Fundamentals of Electrical Engineering I

By:
Don H. Johnson

Online:
<http://cnx.org/content/col10040/1.9.1/ >

C O N N E X I O N S

Rice University, Houston, Texas

problem, we explore a simple kind of modem, in which binary information is represented by the presence or absence of a sinusoid (presence representing a "1" and absence a "0"). Consequently, the modem's transmitted signal that represents a single bit has the form

$$x(t) = A \sin(2\pi f_0 t),\ 0 \le t \le T$$

Within each bit interval T, the amplitude is either A or zero.

a) What is the smallest transmission interval that makes sense with the frequency f_0?
b) Assuming that ten cycles of the sinusoid comprise a single bit's transmission interval, what is the datarate of this transmission scheme?
c) Now suppose instead of using "on-off" signaling, we allow one of several **different** values for the amplitude during any transmission interval. If N amplitude values are used, what is the resulting datarate?
d) The classic communications block diagram applies to the modem. Discuss how the transmitter must interface with the message source since the source is producing letters of the alphabet, not bits.

Problem 1.3: Advanced Modems
To transmit symbols, such as letters of the alphabet, RU computer modems use two frequencies (1600 and 1800 Hz) and several amplitude levels. A transmission is sent for a period of time T (known as the transmission or baud interval) and equals the sum of two amplitude-weighted carriers.

$$x(t) = A_1 \sin(2\pi f_1 t) + A_2 \sin(2\pi f_2 t),\ 0 \le t \le T$$

We send successive symbols by choosing an appropriate frequency and amplitude combination, and sending them one after another.

a) What is the smallest transmission interval that makes sense to use with the frequencies given above? In other words, what should T be so that an integer number of cycles of the carrier occurs?
b) Sketch (using Matlab) the signal that modem produces over several transmission intervals. Make sure you axes are labeled.
c) Using your signal transmission interval, how many amplitude levels are needed to transmit ASCII characters at a datarate of 3,200 bits/s? Assume use of the extended (8-bit) ASCII code.

NOTE: We use a discrete set of values for A_1 and A_2. If we have N_1 values for amplitude A_1, and N_2 values for A_2, we have $N_1 N_2$ possible symbols that can be sent during each T second interval. To convert this number into bits (the fundamental unit of information engineers use to qualify things), compute $\log_2(N_1 N_2)$.

Exercise 2.2 *(Solution on p. 31.)*

Convert $3 - 2j$ to polar form.

2.1.2 Euler's Formula

Surprisingly, the polar form of a complex number z can be expressed mathematically as

$$z = re^{j\theta} \tag{2.2}$$

To show this result, we use **Euler's relations** that express exponentials with imaginary arguments in terms of trigonometric functions.

$$e^{j\theta} = \cos(\theta) + j\sin(\theta) \tag{2.3}$$

$$\cos(\theta) = \frac{e^{j\theta} + e^{-j\theta}}{2} \tag{2.4}$$

$$\sin(\theta) = \frac{e^{j\theta} - e^{-j\theta}}{2j}$$

The first of these is easily derived from the Taylor's series for the exponential.

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

Substituting $j\theta$ for x, we find that

$$e^{j\theta} = 1 + j\frac{\theta}{1!} - \frac{\theta^2}{2!} - j\frac{\theta^3}{3!} + \dots$$

because $j^2 = -1$, $j^3 = -j$, and $j^4 = 1$. Grouping separately the real-valued terms and the imaginary-valued ones,

$$e^{j\theta} = 1 - \frac{\theta^2}{2!} + \dots + j\left(\frac{\theta}{1!} - \frac{\theta^3}{3!} + \dots\right)$$

The real-valued terms correspond to the Taylor's series for $\cos(\theta)$, the imaginary ones to $\sin(\theta)$, and Euler's first relation results. The remaining relations are easily derived from the first. Because of the relationship $r = \sqrt{a^2 + b^2}$, we see that multiplying the exponential in (2.3) by a real constant corresponds to setting the radius of the complex number by the constant.

2.1.3 Calculating with Complex Numbers

Adding and subtracting complex numbers expressed in Cartesian form is quite easy: You add (subtract) the real parts and imaginary parts separately.

$$(z_1 \pm z_2) = (a_1 \pm a_2) + j(b_1 \pm b_2) \tag{2.5}$$

To multiply two complex numbers in Cartesian form is not quite as easy, but follows directly from following the usual rules of arithmetic.

$$\begin{aligned} z_1 z_2 &= (a_1 + jb_1)(a_2 + jb_2) \\ &= a_1 a_2 - b_1 b_2 + j(a_1 b_2 + a_2 b_1) \end{aligned} \tag{2.6}$$

Note that we are, in a sense, multiplying two vectors to obtain another vector. Complex arithmetic provides a unique way of defining vector multiplication.

Exercise 2.3 *(Solution on p. 31.)*

What is the product of a complex number and its conjugate?

2.7 Signals and Systems Problems[10]

Problem 2.1: Complex Number Arithmetic
Find the real part, imaginary part, the magnitude and angle of the complex numbers given by the following expressions.

a) -1
b) