation.

Example 12.4A

What is the ratio of the width $\Delta\omega$ of the passband for a Fabry-Perot resonator relative to the spacing $\omega_{i+1} - \omega_i$ between adjacent resonances? What power transmission coefficient $T^2 = |\underline{E}_t|^2/|\underline{E}_i|^2$ is required for each mirror in order to produce isolated sharp resonances? What is the width Δf [Hz] of each resonance?

Solution: The resonance width $\Delta\omega$ and spacing are given by (12.4.4) and (12.4.2), respectively, so:

$$\Delta\omega/(\omega_{i+1} - \omega_i) = (2P_E/w_T)/(\pi c/L) = \left[(2P_+ T^2)/(2LP_+/c)\right](L/\pi c) = T^2/\pi <\sim 0.3$$

Therefore we require $T^2 < \sim 1$ so that $\Delta\omega <\sim 0.3(\omega_{i+1} - \omega_i)$.

$\Delta f = \Delta\omega/2\pi = (T^2/\pi)(\omega_{i+1} - \omega_i)/2\pi = (T^2/\pi)c/2L$ [Hz]; it approaches zero as T^2/L does.

12.4.4   Interferometers

The *Michelson interferometer* and the *Mach-Zehnder interferometer* are important devices illustrated in Figure 12.4.4(c) and (d), respectively. In both cases an input optical beam is split by a *beam-splitter* into two coherent beams that are reflected by mirrors and then recombined coherently in a second beam-splitter to form two output beams. The intensity of each output beam depends on whether its two input components added in-phase or out-of-phase. The beam-splitters are typically dielectric mirrors coated so that half the power is reflected and half transmitted from the front surface; the rear surface might be anti-reflection coated. Half-silvered mirrors can also be used.

As the position x of a Michelson mirror varies, the output power varies sinusoidally from zero, which results when the two beams cancel at the output, to its peak value when the two beams add in phase. Cancellation requires that the two beams have equal strength. It is interesting to ask where the input power goes when output "A" is zero; the figure suggests the answer. The missing power emerges from the other output; the sum of the powers emerging

from the two outputs equals the input power, less any dissipative losses. This requirement for power conservation translates into a requirement for a specific phase relationship between the various beams.

Consecutive peaks in output strength occur as the mirror moves $\lambda/2$ (typically $\sim 3 \times 10^{-7}$ m); the factor of 1/2 arises because of the round trip taken by the reflected beam. The sinusoidal output power can generally be measured with sufficient accuracy at optical wavelengths to determine relative mirror positions x with accuracies of an angstrom (10^{-10} m), or tiny fractions thereof; thus the Michelson interferometer is a powerful tool for measuring or comparing wavelengths and distances. Another important application is measurement of optical spectra. Since each optical wavelength λ in the input beam produces an additive sinusoidal contribution to the output power waveform A(x) of period $\lambda/2$, the input optical power spectrum is the Fourier transform of A(2x). Because this technique was first used to analyze infrared spectra, it is called *Fourier transform infrared spectroscopy* (FTIR).

If the two output beams in a Mach-Zehnder interferometer add in phase, the output power is maximized and equals the input power, much like the Michelson interferometer. In either type of interferometer the phase of the optical beam in one arm can be modulated by varying the effective dielectric constant and delay of its propagation medium; certain dielectrics are tunable when biased with large electric fields. In this fashion the output beam power can be modulated by varying the voltage V across the propagation medium, as illustrated in Figure 12.4.4(d) and (e). If the device operates near a transmission null, very little change in refractive index is required to produce a large increase in output power. Such devices can modulate optical power at frequencies of 10 GHz or more.

The Mach-Zehnder interferometer configuration in Figure 12.4.4(e) is widely used for modulators because the waveguides can be integrated on a chip together with other optical components. The output is maximum when the two arms have equal phase delays. When the two merging beams are out of phase the excited waveguide mode is not trapped in the output waveguide but radiates away; the radiated wave corresponds to output B in Figure 12.4.4(d). The same integrated configuration can alternatively be used as a notch filter, eliminating an undesired optical wavelength for which the two arm lengths differ by exactly $\lambda/2$, while passing nearby wavelengths.

12.4.5   Optical switches

Optical switches redirect optical beams just as electrical switches redirect currents. One approach is to use MEMS devices that mechanically move mirrors or shutters to redirect the light beams, which usually are narrow, coherent, and laser-produced. Such devices can switch light beams at rates approaching 1 MHz.

Another approach is to direct the light beam at right angles to a dielectric within which ultrasonic acoustic waves (at radio frequencies) are propagating transverse to the light so as to produce a dynamic phase grating through which the light propagates and diffracts. The configuration is that of Figure 12.4.4(b). The acoustic waves compress and decompress the medium in a wavy pattern; the compressed regions have a slightly higher permittivity and

therefore a slightly lower velocity of light. By making the dielectric sufficiently thick, the cumulative phase variation of the light passing through the device can be $\lambda/2$ or more, thus producing strong diffraction at an angle θ corresponding to the wavelength of light λ and the acoustic wavelength λ_a, where $\theta = \sin^{-1}(\lambda/\lambda_a)$ and $\lambda_a \cong 2\lambda$. In practice the cumulative phase variation is often much less than $\lambda/2$ because of improved simplicity, linearity, and the availability of high input powers that can compensate for the reduced diffractive power efficiency. In this fashion the diffracted beam can be steered among several output ports at rates up to ~1 MHz or more, limited largely by the time it takes the acoustic wave to traverse the diffraction zone. Acoustic velocities in solids are roughly 1000-3000 m/s.

A more important method, however, is the use of Mach-Zehnder interferometers (see Section 12.4.4) to modulate input optical streams, varying their intensity by more than 15 dB at rates up to ~10 GHz, limited by the time it takes the signals modulating the electrical phase length modulator of Figure 12.4.4(e) to propagate across that modulator (e.g. nanoseconds). The detected modulator output signal is the product of the optical and modulator signals. The spectrum of this product contains the convolution of the two input spectra, which exhibits upper and lower sidebands that correspond to the radio frequency signal being communicated.

Chapter 13: Acoustics

13.1 Acoustic waves

13.1.1 <u>Introduction</u>

Wave phenomena are ubiquitous, so the wave concepts presented in this text are widely relevant. Acoustic waves offer an excellent example because of their similarity to electromagnetic waves and because of their important applications. Beside the obvious role of acoustics in microphones and loudspeakers, surface-acoustic-wave (SAW) devices are used as radio-frequency (RF) filters, acoustic-wave modulators diffract optical beams for real-time spectral analysis of RF signals, and mechanical crystal oscillators currently control the timing of most computers and clocks. Because of the great similarity between acoustic and electromagnetic phenomena, this chapter also reviews much of electromagnetics from a different perspective.

Section 13.1.2 begins with a simplified derivation of the two main differential equations that characterize linear acoustics. This pair of equations can be combined to yield the acoustic wave equation. Only longitudinal acoustic waves are considered here, not transverse or "shear" waves. These equations quickly yield the group and phase velocities of sound waves, the acoustic impedance of media, and an acoustic Poynting theorem. Section 13.2.1 then develops the acoustic boundary conditions and the behavior of acoustic waves at planar interfaces, including an acoustic Snell's law, Brewster's angle, the critical angle, and evanescent waves. Section 13.2.2 shows how acoustic plane waves can travel within pipes and be guided and manipulated much as plane waves can be manipulated within TEM transmission lines.

Acoustic waves can be totally reflected at firm boundaries, and Section 13.2.3 explains how they can be trapped and guided in a variety of propagation modes closely resembling those in electromagnetic waveguides, where they exhibit cutoff frequencies of propagation and evanescence below cutoff. Section 13.2.4 then explains how these guides can be terminated at their ends with open or closed orifices, thus forming resonators with Q's that can be controlled as in electromagnetic resonators so as to yield band-stop or band-pass filters. The frequencies of acoustic resonances can be perturbed by distorting the shape of the cavity, as governed by nearly the same equation used for electromagnetic resonators except that the electromagnetic energy densities are replaced by acoustic energy density expressions. Section 13.3 discusses acoustic radiation and antennas, including antenna arrays, and Section 13.4 concludes the chapter with a brief introduction to representative electroacoustic devices.

13.1.2 <u>Acoustic waves and power</u>

Most waves other than electromagnetic waves involve perturbations. For example, acoustic waves involve perturbations in the pressure and velocity fields in gases, liquids, or solids. In gases we may express the total pressure p_T, density ρ_T, and velocity $\overline{u}_T$ fields as the sum of a static component and a dynamic perturbation:

$$p_T(\overline{r},t) = P_o + p(\overline{r},t) \ \left[N/m^2\right] \tag{13.1.1}$$

$$\rho_T(\overline{r},t) = \rho_o + \rho(\overline{r},t) \ \left[kg/m^3\right] \tag{13.1.2}$$

$$\overline{u}(\overline{r},t) \ = \overline{U}_o + \overline{u}(\overline{r},t) \ [m/s] \tag{13.1.3}$$

Another complexity is that, unlike electromagnetic variables referenced to a particular location, gases move and compress, requiring further linearization.[73]   Most important is the approximation that the mean velocity $\overline{U}_o = 0$. After these simplifying steps we are left with two linearized acoustic equations, *Newton's law* (f = ma) and *conservation of mass*:

$$\nabla p \cong -\rho_o \partial \overline{u}/\partial t \ \left[N/m^3\right] \qquad \text{(Newton's law)} \tag{13.1.4}$$

$$\rho_o \nabla \bullet \overline{u} + \partial \rho/\partial t \cong 0 \ \left[kg/m^3 s\right] \qquad \text{(conservation of mass)} \tag{13.1.5}$$

Newton's law states that the pressure gradient will induce mass acceleration, while conservation of mass states that velocity divergence $\nabla \bullet \overline{u}$ is proportional to the negative time derivative of mass density.

These two basic equations involve three key variables: p, $\overline{u}$, and ρ; we need the acoustic constitutive relation to reduce this set to two variables. Most acoustic waves involve frequencies sufficiently high that the heating produced by wave compression has no time to escape by conduction or radiation, and thus this heat energy returns to the wave during the subsequent expansion without significant loss. Such *adiabatic processes* involve no heat transfer across populations of particles. The resulting adiabatic *acoustic constitutive relation* states that the fractional change in density equals the fractional change in pressure, divided by a constant γ, called the adiabatic exponent:

$$\partial \rho/\partial p = \rho_o/\gamma P_o \tag{13.1.6}$$

The reason γ is not unity is that gas heats when compressed, which further increases the pressure, so the gas thereby appears to be slightly "stiffer" or more resistant to compression than otherwise. This effect is diminished for gas particles that have internal rotational or vibrational degrees of freedom so the temperature rises less upon compression. Ideal monatomic molecules without such degrees of freedom exhibit $\gamma = 5/3$, and $1 < \gamma < 2$, in general.

Substituting this constitutive relation into the mass equation (13.1.5) replaces the variable ρ with p, yielding the *acoustic differential equations*:

$$\nabla p \cong -\rho_o \partial \overline{u}/\partial t \ \left[N/m^3\right] \qquad \text{(Newton's law)} \tag{13.1.7}$$

---

[73] The Liebnitz identity facilitates taking time derivatives of integrals over volumes deforming in time.

$$\nabla \bullet \overline{u} = -\left(1/\gamma P_o\right)\partial p/\partial t \qquad\qquad\qquad (13.1.8)$$

These two differential equations are roughly analogous to Maxwell's equations (2.1.5) and (2.1.6), and can be combined. To eliminate $\overline{u}$ from Newton's law we operate on it with $(\nabla \bullet)$, and then substitute (13.1.8) for $\nabla \bullet \overline{u}$ to form the *acoustic wave equation*, analogous to the Helmholtz wave equation (2.2.7):

$$\nabla^2 p - \left(\rho_o/\gamma P_o\right)\partial^2 p/\partial t^2 = 0 \qquad\qquad \text{(acoustic wave equation)} \quad (13.1.9)$$

Wave equations state that the second spatial derivative equals the second time derivative times a constant. If the constant is not frequency dependent, then any arbitrary function of an argument that is the sum or difference of terms linearly proportional to time and space will satisfy this equation; for example:

$$p\left(\overline{r},t\right) = p(\omega t - \overline{k} \bullet \overline{r})\ \left[N/m^2\right] \qquad\qquad\qquad (13.1.10)$$

where $p(\bullet)$ is an arbitrary function of its argument $(\bullet)$, and $\overline{k} = k_x \hat{x} + k_y \hat{y} + k_z \hat{z}$; this is analogous to the wave solution (9.2.4) using the notation (9.2.5). Substituting the solution (13.1.10) into the wave equation yields:

$$\left(\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2\right)p(\omega t - \overline{k} \bullet \overline{r}) - \left(\rho_o/\gamma P_o\right)\partial^2 p(\omega t - \overline{k} \bullet \overline{r})/\partial t^2 = 0 \qquad (13.1.11)$$

$$-\left(k_x^2 + k_y^2 + k_z^2\right)p''(\omega t - \overline{k} \bullet \overline{r}) - \left(\rho_o/\gamma P_o\right)\omega^2 p''(\omega t - \overline{k} \bullet \overline{r}) = 0 \qquad\qquad (13.1.12)$$

$$k_x^2 + k_y^2 + k_z^2 = k^2 = \omega^2 \rho_o/\gamma P_o = \omega^2/v_p^2 \qquad\qquad\qquad (13.1.13)$$

This is analogous to the electromagnetic dispersion relation (9.2.8).

As in the case of electromagnetic waves [see (9.5.19) and (9.5.20)], the *acoustic phase velocity* v_p and *acoustic group velocity* v_g are simply related to k:

$$v_p = \omega/k = \left(\gamma P_o/\rho_o\right)^{0.5} = c_s \qquad\qquad \text{(acoustic phase velocity)} \quad (13.1.14)$$

$$v_g = \left(\partial k/\partial \omega\right)^{-1} = \left(\gamma P_o/\rho_o\right)^{0.5} = c_s \qquad\qquad \text{(acoustic group velocity)} \quad (13.1.15)$$

Adiabatic acoustic waves propagating in 0^oC air near sea level experience $\gamma = 1.4$, $\rho_o = 1.29$ [kg/m^3], and $P_o = 1.01 \times 10^5$ [N/m^2], yielding $c_s \cong 330$ [m/s].

In solids or liquids the constitutive relation is:

$$\partial\rho/\partial p = \rho/K \qquad \text{(constitutive relation for solids and liquids)} \quad (13.1.16)$$

K [N m^{-2}] is the *bulk modulus* of the medium. The coefficient 1/K then replaces $1/\gamma P_o$ in (13.1.8–10), yielding the *acoustic velocity in solids* and liquids:

$$c_s = \left(K/\rho_o\right)^{0.5} \quad \left[\text{m s}^{-1}\right] \qquad \text{(acoustic velocity in solids and liquids)} \quad (13.1.17)$$

Typical acoustic velocities are 900 - 2000 m s^{-1} in liquids (~1500 m s^{-1} in water), and 1500–13,000 m s^{-1} in solids (~5900 m s^{-1} in steel).

Analogous to (7.1.25) and (7.1.26), the acoustic differential equations (13.1.8) and (13.1.7) can be simplified for sinusoidal plane waves propagating along the z axis:

$$\nabla \underline{p} \bullet \hat{z} = \frac{d\underline{p}(z)}{dz} = -j\omega\rho_o \underline{u}_z(z) \qquad (13.1.18)$$

$$\nabla \bullet \overline{\underline{u}} = \frac{d\underline{u}_z(z)}{dz} = \frac{-j\omega}{\gamma P_o}\underline{p}(z) \qquad (13.1.19)$$

These can be combined to yield the wave equation for z-axis waves analogous to (7.1.27):

$$\frac{d^2\underline{p}(z)}{dz^2} = -\omega^2 \frac{\rho_o}{\gamma P_o}\underline{p}(z) \qquad (13.1.20)$$

Analogous to (7.1.28) and (7.1.29), the solution is a sum of exponentials of the form:

$$\underline{p}(z) = \underline{p}_+ e^{-jkz} + \underline{p}_- e^{+jkz} \quad \left[\text{N m}^{-2}\right] \qquad (13.1.21)$$

$$\underline{u}_z(z) = -\frac{1}{j\omega\rho_o}\frac{d\underline{p}(z)}{dz} = \frac{k}{\omega\rho_o}\left[\underline{p}_+ e^{-jkz} - \underline{p}_- e^{+jkz}\right] \;[\text{m/s}] \qquad (13.1.22)$$

Note that, unlike electromagnetic waves, where the key fields are vectors transverse to the direction of propagation, the velocity vector for acoustic waves is in the direction of propagation and pressure is a scalar.

Analogous to (7.1.31), the characteristic *acoustic impedance* of a gas is:

$$\eta_s = \frac{\underline{p}(z)}{\underline{u}_z(z)} = \frac{\omega\rho_o}{k} = \rho_o c_s = \sqrt{\gamma\rho_o P_o} \quad \left[\text{N s/m}^3\right] \qquad (13.1.23)$$

The acoustic impedance of air at room temperature is ~425 [N s m^{-3}]. The acoustic impedance for solids and liquids is $\eta_s = \rho_o c_s = (\rho_o K)^{0.5}$ [N s m^{-3}]. Note that the units are not ohms.

The instantaneous acoustic intensity [W m^{-2}] of this plane wave is $p(t)u_z(t)$, the complex power is $\underline{p}\,\overline{\underline{u}}*/2$, and the time average acoustic power is $R_e\{\underline{p}\,\overline{\underline{u}}*/2\}$ [W m^{-2}], analogous to (2.7.41).

We can derive an *acoustic power conservation* law similar to the Poynting theorem (2.7.22) by computing the divergence of $\underline{p}\ \overline{\underline{u}}*$ [W m^{-2}] and substituting in (13.1.18) and (13.1.19):[74]

$$\nabla \bullet (\underline{p}\,\overline{\underline{u}}*) = \overline{\underline{u}}* \bullet \nabla \underline{p} + \underline{p} \nabla \bullet \overline{\underline{u}}* = \overline{\underline{u}}* \bullet (-j\omega \rho_o \overline{\underline{u}}) + j\omega \underline{p}\underline{p}*/\gamma P_o \qquad (13.1.24)$$

$$= -4j\omega\left(\left[\rho_o |\overline{\underline{u}}|^2/4\right] - \left[|\underline{p}|^2/4\gamma P_o\right]\right) = -4j\omega(\langle W_k \rangle - \langle W_p \rangle) \qquad (13.1.25)$$

The time average *acoustic kinetic energy density* of the wave is W_k [J m^{-3}] $= \rho_o |\overline{\underline{u}}|^2/4$, and the time average *acoustic potential energy density* is $W_p = |\underline{p}|^2/4\gamma P_o$. For liquids or solids $\gamma P_o \to K$, so $W_p = |\underline{p}|^2/4K$. If there is no divergence of acoustic radiated power $\underline{p}\,\overline{\underline{u}}*$, then it follows from (13.1.25) that:

$$\langle W_k \rangle = \langle W_p \rangle \qquad \text{(energy balance in a lossless resonator)} \quad (13.1.26)$$

The *acoustic intensity* I [W m^{-2}] of an acoustic plane wave, analogous to (2.7.41), is:

$$I = R_e\{\underline{p}\underline{u}*/2\} = |\underline{p}|^2/2\eta_s = \eta_s |\underline{u}|^2/2 \ \ [\text{W m}^{-2}] \qquad \text{(acoustic intensity)} \quad (13.1.27)$$

where the acoustic impedance $\eta_s = \rho_o c_s$. The instantaneous acoustic intensity is $p(t)u_z(t)$, as noted above.

Example 13.1A
A loud radio radiates 100 acoustic watts at 1 kHz from a speaker 10-cm square near sea level where ρ_o = 1.29 [kg m^{-3}] and $c_s \cong 330$ m s^{-1}. What are the: 1) wavelength, 2) peak pressure, particle velocity, and displacement, and 3) average energy density of this uniform acoustic plane wave in the speaker aperture?

Solution: λ = c_s/f = 330/1000 = 33 cm. (13.1.22) yields $|\underline{u}|$ = $(2I/\eta_s)^{0.5}$, and (13.1.18) says $\eta_s = \rho_o c_s$, so $|\underline{u}|$ = $[200/(1.29\times330)]^{0.5}$ = 0.69 [m s^{-1}]. $\underline{p}$ = $\eta_s\underline{u}$ = 425.7×0.69 = 292 [N m^{-2}]. Note that this acoustic pressure is much less than the ambient pressure $P_o \cong 10^5$ N m^{-2}, as required for linearization of the acoustic equations. Displacement

---

[74] Although these two equations apply to waves propagating in the z direction, their right-hand sides also apply to any direction if the subscript z is omitted.

$\underline{d}$ is the integral of velocity $\underline{u}$, so $\underline{d} = \underline{u}/j\omega$ and the peak-to-peak displacement is $2|\underline{u}|/\omega$ = $2 \times 0.69/2\pi 1000 = 0.22$ mm. The average acoustic energy density stored equals $2\langle W_k \rangle = 2\rho_o |\underline{u}|^2 / 4 = 1.29(0.69)^2 / 2 = 0.31$ [J m^{-3}].

13.2   Acoustic waves at interfaces and in guiding structures and resonators

13.2.1   Boundary conditions and waves at interfaces

The behavior of acoustic waves at boundaries is determined by the acoustic boundary conditions. At rigid walls the normal component of acoustic velocity must clearly be zero, and fluid pressure is unconstrained there. At boundaries between two fluids or gases in equilibrium, both the acoustic pressure $p(\bar{r},t)$ and the normal component of acoustic velocity $\bar{u}_\perp(\bar{r},t)$ must be continuous. If the pressure were discontinuous, then a finite force normal to the interface would be acting on infinitesimal mass, giving it infinite acceleration, which is not possible. If $\bar{u}_\perp$ were discontinuous, then $\partial p/\partial t$ at the interface would be infinite, which also is not possible; (13.1.8) says $\nabla \bullet \bar{u} = -(1/\gamma P_o)\partial p/\partial t$. These *acoustic boundary conditions* at a boundary between media 1 and 2 can be stated as:

$$p_1 = p_2 \qquad \text{(boundary condition for pressure)} \qquad (13.2.1)$$

$$\bar{u}_{1\perp} = \bar{u}_{2\perp} \qquad \text{(boundary condition for velocity)} \qquad (13.2.2)$$

A uniform acoustic plane wave incident upon a planar boundary between two media having different acoustic properties will generally have a transmitted component and a reflected component, as suggested in Figure 13.2.1. The angles of incidence, reflection, and transmission are θ_i, θ_r, and θ_t, respectively.
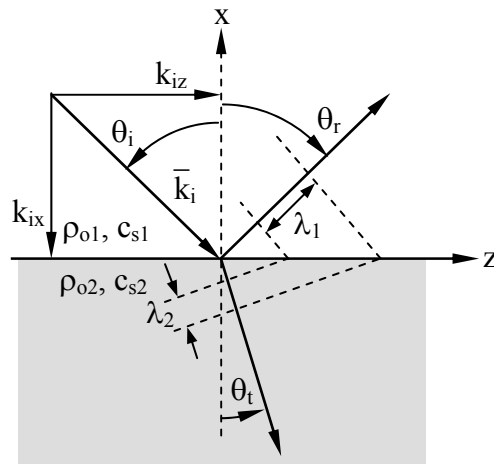
Figure 13.2.1   Acoustic waves at a planar interface with phase matching.

A typical example is the boundary between cold air overlying a lake and warm air above; the warm air is less dense, although the pressures across the boundary must balance. Since $c_s \propto \rho_o^{-0.5}$ and $\eta_s \propto \rho_o^{0.5}$, both the amplitudes and angles of propagation must change at the density discontinuity.

As was done for electromagnetic waves (see Section 9.2.2), we can begin with tentative general expressions for the incident, reflected, and transmitted plane waves:

$$\underline{p}_i(\overline{r}) = \underline{p}_{io}e^{+jk_{ix}x - jk_{iz}z} \qquad \text{(incident wave)} \quad (13.2.3)$$

$$\underline{p}_r(\overline{r}) = \underline{p}_{ro}e^{+jk_{rx}x - jk_{rz}z} \qquad \text{(reflected wave)} \quad (13.2.4)$$

$$\underline{p}_t(\overline{r}) = \underline{p}_{to}e^{+jk_{tx}x - jk_{tz}z} \qquad \text{(transmitted wave)} \quad (13.2.5)$$

At $x = 0$ the pressure is continuous across the boundary (13.2.1), so $\underline{p}_i(\overline{r}) + \underline{p}_r(\overline{r}) = \underline{p}_t(\overline{r})$, which requires the phases (-jkz) to match:

$$k_{iz} = k_{rz} = k_{tz} \equiv k_z \qquad (13.2.6)$$

But k_z is the projection of the $\overline{k}$ on the z axis, so $k_{iz} = k_i \sin\theta_i$, where $k_i = \omega/c_{si}$, and:

$$k_i \sin\theta_i = k_r \sin\theta_r = k_t \sin\theta_t \qquad (13.2.7)$$

$$\theta_i = \theta_r \qquad (13.2.8)$$

$$\sin\theta_t / \sin\theta_i = c_{si}/c_{st} \qquad \text{(\textit{acoustic Snell's law})} \quad (13.2.9)$$

Thus acoustic waves refract at boundaries like electromagnetic waves (9.2.26).

Acoustic waves can also be evanescent for $\theta_i > \theta_c$, where the critical angle θ_c is the angle of incidence (9.2.30) required by Snell's law when $\theta_t = 90°$:

$$\theta_c = \sin^{-1}(c_{si}/c_{st}) \qquad \text{(\textit{acoustic critical angle})} \quad (13.2.10)$$

When $\theta_i > \theta_c$, then k_{tx} becomes imaginary, analogous to (9.2.32), the transmitted acoustic wave is evanescent, and there is total reflection of the incident acoustic wave. Thus:

$$k_{tx} = \pm j(k_t^2 - k_z^2)^{0.5} = \pm j\alpha \qquad (13.2.11)$$

$$\underline{p}_t(x,z) = \underline{p}_{to}e^{-\alpha x - jk_z z} \qquad (13.2.12)$$

It follows from the complex version of (13.1.7) that:

$$\overline{\underline{u}}_t = -\nabla \underline{p}_t \big/ j\omega\rho_o = \left(\alpha\hat{x} + jk_z\hat{z} \right) \underline{p}_t \big/ j\omega\rho_o \tag{13.2.13}$$

The complex power flow in this *acoustic evanescent wave* is $\underline{p}\underline{u}^*$, analogous to (9.2.35), so the power flowing in the -x direction is imaginary and the time-average real power flow is:

$$R_e\left\{\underline{p}\overline{\underline{u}}^*\right\}\big/2 = \hat{z}\left(k_z/2\omega\rho_o \right) \left| \underline{p}_{to} \right|^2 e^{-2\alpha z} \ \left[\text{W m}^{-2} \right] \tag{13.2.14}$$

The fraction of power reflected from an acoustic boundary can be found by applying the boundary conditions and solving for the unknown reflected amplitude. If we define $\underline{p}_{ro}$ and $\underline{p}_{to}$ as $\underline{\Gamma}\underline{p}_{io}$ and $\underline{T}\underline{p}_{io}$, respectively, then matching boundary conditions at $x = z = 0$ yields:

$$p_{io} + p_{ro} = p_{to} \Rightarrow 1 + \underline{\Gamma} = \underline{T} \tag{13.2.15}$$

We need an additional boundary condition, and may combine $\overline{\underline{u}} = -\nabla\underline{p}\big/j\omega\rho_o$ (13.1.7) with the expression for $\underline{p}$ (13.2.3) to yield:

$$\overline{\underline{u}}_i = \left[\left(-jk_{xi}\hat{x} + jk_z\hat{z} \right)\big/j\omega\rho_{oi} \right] \underline{p}_{io} e^{+jk_{ix}x - jk_{iz}z} \tag{13.2.16}$$

Similar expressions for $\overline{\underline{u}}_r$ and $\overline{\underline{u}}_t$ can be found, and enforcing continuity of $\overline{\underline{u}}_\perp$ across the boundary at $x = z = 0$ yields:

$$\frac{k_{xi}}{\omega\rho_{oi}} - \underline{\Gamma}\frac{k_{xi}}{\omega\rho_{oi}} = \frac{k_{xt}}{\omega\rho_{ot}}\underline{T} \tag{13.2.17}$$

$$1 - \underline{\Gamma} = \underline{T}\frac{k_{xt}\rho_{oi}}{k_{xi}\rho_{ot}} = \underline{T}\frac{\eta_i\cos\theta_t}{\eta_t\cos\theta_i} \equiv \frac{\underline{T}}{\eta_n} \tag{13.2.18}$$

where we define the normalized angle-dependent acoustic impedance $\eta_n \equiv (\eta_t\cos\theta_i/\eta_i\cos\theta_t)$ and we recall $k_{xt} = k_t\cos\theta_t$, $k_t = \omega/c_{st}$, and $\eta_t = c_{st}\rho_{ot}$. Combining (13.2.15) and (13.2.18) yields:

$$\underline{\Gamma} = \frac{\eta_n - 1}{\eta_n + 1} \tag{13.2.19}$$

$$\underline{T} = 1 + \underline{\Gamma} = \frac{2\eta_n}{\eta_n + 1} \tag{13.2.20}$$

These expressions for $\underline{\Gamma}$ and $\underline{T}$ are essentially the same as for electromagnetic waves, (7.2.31) and (7.2.32), although the expressions for η_n are different. The fraction of acoustic power reflected is $|\underline{\Gamma}|^2$. Acoustic impedance $\underline{\eta}(z)$ for waves propagating perpendicular to boundaries therefore also are governed by (7.2.24):

$$\underline{\eta}(z) = \eta_o \frac{\eta_L - j\eta_o \tan kz}{\eta_o - j\underline{\eta}_L \tan kz} \tag{13.2.21}$$

The Smith chart method of Section 7.3.2 can also be used.

There can even be an *acoustic Brewster's angle* θ_B when $\underline{\Gamma} = 0$, analogous to (9.2.75). Equation (13.2.19) suggests this happens when $\eta_n = 1$ or, from (13.2.18), when $\eta_i \cos\theta_t = \eta_t \cos\theta_B$. After some manipulation it can be shown that Brewster's angle is:

$$\theta_B = \tan^{-1} \sqrt{\frac{(\eta_t/\eta_i)^2 - 1}{1 - (c_{s_t}/c_{s_i})^2}} \tag{13.2.22}$$

Example 13.2A
A typical door used to block out sounds might be 3 cm thick and have a density of 1000 kg m^{-3}, large compared to 1.29 kg m^{-3} for air. If $c_s = 330$ m s^{-1} in air and 1000 m s^{-1} in the door, what are their respective acoustic impedances, η_a and η_d? What fraction of 500-Hz normally incident acoustic power would be reflected by the door? The fact that the door is not gaseous is irrelevant here if it is free to move and not secured to its door jamb.

Solution: The acoustic impedance $\eta = \rho_o c_s = 425.7$ in air and 10^6 in the door (13.1.23). The impedance at the front surface of the door given by (13.2.21) is $\eta_{fd} = \eta_d(\eta_a - j\eta_d \tan kz)/(\eta_d - j\eta_a \tan kz)$, where $k = 2\pi/\lambda_d$ and $z = 0.03$. $\lambda_d = c_d/f = 1000/500 = 2$, so $kz = \pi z = 0.094$, and $\tan kz = 0.095$. Thus $\eta_{fd} = 425.7 + 3.84$ and $\eta_{fd}/\eta_a = 1.0090$. Using (13.2.19) the reflected power fraction $= |\Gamma|^2 = |(\eta_n - 1)/(\eta_n + 1)|^2$ where $\eta_n = \eta_{fd}/\eta_a = 1.0090$, we find $|\Gamma|^2 \cong 2\times 10^{-5}$. Virtually all acoustic power passes through. If this solid door were secure in its frame, shear forces (neglected here) would lead to far better acoustic isolation.

13.2.2   Acoustic plane-wave transmission lines

Acoustic plane waves guided within tubes of constant cross-section satisfy the boundary conditions posed by stiff walls: 1) $u_\perp = 0$, and 2) any $u_{//}$ and p is permitted. If these tubes curve slowly relative to a wavelength then their plane-wave behavior is preserved. The viscosity of gases is sufficiently low that frictional losses at the wall can usually be neglected in small acoustic devices. The resulting waves are governed by the acoustic wave equation (13.1.20), which has the solutions for p, u_z, and η given by (13.1.21), (13.1.22), and (13.1.23), respectively. Wave intensity is governed by (13.1.26), the complex reflection coefficient $\underline{\Gamma}$ is given by

(13.2.19), and impedance transformations are governed by (13.2.21). This set of equations is adequate to solve most acoustic transmission line problems in single tubes once we model their terminations.

Two acoustic terminations for tubes are easily treated: closed ends and open ends. The boundary condition posed at the closed end of an acoustic pipe is simply that u = 0. At an open end the pressure is sufficiently released that p ≅ 0 there. If we intuitively relate acoustic velocity u(t,z) to current i(z,t) in a TEM line, and p(z,t) to voltage v(t,z), then a closed pipe is analogous to an open circuit, and an open pipe is analogous to a short circuit (the reverse of what we might expect).[75] Standing waves exist in either case, with $\lambda/2$ separations between pressure nulls or between velocity nulls.

13.2.3 Acoustic waveguides

Acoustic waveguides are pipes that convey sound in one or more waveguide modes. Section 13.2.2 considered only the special case where the waves were uniform and the acoustic velocity $\overline{u}$ was confined to the ±z direction. More generally the wave pressure and velocity must satisfy the acoustic wave equation, analogous to (2.3.21):

$$\left(\nabla^2 + \omega^2/c_s^2\right)\left\{\frac{\underline{p}}{\underline{u}}\right\} = 0 \tag{13.2.23}$$

Solutions to (13.2.23) in cartesian coordinates are appropriate for rectangular waveguides, as discussed in Section 9.3.2. Assume that two of the walls are at x = 0 and y = 0. Then a wave propagating in the +z direction might have the general form:

$$\underline{p}(x,y,z) = \underline{p}_o \left\{\begin{matrix} \sin k_x x \\ \cos k_x x \end{matrix}\right\}\left\{\begin{matrix} \sin k_y y \\ \cos k_y y \end{matrix}\right\} e^{-jk_z z} \tag{13.2.24}$$

The choice between sine and cosine is dictated by boundary conditions on $\overline{u}$, which can be found using $\overline{u} = -\nabla \underline{p}/j\omega\rho_o$ (13.2.13). Since the velocity $\overline{u}$ perpendicular to the waveguide walls at x = 0 and y = 0 must be zero, so must be the gradient ∇p in the same perpendicular x and y directions at the walls. Only the cosine factors in (13.2.24) have this property, so the sine factors must be zero, yielding:

$$\underline{p} = \underline{p}_o \cos k_x x \cos k_y y \, e^{-jk_z z} \tag{13.2.25}$$

$$\overline{H} = \left[\hat{x} \; k_z \{\sin k_x x \; or \; \cos k_x x\} \right. $$
$$\left. -\hat{y}\left(jk_y/k_o\right)\cos k_x x \; \sin k_y y + \hat{z}\left(k_z/k_o\right)\cos k_x x \; \cos k_y y\right] e^{-jk_z z} \tag{13.2.26}$$

---

[75] Although methods directly analogous to TEM transmission lines can also be used to analyze tubes of different cross-sections joined at junctions, the subtleties place this topic outside the scope of this text.

Since $\bar{\underline{u}}_\perp$ (i.e. $\underline{u}_x$ and $\underline{u}_y$) must also be zero at the walls located at $x = a$ and $y = b$, it follows that $k_x a = m\pi$, and $k_y b = n\pi$, where m and n are integers: 0,1,2,3,...   Substitution of any of these solutions (13.2.25) into the wave equation (13.1.9) yields:

$$k_x^2 + k_y^2 + k_z^2 = (m\pi/a)^2 + (n\pi/b)^2 + (2\pi/\lambda_z)^2 = k_s^2 = \omega^2 \rho_o / \gamma P_o = \omega^2 / c_s^2 = (2\pi/\lambda_s)^2 \quad (13.2.27)$$

$$k_{z_{mn}} = \sqrt{k_s^2 - k_x^2 - k_y^2} = \sqrt{\left(\frac{\omega}{c_s}\right)^2 - \left(\frac{m\pi}{a}\right)^2 - \left(\frac{n\pi}{b}\right)^2} \quad \rightarrow \quad \pm j\alpha \text{ at } \omega_n \qquad (13.2.28)$$

Therefore each acoustic mode A_{mn} has its own cutoff frequency ω_{mn} where k_z becomes imaginary. Thus each mode becomes evanescent for frequencies below its cutoff frequency f_{mn}, analogous to (9.3.22), where:

$$f_{mn} = \omega_{mn}/2\pi = \left[(c_s m/2a)^2 + (c_s n/2b)^2\right]^{0.5} \quad [Hz] \quad \text{(cutoff frequency)} \quad (13.2.29)$$

$$\lambda_{mn} = c_s/f_{mn} = \left[(m/2a)^2 + (n/2b)^2\right]^{-0.5} \quad [m] \qquad \text{(cutoff wavelength)} \quad (13.2.30)$$

Below the cutoff frequency f_{mn} for each acoustic mode the *evanescent acoustic mode* propagates as $e^{-jk_z z} = e^{-\alpha z}$, analogous to (9.3.31), where the wave decay rate is:

$$\alpha = \left[(m\pi/a)^2 + (n\pi/b)^2 - (\omega_{mn}/c_s)^2\right]^{0.5} \qquad (13.2.31)$$

The total wave in any acoustic waveguide is that superposition of separate modes which matches the given boundary conditions and sources, where one (A_{00}) or more modes always propagate and an infinite number ($m\rightarrow\infty$, $n\rightarrow\infty$) are evanescent and reactive. The expression for $\underline{p}$ follows from (13.2.25) where $e^{-jk_z z} \rightarrow e^{-\alpha z}$, and the expression for $\bar{\underline{u}}$ follows from $\bar{\underline{u}} = -\nabla\underline{p}/(j\omega\rho_o)$ (13.2.13).

13.2.4   Acoustic resonators

Any closed container trapping acoustic energy exhibits resonances just as do low-loss containers of electromagnetic radiation. We may consider a rectangular room, or perhaps a smaller box, as a rectangular acoustic waveguide terminated at its ends by walls (velocity nulls for u_z). The acoustic waves inside must obey (13.2.27):

$$k_x^2 + k_y^2 + k_z^2 = (m\pi/a)^2 + (n\pi/b)^2 + (q\pi/d)^2 = \omega^2/c_s^2 \qquad (13.2.32)$$

where $k_z = 2\pi/\lambda_z$ has been replaced by $k_z = q\pi/d$ using the constraint that if the box is short- or open-circuited at both ends then its length d must be an integral number q of half-wavelengths $\lambda_z/2$; therefore $d = q\lambda_z/2$ and $2\pi/\lambda_z = q\pi/d$. Thus, analogous to (9.4.3), the *acoustic resonant frequencies* of a closed box of dimensions a,b,d are:

$$f_{mnq} = c_s\left[(m/2a)^2 + (n/2b)^2 + (q/2d)^2\right]^{0.5} \; [Hz] \; \text{(resonant frequencies)} \quad (13.2.33)$$

A simple geometric construction yields the mode density (modes/Hz) for both acoustic and electromagnetic rectangular acoustic resonators of volume V = abd, as suggested in Figure 13.2.2.
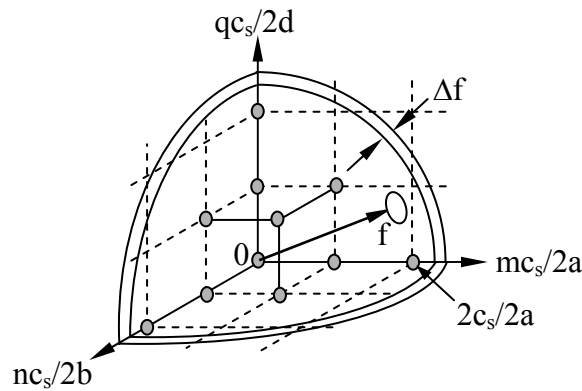
Figure 13.2.2   Resonant modes of a rectangular cavity.

Each resonant mode A_{mnq} corresponds to one set of quantum numbers m,n,q and to one cell in the figure. Referring to (13.2.11) it can be seen that frequency in the figure corresponds to the length of a vector from the origin to the mode A_{mnq} of interest. The total number N_o of acoustic modes with resonances at frequencies less than f_o is approximately the volume of the eighth-sphere shown in the figure, divided by the volume of each cell of dimension $(c_s/2a) \times (c_s/2b) \times (c_s/2d)$, where each cell corresponds to one acoustic mode. The approximation improves as f increases. Thus:

$$N_o \cong \left[4\pi f_o^3/(3 \times 8)\right]/\left(c_s^3/8abd\right) = 4\pi f_o^3 V/3c_s^3 \; [\text{modes} < f_o] \quad (13.2.34)$$

In the electromagnetic case each set of quantum numbers m,n,q corresponds to both a TE and a TM resonant mode of a rectangular cavity, so N_o is then doubled:

$$^3 N_o \cong 8\pi f_o \; V/3C^2 \qquad \text{(electromagnetic modes} < f_o) \quad (13.2.35)$$

The number density n_o of acoustic modes per Hertz is the volume of a thin shell of thickness Δf, again divided by the volume of each cell:

$$n_o \cong \Delta f \times 4\pi f^2 \Big/ \left(8c_s^3 \Big/ 8abd\right) = \Delta f \, 4\pi f^2 V \Big/ c_s^3 \quad \left[\text{modes Hz}^{-1}\right] \qquad (13.2.36)$$

Thus the modes of a resonator overlap more and tend to blend together as the frequency increases. The density of electromagnetic modes in a similar cavity is again twice that for acoustic modes.

Typical examples of acoustic resonators include musical instruments such as horns, woodwinds, organ pipes, and the human vocal tract. Rooms with reflective walls are another example. In each case if we wish to excite a particular mode efficiently the source must not only excite it with the desired frequency, but also from a favorable location.

One way to identify favorable locations for modal excitation is to assume the acoustic source exerts pressure p across a small aperture at the wall or interior of the resonator, and then to compute the incremental acoustic intensity transferred from that source to the resonator using (13.1.27):

$$I = R_e \left\{pu^* \Big/ 2\right\} \qquad (13.2.37)$$

In this expression we assume $\bar{u}$ is dominated by waves already present in the resonator at the resonant frequency of interest and that the vector $\bar{u}$ is normal to the surface across which p is applied. Therefore pressure sources located at velocity nulls for a particular mode transfer no power and no excitation occurs. Conversely, power transfer is maximized if pressure is applied at velocity maxima. Similarly, acoustic velocity sources are best located at a pressure maximum of a desired mode. For example, all acoustic modes have pressure maxima at the corners of rectangular rooms, so velocity loudspeakers located there excite all modes equally.

The converse is also true. If we wish to damp certain acoustic modes we may put absorber at their velocity or pressure maxima, depending on the type of absorber used. A wire mesh that introduces drag damps high velocities, and surfaces that reflect waves weakly (such as holes in pipes) damp pressure maxima. High frequency modes are more strongly damped in humid atmospheres than are low frequency modes, but such bulk absorption mechanisms do not otherwise discriminate among them.

Because the pressure and velocity maxima are located differently for each mode, each mode typically has a different Q, which is the number of radians before the total stored energy w_T decays by a factor of e^{-1}. Therefore the Q of any particular mode m,n,p is (7.4.34):

$$Q = \omega_o w_T / P_d \qquad \text{(acoustic Q)} \quad (13.2.38)$$

The resonant frequencies and stored energies are given by (13.2.33) and (13.1.25), respectively, where it suffices to compute either the maximum stored kinetic or potential energy, for they are equal. The power dissipated P_d can be found by integrating the intensity expression (13.2.37) over the soft walls of the resonator, and adding any dissipation occurring in the interior.

Small changes in resonator shape can perturb acoustic resonant frequencies, much like electromagnetic resonances are perturbed. Whether a gentle indentation increases or lowers a particular resonant frequency depends on whether the time average acoustic pressure for the mode of interest is positive or negative at that indentation. It is useful to note that acoustic energy is quantized, where each *phonon* has energy hf Joules where h is Planck's constant; this is directly analogous to a photon at frequency f. Therefore the total acoustic energy in a resonator at frequency f is:

$$w_T = nhf \ [J] \tag{13.2.39}$$

If the cavity shape changes slowly relative to the frequency, the number n of acoustic phonons remains constant and any change in w_T results in a corresponding change in f. The work Δw_T done on the phonon field when cavity walls move inward Δz is positive if the time average acoustic pressure P_a is outward (positive), and negative if that pressure is inward or negative: $\Delta w_T = P_a \Delta z$. It is well known that gaseous flow parallel to a surface pulls on that surface as a result of the Bernoulli effect, which is the same effect that explains how airplane wings are supported in flight and how aspirators work. Therefore if an acoustic resonator is gently indented at a velocity maximum for a particular resonance, that resonant frequency f will be reduced slightly because the phonon field pulling the wall inward will have done work on the wall. All acoustic velocities at walls must be parallel to them. Conversely, if the indentation occurs near pressure maxima for a set of modes, the net acoustic force is outward and therefore the indentation does work on the phonon field, increasing the energy and frequency of those modes.

The most pervasive example of this phenomenon is human speech, which employs a vocal tract perhaps 16 cm long, typically less in women and all children. One end is excited by brief pulses in air pressure produced as the vocal chords vibrate at the pitch frequency of any vowel being uttered. The resulting train of periodic pressure pulses with period T has a frequency spectrum consisting of impulses spaced at T^{-1} Hz, typically below 500 Hz. The vocal tract then accentuates those impulses falling near any resonance of that tract.
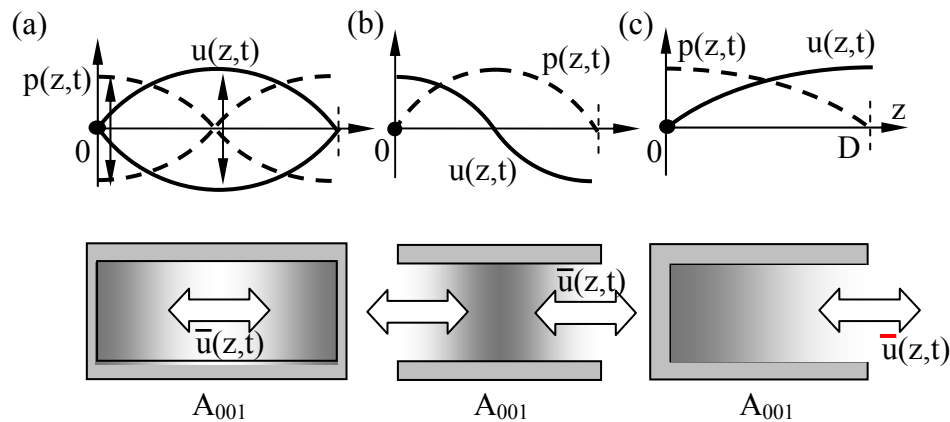
Figure 13.2.3 Acoustic resonances in tubes.

Figure 13.2.3 illustrates the lowest frequency acoustic resonances possible in pipes that are: (a) closed at both ends, (b) open at both ends, and (c) closed at one end and open at the other; each mode is designated A_{001} for its own structure, where 00 corresponds to the fact that the acoustic wave is uniform in the x-y plane, and 1 indicates that it is the lowest non-zero-frequency resonant mode. Resonator (a) is capable of storing energy at zero frequency by pressurization (in the A_{000} mode), and resonator (b) could store energy in the A_{000} mode if there were a steady velocity in one direction through the structure; these A_{000} modes are generally of no interest, and some experts do not consider them modes at all.

A sketch of the human vocal tract appears in Figure 13.2.4(a); at resonance it is generally open at the mouth and closed at the vocal chords, analogous to the resonator pictured in Figure 13.2.3(c). This structure resonates when its length D corresponds to one-quarter wavelength, three-quarters wavelength, or generally (2n-1)/4 wavelengths for the A_{00n} mode, as sketched in Figure 13.2.3(b). For a vocal tract 16 cm long and a velocity of sound $c_s = 340$ m s^{-1}, the lowest resonant frequency $f_{001} = c_s/\lambda_{001} = 340/(4 \times 0.16) = 531$ Hz. The next resonances, f_2 and f_3, fall at 1594 and 2656 Hz, respectively.
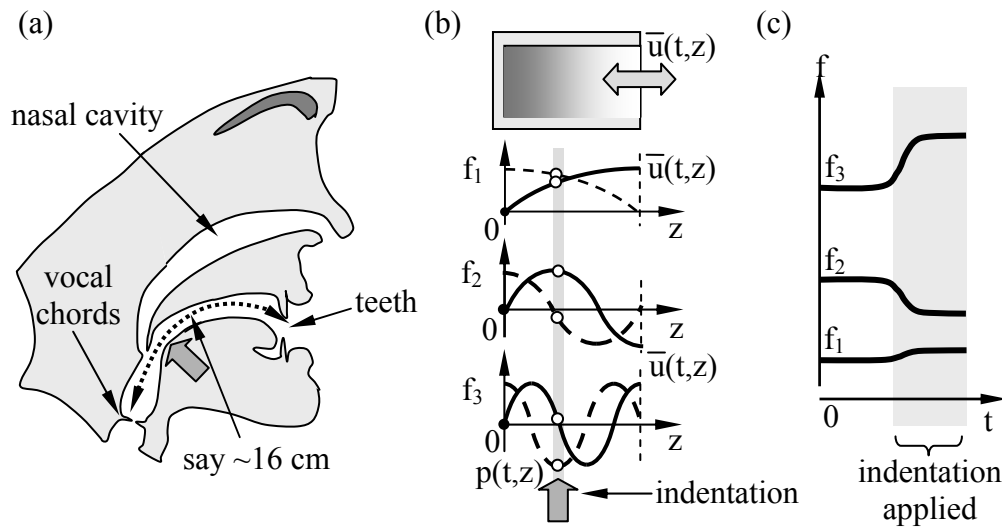
Figure 13.2.4   Human vocal tract.

If the tongue now indents the vocal tract at the arrow indicated in (a) and (b) of Figure 13.2.4, then the three illustrated resonances will all shift as indicated in (c) of the figure. The resonance f_1 shifts only slightly upward because the indentation occurs between the peaks for velocity and pressure, but nearer to the pressure peak. The resonances shift more significantly down for f_2 and up for f_3 because this indentation occurs near velocity and pressure maximum for these two resonances, respectively, while occurring near a null for the complimentary variable. By simply controlling the width of the vocal tract at various positions using the tongue and teeth, these tract resonances can be modulated up and down to produce our full range of vowels.

These resonances are driven by periodic impulses of air released by the vocal cords at a pitch controlled by the speaker. The pitch is a fraction of the lowest vocal tract resonant frequency, and the impulses are sufficiently brief that their harmonics range up to 5 kHz and

more. Speech also includes high-pitched broadband noise caused by turbulent air whistling past the teeth or other obstacles as in the consonants s, h, and f, and impulsive spikes caused by temporary tract closures, as in the consonants b, d, and p. Speech therefore includes both voiced (driven by vocal chord impulses) and unvoiced components. The spectral content of most consonants can similarly be modulated by the vocal tract and mouth.

It is possible to change the composition of the air in the vocal tract, thus altering the velocity of sound c_s and the resonant frequencies of the tract, which are proportional to c_s (13.2.30). Thus when breathing helium all tract resonance frequencies increase by a noticeable fraction, equivalent to shortening the vocal tract. Note that pitch is not significantly altered by helium because the natural pitch of the vocal chords is determined instead primarily by their tension, composition, and length.

13.3   Acoustic radiation and antennas

Any mechanically vibrating surface can radiate acoustic waves. As in the case of electromagnetic waves, it is easiest to understand a point source first, and then to superimpose such radiators in combinations that yield the total desired radiation pattern. Reciprocity applies to linear acoustics, so the receiving and transmitting properties of acoustic antennas are proportional, as they are for electromagnetic waves; i.e. $G(\theta,\phi) \propto A(\theta,\phi)$.

The acoustic wave equation for pressure permits analysis of an *acoustic monopole* radiator:

$$\left[\nabla^2 + (\omega/c_s)^2\right]\underline{p} = 0 \tag{13.3.1}$$

If the acoustic radiator is simply an isolated sphere with a sinusoidally oscillating radius $\underline{a}$, then the source is spherically symmetric and so is the solution; thus $\partial/\partial\theta = \partial/\partial\phi = 0$. If we define $\omega/c_s = k$, then (13.3.1) becomes:

$$\left[r^{-2}d(r^2 d/dr) + k^2\right]\underline{p} = \left[d^2/dr^2 + 2r^{-1}d/dr + k^2\right]\underline{p} = 0 \tag{13.3.2}$$

This can be rewritten more simply as:

$$d^2(r\underline{p})/dr^2 + k^2(r\underline{p}) = 0 \tag{13.3.3}$$

This equation is satisfied if $r\underline{p}$ is an exponential, so a radial acoustic wave propagating outward would have the form:

$$\underline{p}(r) = \underline{K}r^{-1}e^{-jkr} \quad \left[N\ m^{-2}\right] \tag{13.3.4}$$

The associated acoustic velocity $\underline{u}(r)$ follows from the complex form of Newton's law (13.1.7): $\nabla\underline{p} \cong -j\omega\rho_o\underline{u}$ [N m^{-3}]:

$$\bar{u}(r) = -\nabla \underline{p}/j\omega r_o = \hat{r}\underline{K}(\eta_s r)^{-1}\left[1+(jkr)^{-1}\right]e^{-jkr} \qquad (13.3.5)$$

The first and second terms in the solution (13.3.5) correspond to the acoustic far field and *acoustic near field*, respectively. When kr >> 1 or, equivalently, r >> $\lambda/2\pi$, then the near field term can be neglected, so that the far-field velocity corresponding to (13.3.4) is:

$$\bar{u}_{ff}(r) = \hat{r}\underline{K}(\eta_s r)^{-1}e^{-jkr} \; \left[m\;s^{-1}\right] \qquad \text{(far-field acoustic velocity)} \quad (13.3.6)$$

The near-field velocity from (13.3.5) is:

$$\bar{u}_{nf} = -j\underline{K}\hat{r}\left(k\eta_s r^2\right)^{-1}e^{-jkr} \; \left[m\;s^{-1}\right] \qquad \text{(near-field acoustic velocity)} \quad (13.3.7)$$

Since k = ω/c_s, the near-field velocity is proportional to ω^{-1}, and becomes very large at low frequencies. Thus a velocity *microphone*, i.e., one that responds to acoustic velocity u rather than to pressure, will respond much more strongly to low frequencies than to high ones when the microphone is held close to one's lips (r $<<\lambda/2\pi$); this effect is usually compensated electronically. The advantage of velocity microphones is that they are largely deaf to ambient noise originating in their far field (r >> $\lambda/2\pi$), although they are sensitive to local wind turbulence.

The acoustic intensity I(r) can be computed using (13.1.22) for a sphere of radius $\underline{a}$ oscillating with a surface velocity $\underline{u}_o$ at r = a. In this case $\bar{u}(a) = \hat{r}\underline{u}_o$, and substituting this value for $\bar{u}$ into (13.3.7) yields the constant $\underline{K} = j\underline{u}_o\eta_s a^2$; this near-field equation is appropriate only if a << $\lambda/2\pi$. Thus, using (13.3.4) and (13.3.6), the far field intensity is:

$$I = R_e\{\underline{pu}^*\}/2 = |\underline{K}|^2/2\eta_s r^2 = \eta_s\left|2\pi\underline{u}_o a^2\right|^2/2 \; \left[W\;m^{-2}\right] \qquad (13.3.8)$$

Integrating I over a sphere of radius r yields the total acoustic power transmitted:

$$P_t = 2\pi\eta_s\left|\omega a^2\underline{u}_o/c_s\right|^2 \; [W] \qquad \text{(acoustic power radiated)} \quad (13.3.9)$$

where $2\pi/\lambda = \omega/c_s$ has been substituted. Thus P_t is proportional to $\eta_s\omega^2 a^4(u_o/c_s)^2$. This suggests the importance of using a high frequency ω and large radius a if substantial power is to be radiated using a velocity source $\underline{u}_o$.

If we imagine a Thevenin equivalent acoustic source providing a "current" of u_o, then, using (13.3.9), the *acoustic radiation resistance* of this acoustic antenna is:

$$R_r = P_t/\left(|\underline{u}_o|^2/2\right) = 4\pi\eta_s\left(ka^2\right)^2 \; \left[kg\;s^{-1}\right] \qquad (13.3.10)$$

Arrays of such acoustic sources can synthesis a wide variety of antenna patterns because superposition applies and thus acoustic pressure and velocities will tend to cancel in some directions and add in others. For example, two such equal sources spaced distance d along the z axis, close compared to a wavelength and driven out of phase, would radiate the far-field pressure:

$$\underline{p}(r) \cong \left(jk\eta_s a^2 \underline{u}_o/r\right)\left(e^{-jkr_1} - e^{-jkr_2}\right) = \left(2k\eta_s a^2 \underline{u}_o/r\right)\sin\left[(kd/2)\cos\theta\right]e^{-jkr} \qquad (13.3.11)$$

where $r_{1,2} \cong r \pm (d/2)\cos\theta$. In the limit where $kd = 2\pi d/\lambda \ll 1$, (13.3.11) becomes:

$$p(r) \cong \left(k^2 d\eta_s a^2 \underline{u}_o/r\right)\cos\theta \; e^{-jkr} \qquad (13.3.12)$$

The radiated intensity $I(\theta)$ for this acoustic dipole is sketched in Figure 13.3.1(a), and is proportional to p^2 and therefore to k^4 and ω^4.
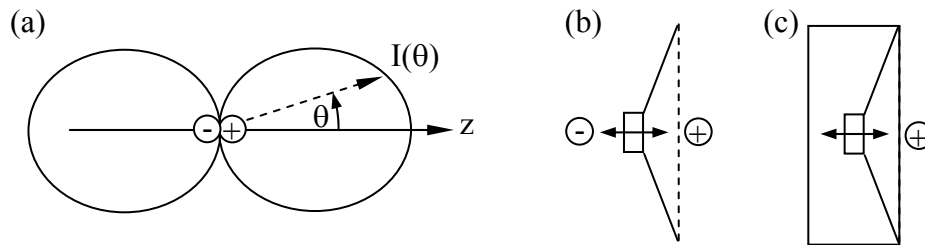
Figure 13.3.1   Acoustic radiators: (a) dipole, (b) loudspeaker, (c) baffled loudspeaker.

Thus it radiates poorly at low frequencies. Its *acoustic antenna gain* $G(\theta)$ is $3\cos^2\theta$, which can be computed by comparing the acoustic intensity I to the total acoustic power radiated P_t, just as is done for electromagnetic antennas. That is, the acoustic gain over an isotropic radiator is:

$$G(\theta,\phi) = I(\theta,\phi,r)/\left[P_t/4\pi r^2\right] \qquad \text{(acoustic antenna gain)} \quad (13.3.13)$$

$$P_t = \int_0^{2\pi}\int_0^{\pi} I(\theta,\phi,r)r^2 \; \sin\theta \; d\theta \; d\phi \; [W] \qquad (13.3.14)$$

A common way to produce this dipole acoustic pattern is illustrated in Figure 13.3.1(b) for the case of a *loudspeaker* with no baffling to block radiation from the back side of its vibrating speaker cone; the back side is clearly 180° out of phase with the velocity of the front side. The radiation from an unbaffled loudspeaker can unfortunately reflect from the walls of the room and interfere with the sound from the front side, reinforcing those frequencies for which the two rays add in phase, and diminishing those frequencies for which they are out of phase. As a result, most good loudspeakers are baffled so the reverse wave is trapped and cannot interfere with the primary wave radiated forward. This alters the acoustic impedance of the loudspeaker, but it can

be electrically compensated. The result is an acoustic monopole that radiates total power in proportion to p^2, k^2, and therefore ω^2, rather than ω^4 as for the dipole.

A linear array of monopole acoustic sources of total length L has a diffraction pattern similar to that for an array of Hertzian dipoles. If the sources are all in phase, then they radiate maximum power broadside ($\theta \equiv 0$) where all rays remain in phase. They exhibit their first null at $\theta \cong \pm\lambda/L$. See Section 10.4 for more discussion of arrays of radiators. *Acoustic array microphones* have similar directional patterns, and microphones feeding parabolic reflectors of large dimension L have even higher gains, where the gain of an acoustic antenna is proportional to its effective area. The effective area of a parabolic reflector large compared to a wavelength is approximately its physical cross-section if it is uniformly illuminated without spillover, as shown in (11.1.25) for electromagnetic waves.

13.4   Electrodynamic-acoustic devices

13.4.1   Magneto-acoustic devices

One of the most common electro-acoustic devices is the loudspeaker, where larger units typically employ a magnetic solenoid (see Section 6.4.1) to drive a large lightweight cone that pushes air with the driven waveform. The frequency limits are within the mechanical resonances of the system, which are the natural frequencies of oscillation of the cone. The low frequency mechanical limit is typically set by the resonance of the rigid cone oscillating within its support structure. An upper mechanical limit is set by the natural resonance modes of the cone itself, which are lower for larger cones because the driven waves typically propagate outward from the driven center, and can reflect from the outer edge of the cone, setting up standing waves. The amplitude limit is typically set by the strength of the system and its linearity. As shown in Section 6.1.2, mechanical motion can generate electric voltages in the same systems, so they also function as microphones.

Another magneto-acoustic device uses magnetostriction, which is the shrinkage of some magnetic materials when exposed to large magnetic fields. They are used when small powerful linear motions are desired, typically on the order of microns. To obtain larger motions the drive head can be connected to a mechanically tapered acoustic transmission line resembling a small solid version of a trumpet horn that smoothly matches the high mechanical impedance of the driver over a large area to the low mechanical impedance of the small tip. The small tip moves much greater distances because acoustic power is conserved if the taper is slow compared to a quarter-wavelength, much like a series of quarter-wave transformers being used for impedance transformation; small tips moving large distances convey the same power as large areas moving small distances. Such acoustic-transmission-line transformers can be used in either direction, depending on whether high displacements or high forces are desired.

13.4.2   Electro-acoustic devices

The simplest electro-acoustic device is perhaps a capacitor with one plate that is free to move and push air in response to time-varying electric forces on it, as discussed in Section 6.2.2.

These can be implemented macroscopically or within micro-electromechanical systems (MEMS).

Some materials such as quartz are piezo-electric and shrink or distort when high voltages are place across them. Because this warping yields little heat, periodic excitation of quartz crystals can cause them to resonate with a very high Q, making them useful for time-keeping purposes in watches, computers, and other electronic devices. These mechanical resonances for common crystals are in the MHz range and have stabilities that are $\sim 10^{-4}$–10^{-6}, depending mostly on temperature stability; larger crystals resonate at lower frequencies. They can also be designed to drive tiny resonant loudspeakers at high acoustic frequencies and efficiencies for watch alarms, etc.

By reciprocity, good piezo-electric actuators are also good sensors and can be used as microphones. Mechanical distortion of such materials generates small measurable voltages. The same is true when the plate separation of capacitors is varied, as shown in Section 6.6.1. Mechanically tapered solid acoustic waveguides can also be used for impedance transformations between low-force/high-motion terminals and high-force/low motion terminals, as noted in Section 13.4.1. Levers can also be used for the same purpose.

13.4.3   Opto-acoustic-wave transducers

When transparent materials are compressed their permittivity generally increases, slowing lightwaves passing through. This phenomenon has been used to compute Fourier transforms of broadband signals that are converted to acoustic waves propagating down the length of a transparent rectangular rod. A uniform plane wave from a laser then passes through the rod at right angles to it and to the acoustic beam, and thereby experiences local phase lags along those portions of the rod where the acoustic wave has temporarily compressed it. If the acoustic wave is at 100 MHz and the velocity of sound in the bar is 1000 m/s, then the acoustic wavelength is 10 microns. If the laser has wavelength 1 micron, then the laser light will pass straight through the bar and will also diffract at angles $\pm \lambda_{laser}/\lambda_{acoustic} = 10^{-6}/10^{-5} = 0.1$ radian. Several other beams will emerge too, at $\sim\pm0.2$, 0.3, etc. [radians]. There will therefore be a diffracted laser beam at an angle unique to each Fourier component of the acoustic signal, the strength of which depends on the magnitude of the associated optical phase delays along the rod. Lenses can then focus these various plane waves to make the power density spectrum more visible.

If several exit ports are provided for the emerging light beams, one per angle, the laser beam can effectively be switched at acoustic speeds among those ports. If 100 exit ports are provided, then the rod length L should be at least 100 wavelengths, or 1mm for the case cited above. At an acoustic velocity c_s of 1000 m/s a new wave can enter the device after $L/c_s = 10^{-3}/1000 = 10^{-6}$ seconds.

13.4.4   Surface-wave devices

Only compressive acoustic waves have been discussed so far, but acoustic shear waves can also be generated in solids, and exhibit most of the same wave phenomena as compressive waves, such as guidance and resonance. The dominant velocity in a shear wave is transverse to the

direction of wave propagation. By generating shear waves on the surface of quartz devices, and by periodically loading those surfaces mechanically with slots or metal, multiple reflections are induced that, depending on their spacing relative to a wavelength, permit band-pass and band-stop filters to be constructed, as well as transformers, resonators, and directional couplers. Because quartz has such high mechanical Q, it is often used to construct high-Q resonators at MHz frequencies.

Appendix A: Numerical Constants

A.1   Fundamental Constants

c	velocity of light	2.998×10^{8} m/s
ε_o	permittivity of free space	8.854×10^{-12} F/m
μ_o	permeability of free space	$4\pi \times 10^{-7}$ H/m
η_o	characteristic impedance of free space	$376.7\ \Omega$
e	charge of an electron, (-e.v./Joule)	-1.6008×10^{-19} C
m	mass of an electron	9.1066×10^{-31} kg
m_p	mass of a proton	1.6725×10^{-27} kg
h	Planck constant	6.624×10^{-34} J·s
k	Boltzmann constant	1.3805×10^{-23} J/K
N_o	Avogadro's constant	6.022×10^{23} molec/mole
R	Universal gas constant	8. 31 J/mole·K

A.2   Electrical Conductivity σ, S/m

Silver	6.14×10^{7}	Monel	0.24×10^{7}
Copper	5.80×10^{7}	Mercury	0.1×10^{7}
Gold	4.10×10^{7}	Sea Water	$3-5$
Aluminum	3.54×10^{7}	Distilled Water	2×10^{-4}
Tungsten	1.81×10^{7}	Bakelite	$10^{-8} - 10^{-10}$
Brass	1.57×10^{7}	Glass	10^{-12}
Nickel	1.28×10^{7}	Mica	$10^{-11} - 10^{-15}$
Iron (pure)	1.0×10^{7}	Petroleum	10^{-14}
Steel	$0.5 - 1.0 \times 10^{7}$	Fused Quartz	$<2 \times 10^{-17}$
Lead	0.48×10^{7}		

A.3   Relative Dielectric Constant $\varepsilon/\varepsilon_o$ at 1 MHz

Vacuum	1.00	Vycor glass	3.8
Styrofoam (25% filler)	1.03	Low-loss glass	4.1
Firwood	1.8 – 2.0	Ice	4.15
Paper	2.0 – 3.0	Pyrex glass	5.1
Petroleum	2.1	Muscovite (mica)	5.4
Paraffin	2.1	Mica	5.6 – 6.0
Teflon	2.1	Magnesium silicate	5.7 – 6.4
Vaseline	2.16	Porcelain	5.7
Rubber	2.3 – 4.0	Aluminum oxide	8.8
Polystyrene	2.55	Diamond	16.5
Sandy soil	2.6	Ethyl alcohol	24.5
Plexiglas	2.6 – 3.5	Distilled water	81.1
Fused quartz	3.78	Titanium dioxide	100

A.4   Relative Permeability μ/μ_o

Vacuum	1
Biological tissue	1
Cold steel	2,000
Iron (99.91%)	5,000
Purified iron (99.95%)	180,000
mu metal (FeNiCrCu)	100,000
Supermalloy (FeNiMoMn)	800,000

Appendix B: Complex Numbers and Sinusoidal Representation

Most linear systems that store energy exhibit frequency dependence and therefore are more easily characterized by their response to sinusoids rather than to arbitrary waveforms. The resulting system equations contain many instances of $A\cos(\omega t + \phi)$, where A, ω, and ϕ are the amplitude, frequency, and phase of the sinsusoid, respectively. $A\cos(\omega t + \phi)$ can be replaced by $\underline{A}$ using *complex notation*, indicated here by the underbar and reviewed below; it utilizes the arbitrary definition:

$$j \equiv (-1)^{0.5} \tag{B.1}$$

This arbitrary non-physical definition is exploited by De Moivre's theorem (B.4), which utilizes a unique property of $e = 2.71828$:

$$e^{\phi} = 1 + \phi + \phi^2/2! + \phi^3/3! + ... \tag{B.2}$$

Therefore:

$$
\begin{aligned}
e^{j\phi} &= 1 + j\phi - \phi^2/2! - j\phi^3/3! + \phi^4/4! + j\phi^5/5! - ... \\
&= \left[1 - \phi^2/2! + \phi^4/4! - ... \right] + \left[j\phi - j\phi^3/3! + j\phi^5/5!... \right]
\end{aligned}
\tag{B.3}
$$

$$e^{j\phi} = \cos\phi + j\sin\phi \tag{B.4}$$

This is a special instance of a general *complex number* $\underline{A}$:

$$\underline{A} = A_r + jA_i \tag{B.5}$$

where the real part is $A_r \equiv R_e\{\underline{A}\}$ and the imaginary part is $A_i \equiv I_m\{\underline{A}\}$.

It is now easy to use (B.4) and (B.5) to show that[76]:

$$A\cos(\omega t + \phi) = R_e\left\{ Ae^{j(\omega t + \phi)} \right\} = R_e\left\{ Ae^{j\phi}e^{j\omega t} \right\} = R_e\left\{ \underline{A}e^{j\omega t} \right\} = A_r\cos\omega t - A_i\sin\omega t \tag{B.6}$$

where:

$$\underline{A} = Ae^{j\phi} = A\cos\phi + jA\sin\phi = A_r + jA_i \tag{B.7}$$

---

[76] The physics community differs and commonly defines $A\cos(\omega t + \phi) = R_e\{Ae^{-j(\omega t + \phi)}\}$ and $A_i \equiv$ -$A\sin\phi$, where the rotational direction of ϕ is reversed in Figure B.1. Because phase is reversed in this alternative notation, the impedance of an inductor L becomes -$j\omega L$, and that of a capacitor becomes $j/\omega C$. In this notation j is commonly replaced by -i.

$$A_r \equiv A\cos\phi, \quad A_i \equiv A\sin\phi \tag{B.8}$$

The definition of $\underline{A}$ given in (B.8) has the useful geometric interpretation shown in Figure B.1(a), where the magnitude of the *phasor* $\underline{A}$ is simply the given amplitude A of the sinusoid, and the angle ϕ is its phase.
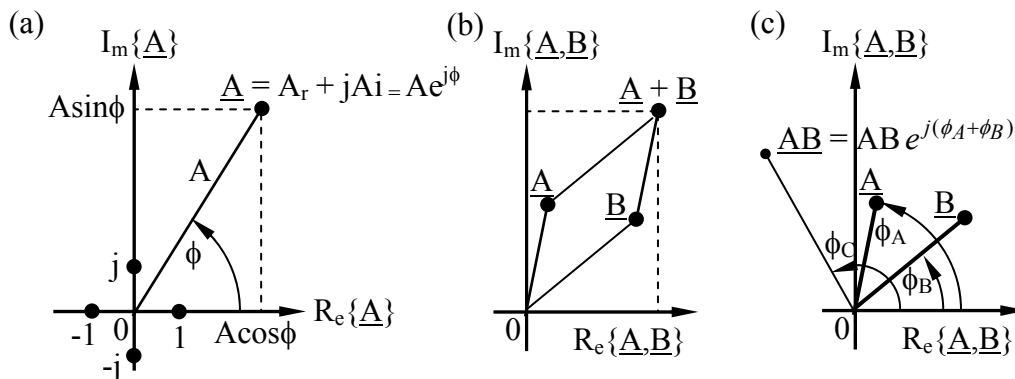
Figure B.1   Representation of phasors in the complex plane.

When $\phi = 0$ we have $R_e\{\underline{A}e^{j\omega t}\} = A\cos\omega t$, and when $\phi = \pi/2$ we have $-A\sin\omega t$. Advances in time alter the phasor $\underline{A}$ in the same sense as advances in ϕ; the phasor rotates counterclockwise. The utility of this diagram is partly that the signal of interest, $R_e\{\underline{A}e^{j\omega t}\}$, is simply the projection of the phasor $\underline{A}e^{j\omega t}$ on the real axis. It also makes clear that:

$$A = \left(A_r^2 + A_i^2\right)^{0.5} \tag{B.9}$$

$$\phi = \tan^{-1}\left(A_i/A_r\right) \tag{B.10}$$

It is also easy to see, for example, that $e^{j\pi} = -1$, and that $\underline{A} = jA$ corresponds to $-A\sin\omega t$.

Examples of equivalent representations in the time and complex domains are:

$$\begin{aligned} A\cos\omega t \;&\leftrightarrow\; A \\ -A\sin\omega t \;&\leftrightarrow\; jA \\ A\cos(\omega t + \phi) \;&\leftrightarrow\; Ae^{j\phi} \\ A\sin(\omega t + \phi) \;&\leftrightarrow\; -jAe^{j\phi} = Ae^{j(\phi - \pi/2)} \end{aligned}$$

Complex numbers behave as vectors in some respects, where addition and multiplication are also illustrated in Figure B.1(b) and (c), respectively:

$$\underline{A} + \underline{B} = \underline{B} + \underline{A} = A_r + B_r + j(A_i + B_i) \tag{B.11}$$

$$\underline{A}\underline{B} = \underline{B}\underline{A} = (A_r B_r - A_i B_i) + j(A_r B_i + A_i B_r) = AB\, e^{j(\phi_A + \phi_B)} \tag{B.12}$$

$$\underline{A}^* = A_r - jA_i = A\, e^{-j\phi_A} \tag{B.13}$$

We can easily solve for the real and imaginary parts of $\underline{A}$:

$$A_r = \left(\underline{A} + \underline{A}^*\right)\!/2, \qquad A_i = \left(\underline{A} - \underline{A}^*\right)\!/2 \tag{B.14}$$

Ratios of complex numbers can also be readily computed:

$$\underline{A}/\underline{B} = (A/B)e^{j(\phi_A - \phi_B)} = \underline{A}\underline{B}^* / \underline{B}\underline{B}^* = \underline{A}\underline{B}^* / |\underline{B}|^2 \tag{B.15}$$

Even an nth root of $\underline{A} = Ae^{j\phi}$ can be simply found:

$$\underline{A}^{1/n} = A^{1/n} e^{j\phi/n} \tag{B.16}$$

where n legitimate roots exist and are:

$$\underline{A}^{1/n} = A^{(1/n)} e^{(j\phi/n)} e^{(j2\pi m/n)} \tag{B.17}$$

for m = 0, 1, …, n − 1.

Appendix C: Mathematical Identities

$$\overline{A} = \hat{x}A_x + \hat{y}A_y + \hat{z}A_z$$

$$\overline{A}\bullet\overline{B} = A_xB_x + A_yB_y + A_zB_z = \hat{a}\times\hat{b}\left|\overline{A}\right|\left|\overline{B}\right|\cos\theta$$

$$\overline{A}\times\overline{B} = \det\begin{vmatrix} \hat{x} & \hat{y} & \hat{z} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix}$$

$$= \hat{x}\left(A_yB_z - A_zB_y\right) + \hat{y}\left(A_zB_x - A_xB_z\right) + \hat{z}\left(A_xB_y - A_yB_x\right)$$

$$= \hat{a}\times\hat{b}\left|\overline{A}\right|\left|\overline{B}\right|\sin\theta$$

$$\overline{A}\bullet(\overline{B}\times\overline{C}) = \overline{B}\bullet(\overline{C}\times\overline{A}) = \overline{C}\bullet(\overline{A}\times\overline{B})$$

$$\overline{A}\times(\overline{B}\times\overline{C}) = (\overline{A}\bullet\overline{C})\overline{B} - (\overline{A}\bullet\overline{B})\overline{C}$$

$$(\overline{A}\times\overline{B})\bullet(\overline{C}\times\overline{D}) = (\overline{A}\bullet\overline{C})(\overline{B}\bullet\overline{D}) - (\overline{A}\bullet\overline{D})(\overline{B}\bullet\overline{C})$$

$$\nabla\times\nabla\Psi = 0$$

$$\nabla\bullet(\nabla\times\overline{A}) = 0$$

$$\nabla\times(\nabla\times\overline{A}) = \nabla(\nabla\bullet\overline{A}) - \nabla^2\overline{A}$$

$$-\overline{A}\times(\nabla\times\overline{A}) = (\overline{A}\bullet\nabla)\overline{A} - \frac{1}{2}\nabla(\overline{A}\bullet\overline{A})$$

$$\nabla(\Psi\Phi) = \Psi\nabla\Phi + \Phi\nabla\Psi$$

$$\nabla\bullet(\Psi\overline{A}) = \overline{A}\bullet\nabla\Psi + \Psi\nabla\bullet\overline{A}$$

$$\nabla\times(\Psi\overline{A}) = \nabla\Psi\times\overline{A} + \Psi\nabla\times\overline{A}$$

$$\nabla^2\Psi = \nabla\bullet\nabla\Psi$$

$$\nabla(\overline{A}\bullet\overline{B}) = (\overline{A}\bullet\nabla)\overline{B} + (\overline{B}\bullet\nabla)\overline{A} + \overline{A}\times(\nabla\times\overline{B}) + \overline{B}\times(\nabla\times\overline{A})$$

$$\nabla\bullet(\overline{A}\times\overline{B}) = \overline{B}\bullet(\nabla\times\overline{A}) - \overline{A}\bullet(\nabla\times\overline{B})$$

$$\nabla\times(\overline{A}\times\overline{B}) = \overline{A}(\nabla\bullet\overline{B}) - \overline{B}(\nabla\bullet\overline{A}) + (\overline{B}\bullet\nabla)\overline{A} - (\overline{A}\bullet\nabla)\overline{B}$$

Cartesian Coordinates (x,y,z):

$$\nabla\Psi = \hat{x}\frac{\partial\Psi}{\partial x} + \hat{y}\frac{\partial\Psi}{\partial y} + \hat{z}\frac{\partial\Psi}{\partial z}$$

$$\nabla\cdot\overline{A} = \frac{\partial A_x}{\partial x} + \frac{\partial A_y}{\partial y} + \frac{\partial A_z}{\partial z}$$

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

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

Cylindrical coordinates (r,ϕ,z):

$$\nabla\Psi = \hat{\rho}\frac{\partial\Psi}{\partial r} + \hat{\phi}\frac{1}{r}\frac{\partial\Psi}{\partial y} + \hat{z}\frac{\partial\Psi}{\partial z}$$

$$\nabla\cdot\overline{A} = \frac{1}{r}\frac{\partial(rA_r)}{\partial r} + \frac{1}{r}\frac{\partial A_\phi}{\partial\phi} + \frac{\partial A_z}{\partial z}$$

$$\nabla\times\overline{A} = \hat{r}\left(\frac{1}{r}\frac{\partial A_z}{\partial\phi} - \frac{\partial A_\phi}{\partial z}\right) + \hat{\phi}\left(\frac{\partial A_r}{\partial z} - \frac{\partial A_z}{\partial r}\right) + \hat{z}\frac{1}{r}\left(\frac{\partial(rA_\phi)}{\partial r} - \frac{\partial A_r}{\partial\phi}\right) = \frac{1}{r}\det\begin{vmatrix} \hat{r} & r\hat{\phi} & \hat{z} \\ \partial/\partial r & \partial/\partial\phi & \partial/\partial z \\ A_r & rA_\phi & A_z \end{vmatrix}$$

$$\nabla^2\Psi = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial\Psi}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2\Psi}{\partial\phi^2} + \frac{\partial^2\Psi}{\partial z^2}$$

Spherical coordinates (r,θ,ϕ):

$$\nabla\Psi = \hat{r}\frac{\partial\Psi}{\partial r} + \hat{\theta}\frac{1}{r}\frac{\partial\Psi}{\partial\theta} + \hat{\phi}\frac{1}{r\sin\theta}\frac{\partial\Psi}{\partial\phi}$$

$$\nabla\cdot\overline{A} = \frac{1}{r^2}\frac{\partial(r^2 A_r)}{\partial r} + \frac{1}{r\sin\theta}\frac{\partial(\sin\theta A_\theta)}{\partial\theta} + \frac{1}{r\sin\theta}\frac{\partial A_\phi}{\partial\phi}$$

$$\nabla\times\overline{A} = \hat{r}\frac{1}{r\sin\theta}\left(\frac{\partial(r\sin\theta A_\phi)}{\partial\theta} - \frac{\partial A_\theta}{\partial\phi}\right) + \hat{\theta}\left(\frac{1}{r\sin\theta}\frac{\partial A_r}{\partial\phi} - \frac{1}{r}\frac{\partial(rA_\phi)}{\partial r}\right) + \hat{\phi}\frac{1}{r}\left(\frac{\partial(rA_\theta)}{\partial r} - \frac{\partial A_r}{\partial\theta}\right)$$

$$= \frac{1}{r^2\sin\theta}\det\begin{vmatrix} \hat{r} & r\hat{\theta} & r\sin\theta\,\hat{\phi} \\ \partial/\partial r & \partial/\partial\theta & \partial/\partial\phi \\ A_r & rA_\theta & r\sin\theta A_\phi \end{vmatrix}$$

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

Gauss' Divergence Theorem:

$$\int_V \nabla \cdot \overline{G} \, dv = \oint_A \overline{G} \cdot \hat{n} \, da$$

Stokes' Theorem:

$$\int_A (\nabla \times \overline{G}) \cdot \hat{n} \, da = \oint_C \overline{G} \cdot d\overline{\ell}$$

Fourier Transforms for pulse signals h(t):

$$\underline{H}(f) = \int_{-\infty}^{\infty} h(t) e^{-j2\pi ft} \, dt$$

$$h(t) = \int_{-\infty}^{\infty} \underline{H}(f) e^{+j2\pi ft} \, df$$

Appendix D: Basic Equations for Electromagnetics and Applications

Fundamentals

$\bar{f} = q(\bar{E} + \bar{v} \times \mu_o \bar{H})[N]$

$\nabla \times \bar{E} = -\partial\bar{B}/\partial t$

$\oint_c \bar{E} \bullet d\bar{s} = -\dfrac{d}{dt}\int_A \bar{B} \bullet d\bar{a}$

$\nabla \times \bar{H} = \bar{J} + \partial\bar{D}/\partial t$

$\oint_c \bar{H} \bullet d\bar{s} = \int_A \bar{J} \bullet d\bar{a} + \dfrac{d}{dt}\int_A \bar{D} \bullet d\bar{a}$

$\nabla \bullet \bar{D} = \rho \rightarrow \oint_A \bar{D} \bullet d\bar{a} = \int_V \rho dv$

$\nabla \bullet \bar{B} = 0 \rightarrow \oint_A \bar{B} \bullet d\bar{a} = 0$

$\nabla \bullet \bar{J} = -\partial\rho/\partial t$

$\bar{E}$ = electric field (Vm^{-1})

$\bar{H}$ = magnetic field (Am^{-1})

$\bar{D}$ = electric displacement (Cm^{-2})

$\bar{B}$ = magnetic flux density (T)

Tesla (T) = Weber m^{-2} = 10,000 gauss

ρ = charge density (Cm^{-3})

$\bar{J}$ = current density (Am^{-2})

σ = conductivity (Siemens m^{-1})

$\bar{J}_s$ = surface current density (Am^{-1})

ρ_s = surface charge density (Cm^{-2})

$\varepsilon_o = 8.85 \times 10^{-12}$ Fm^{-1}

$\mu_o = 4\pi \times 10^{-7}$ Hm^{-1}

$c = (\varepsilon_o\mu_o)^{-0.5} \cong 3 \times 10^8$ ms^{-1}

$e = -1.60 \times 10^{-19}$ C

$\eta_o \cong 377$ ohms $= (\mu_o/\varepsilon_o)^{0.5}$

$\left(\nabla^2 - \mu\varepsilon\partial^2/\partial t^2\right)\bar{E} = 0$ [Wave Eqn.]

$E_y(z,t) = E_+(z-ct) + E_-(z+ct) = R_e\{\underline{E}_y(z)e^{j\omega t}\}$

$H_x(z,t) = \eta_o^{-1}[E_+(z-ct)-E_-(z+ct)]$ [or(ωt-kz) or (t-z/c)]

$\int_A (\bar{E} \times \bar{H}) \bullet d\bar{a} + (d/dt)\int_V \left(\varepsilon|\bar{E}|^2\big/2 + \mu|\bar{H}|^2\big/2\right)dv$

$\qquad = -\int_V \bar{E} \bullet \bar{J}\, dv$ (Poynting Theorem)

Media and Boundaries

$\bar{D} = \varepsilon_o\bar{E} + \bar{P}$

$\nabla \bullet \bar{D} = \rho_f$, $\tau = \varepsilon/\sigma$

$\nabla \bullet \varepsilon_o\bar{E} = \rho_f + \rho_p$

$\nabla \bullet \bar{P} = -\rho_p$, $\bar{J} = \sigma\bar{E}$

$\bar{B} = \mu\bar{H} = \mu_o\left(\bar{H} + \bar{M}\right)$

$\varepsilon = \varepsilon_o\left(1 - \omega_p^2\big/\omega^2\right)$

$\omega_p = \left(Ne^2\big/m\varepsilon_o\right)^{0.5}$

$\varepsilon_{eff} = \varepsilon\left(1 - j\sigma/\omega\varepsilon\right)$

skin depth $\delta = (2/\omega\mu\sigma)^{0.5}$ [m]

$\bar{E}_{1//} - \bar{E}_{2//} = 0$

$\bar{H}_{1//} - \bar{H}_{2//} = \bar{J}_s \times \hat{n}$

$B_{1\perp} - B_{2\perp} = 0$

$(D_{1\perp} - D_{2\perp}) = \rho_s$

$\hookrightarrow 0 =$ if $\sigma = \infty$
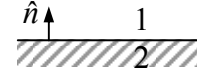
Electromagnetic Quasistatics

$\nabla^2\Phi = 0$

$KCL : \sum_i I_i(t) = 0$ at node

$KVL : \sum_i V_i(t) = 0$ around loop

$C = Q/V = A\varepsilon/d$ [F]

$L = \Lambda/I$

$i(t) = C\, dv(t)/dt$

$v(t) = L\, di(t)/dt = d\Lambda/dt$

$C_{parallel} = C_1 + C_2$

$C_{series} = (C_1^{-1} + C_2^{-1})^{-1}$

$w_e = Cv^2(t)/2$; $w_m = Li^2(t)/2$

$L_{solenoid} = N^2\mu A/W$

$\tau = RC$, $\tau = L/R$

$\Lambda = \int_A \bar{B} \bullet d\bar{a}$ (per turn)

$\bar{f} = q\left(\bar{E} + \bar{v} \times \mu_o\bar{H}\right)[N]$

$f_z = -dw_T/dz$

$\bar{F} = \bar{I} \times \mu_o\bar{H}$ [Nm^{-1}]

$\bar{E}_e = -\bar{v} \times \mu_o\bar{H}$ inside wire

$P = \omega T = W_T dV_{olume}/dt$ [W]

Max $f/A = B^2/2\mu_o$, $D^2/2\varepsilon_o$ [Nm^{-2}]

$vi = \dfrac{dw_T}{dt} + f\dfrac{dz}{dt}$

Electromagnetic Waves

$\left(\nabla^2 - \mu\varepsilon\partial^2/\partial t^2\right)\bar{E} = 0$ [Wave Eqn.]

$\left(\nabla^2 + k^2\right)\underline{\bar{E}} = 0$, $\underline{\bar{E}} = \underline{\bar{E}}_o e^{-j\bar{k}\bullet\bar{r}}$

$k = \omega(\mu\varepsilon)^{0.5} = \omega/c = 2\pi/\lambda$

$k_x^2 + k_y^2 + k_z^2 = k_o^2 = \omega^2\mu\varepsilon$

$v_p = \omega/k$, $v_g = (\partial k/\partial\omega)^{-1}$

$\theta_r = \theta_i$

$\sin\theta_t / \sin\theta_i = k_i/k_t = n_i/n_t$

$\theta_c = \sin^{-1}(n_t/n_i)$

$\theta > \theta_c \Rightarrow \underline{\overline{E}}_t = \underline{\overline{E}}_i \underline{T} e^{+\alpha x - jk_z z}$

$\underline{\overline{k}} = \overline{k}' - j\overline{k}''$

$\underline{\Gamma} = \underline{T} - 1$

$\underline{T}_{TE} = 2\big/\big(1 + [\eta_o \cos\theta_t / \eta_t \cos\theta_i]\big)$

$\underline{T}_{TM} = 2\big/\big(1 + [\eta_t \cos\theta_t / \eta_i \cos\theta_i]\big)$

$\theta_B = \tan^{-1}(\varepsilon_t/\varepsilon_i)^{0.5}$ for TM

$P_d \cong |\underline{\overline{J}}_S|^2 \big/ 2\sigma\delta \ [\text{Wm}^{-2}]$

$\overline{E} = -\nabla\phi - \partial\overline{A}/\partial t, \quad \overline{B} = \nabla\times\overline{A}$

$\underline{\Phi}(r) = \int_{V'}\big(\underline{\rho}(\overline{r})e^{-jk|\overline{r}'-\overline{r}|}\big/4\pi\varepsilon_o |\overline{r}'-\overline{r}|\big)dv'$

$\underline{\overline{A}}(r) = \int_{V'}\big(\mu_o \underline{\overline{J}}(\overline{r})e^{-jk|\overline{r}'-\overline{r}|}\big/4\pi|\overline{r}'-\overline{r}|\big)dv'$

$\underline{\overline{E}}_{ff} = \hat{\vartheta}(j\eta_o k\underline{I}d/4\pi r)e^{-jkr}\sin\theta$

$\nabla^2\underline{\Phi} + \omega^2\mu_o\varepsilon_o\underline{\Phi} = -\rho/\varepsilon_o$

$\nabla^2\underline{\overline{A}} + \omega^2\mu_o\varepsilon_o\underline{\overline{A}} = -\mu_o\underline{\overline{J}}$

Forces, Motors, and Generators

$\overline{f} = q(\overline{E} + \overline{v}\times\mu_o\overline{H})\,[\text{N}]$

$f_z = -dw_T/dz$

$\overline{F} = \overline{I}\times\mu_o\overline{H}\ [\text{Nm}^{-1}]$

$\overline{E}_e = -\overline{v}\times\mu_o\overline{H}$ inside wire

$P = \omega T = W_T dV_{olume}/dt\ [\text{W}]$

Max $f/A = B^2/2\mu_o, D^2/2\varepsilon_o\ [\text{Nm}^{-2}]$

$vi = \dfrac{dw_T}{dt} + f\dfrac{dz}{dt}$

$f = ma = d(mv)/dt$

$x = x_o + v_o t + at^2/2$

$P = fv\ [\text{W}] = T\omega$

$w_k = mv^2/2$

$T = I\,d\omega/dt$

$I = \sum_i m_i r_i^2$

Circuits

$\text{KCL}: \sum_i I_i(t) = 0$ at node

$\text{KVL}: \sum_i V_i(t) = 0$ around loop

$C = Q/V = A\varepsilon/d\ [\text{F}]$

$L = \Lambda/I$

$i(t) = C\,dv(t)/dt$

$v(t) = L\,di(t)/dt = d\Lambda/dt$

$C_{parallel} = C_1 + C_2$

$C_{series} = (C_1^{-1} + C_2^{-1})^{-1}$

$w_e = Cv^2(t)/2;\ w_m = Li^2(t)/2$

$L_{solenoid} = N^2\mu A/W$

$\tau = RC,\ \tau = L/R$

$\Lambda = \int_A \overline{B}\bullet d\overline{a}$ (per turn)

$\underline{Z}_{series} = R + j\omega L + 1/j\omega C$

$\underline{Y}_{par} = G + j\omega C + 1/j\omega L$

$Q = \omega_o w_T/P_{diss} = \omega_o/\Delta\omega$

$\omega_o = (LC)^{-0.5}$

$\langle v^2(t)\rangle\big/R = kT$

Limits to Computation Speed

$dv(z)/dz = -L\,di(z)/dt$

$di(z)/dz = -C\,dv(z)/dt$

$d^2v/dz^2 = LC\,d^2v/dt^2$

$v(z,t)\ = f_+(t - z/c) + f_-(t + z/c)$

$\qquad = g_+(z - ct) + g_-(z - ct)$

$i(t,z) = Y_o[f_+(t - z/c) - f_-(t + z/c)]$

$c = (LC)^{-0.5} = 1\big/\sqrt{\mu\varepsilon}$

$Z_o = Y_o^{-1} = (L/C)^{0.5}$

$\Gamma_L = f_-/f_+ = (R_L - Z_o)/(R_L + Z_o)$

$v(z,t) = g_+(z - ct) + g_-(z + ct)$

$V_{Th} = 2f_+(t),\ R_{Th} = Z_o$

Power Transmission

$(d^2/dz^2 + \omega^2 LC)\underline{V}(z) = 0$

$\underline{V}(z) = \underline{V}_+ e^{-jkz} + \underline{V}_- e^{+jkz}$

$\underline{I}(z) = Y_o[\underline{V}_+ e^{-jkz} - \underline{V}_- e^{+jkz}]$

$k = 2\pi/\lambda = \omega/c = \omega(\mu\varepsilon)^{0.5}$

$\underline{Z}(z) = \underline{V}(z)/\underline{I}(z) = Z_o\underline{Z}_n(z)$

$\underline{Z}_n(z) = [1 + \underline{\Gamma}(z)]\big/[1 - \underline{\Gamma}(z)] = R_n + jX_n$

$\underline{\Gamma}(z) = (\underline{V}_-/\underline{V}_+)e^{2jkz} = [\underline{Z}_n(z) - 1]\big/[\underline{Z}_n(z) + 1]$

$\underline{Z}(z) = Z_o(\underline{Z}_L - jZ_o\tan kz)\big/(\underline{Z}_o - jZ_L\tan kz)$

$\text{VSWR} = |\underline{V}_{max}|\big/|\underline{V}_{min}| = R_{max}$

Wireless Communications and Radar

$G(\theta,\phi) = P_r/(P_R/4\pi r^2)$

$P_R = \int_{4\pi} P_r(\theta,\phi,r)r^2\sin\theta\ d\theta d\phi$

$P_{rec} = P_r(\theta,\phi)A_e(\theta,\phi)$

$A_e(\theta,\phi) = G(\theta,\phi)\lambda^2/4\pi$

$R_r = P_R/\langle i^2(t)\rangle$

$E_{ff}(\theta \cong 0) = (je^{jkr}/\lambda r)\int_A E_t(x,y)e^{jk_x x + jk_y y}dxdy$

$P_{rec} = P_R(G\lambda/4\pi r^2)^2 \sigma_s/4\pi$

$\overline{E} = \sum_i a_i \overline{E}_i e^{-jk\overline{r}_i} = $ (element factor)(array f)

$E_{bit} \geq \sim 4\times 10^{-20}$ [J]

$\underline{Z}_{12} = \underline{Z}_{21}$ if reciprocity

$(d^2/dz^2 + \omega^2 LC)\underline{V}(z) = 0$

$\underline{V}(z) = \underline{V}_+ e^{-jkz} + \underline{V}_- e^{+jkz}$

$\underline{I}(z) = Y_o[\underline{V}_+ e^{-jkz} - \underline{V}_- e^{+jkz}]$

$k = 2\pi/\lambda = \omega/c = \omega(\mu\varepsilon)^{0.5}$

$\underline{Z}(z) = \underline{V}(z)/\underline{I}(z) = Z_o \underline{Z}_n(z)$

$\underline{Z}_n(z) = [1+\underline{\Gamma}(z)]/[1-\underline{\Gamma}(z)] = R_n + jX_n$

$\underline{\Gamma}(z) = (\underline{V}_-/\underline{V}_+)e^{2jkz} = [\underline{Z}_n(z)-1]/[\underline{Z}_n(z)+1]$

$\underline{Z}(z) = Z_o(\underline{Z}_L - jZ_o\tan kz)/(\underline{Z}_o - jZ_L\tan kz)$

$VSWR = |\underline{V}_{max}|/|\underline{V}_{min}| = R_{max}$

$\theta_r = \theta_i$

$\sin\theta_t/\sin\theta_i = k_i/k_t = n_i/n_t$

$\theta_c = \sin^{-1}(n_t/n_i)$

$\theta > \theta_c \Rightarrow \overline{E}_t = \overline{E}_i \underline{T}e^{+\alpha x - jk_z z}$

$\overline{\underline{k}} = \overline{k}' - j\overline{k}''$

$\underline{\Gamma} = \underline{T} - 1$

At ω_o, $\langle w_e \rangle = \langle w_m \rangle$

$\langle w_e \rangle = \int_V (\varepsilon|\underline{E}|^2/4)dv$

$\langle w_m \rangle = \int_V (\mu|\underline{H}|^2/4)dv$

$Q_n = \omega_n w_{Tn}/P_n = \omega_n/2\alpha_n$

$f_{mnp} = (c/2)([m/a]^2 + [n/b]^2 + [p/d]^2)^{0.5}$

$s_n = j\omega_n - \alpha_n$

Optical Communications

$E = hf$, photons or phonons

$hf/c = $ momentum [kg ms^{-1}]

$dn_2/dt = -[An_2 + B(n_2 - n_1)]$

Acoustics

$P = P_o + p$, $\overline{U} = \overline{U}_o + u$ $(\overline{U}_o = 0$ here$)$

$\nabla p = -\rho_o \partial\overline{u}/\partial t$

$\nabla\bullet\overline{u} = -(1/\gamma P_o)\partial p/\partial t$

$(\nabla^2 - k^2\partial^2/\partial t^2)p = 0$

$k^2 = \omega^2/c_s^2 = \omega^2\rho_o/\gamma P_o$

$c_s = v_p = v_g = (\gamma P_o/\rho_o)^{0.5}$ or $(K/\rho_o)^{0.5}$

$\eta_s = p/u = \rho_o c_s = (\rho_o\gamma P_o)^{0.5}$ gases

$\eta_s = (\rho_o K)^{0.5}$ solids, liquids

$p, \overline{u}_\perp$ continuous at boundaries

$\underline{p} = \underline{p}_+ e^{-jkz} + \underline{p}_- e^{+jkz}$

$\underline{u}_z = \eta_s^{-1}(\underline{p}_+ e^{-jkz} - \underline{p}_- e^{+jkz})$

$\int_A \overline{up}\bullet d\overline{a} + (d/dt)\int_V (\rho_o|\overline{u}|^2/2 + p^2/2\gamma P_o)dV$

Mathematical Identities

$\sin^2\theta + \cos^2\theta = 1$

$\cos\alpha + \cos\beta = 2\cos[(\alpha+\beta)/2]\cos[(\alpha-\beta)/2]$

$\underline{H}(f) = \int_{-\infty}^{+\infty} h(t)e^{-j\omega t}dt$

$e^x = 1 + x + x^2/2! + x^3/3! + ...$

$\sin\alpha = (e^{j\alpha} - e^{-j\alpha})/2j$

$\cos\alpha = (e^{j\alpha} + e^{-j\alpha})/2$

Vector Algebra

$\nabla = \hat{x}\partial/\partial x + \hat{y}\partial/\partial y + \hat{z}\partial/\partial z$

$\overline{A}\bullet\overline{B} = A_x B_x + A_y B_y + A_z B_z$

$\nabla^2\phi = (\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2)\phi$

$\nabla\bullet(\nabla\times\overline{A}) = 0$

$\nabla\times(\nabla\times\overline{A}) = \nabla(\nabla\bullet\overline{A}) - \nabla^2\overline{A}$

Gauss and Stokes' Theorems

$\iiint_V (\nabla\bullet\overline{G})dv = \oiint_A \overline{G}\bullet d\overline{a}$

$\oiint_A (\nabla\times\overline{G})\bullet d\overline{a} = \oint_c \overline{G}\bullet d\overline{s}$

Complex Numbers and Phasors

$v(t) = R_e\{\underline{V}e^{j\omega t}\}$ where $\underline{V} = |V|e^{j\phi}$

$e^{j\omega t} = \cos\omega t + j\sin\omega t$

Spherical Trigonometry

$\int_{4\pi} r^2\sin\theta\ d\theta d\phi = 4\pi$

Appendix E: Frequently Used Trigonometric and Calculus Expressions

$\sin\theta = a/c$

$\cos\theta = b/c$

$\tan\theta = a/b$

$a^2 + b^2 = c^2$

$\sin^2\theta + \cos^2\theta = 1$

$e^{j\theta} = \cos\theta + j\sin\theta$

$(d/d\theta)\sin\theta = \cos\theta$

$(d/d\theta)\cos\theta = -\sin\theta$

$(d/dx)e^{f(x)} = [df(x)/dx]\, e^{f(x)}$

$a^x = (e^{\ln a})^x$

$(d/dx)x^n = nx^{n-1}$

$(d/dx)AB = A(dB/dx) + B(dA/dx)$

$(d/dx)f_1[f_2(\theta)] = [df_1/df_2][df_2(\theta)/d\theta]d\theta/dx$

$(d/dx)\sin[f(\theta)] = \cos[f(\theta)][df(\theta)/d\theta]d\theta/dx$

$\int \sin\theta\, d\theta = -\cos\theta$

$\int \cos\theta\, d\theta = \sin\theta$

$\int e^{ax}\, dx = e^{ax}/a$

$\int x^n\, dx = x^{n+1}/(n+1)$

Index

6.013 Electromagnetics and Applications
Spring 2009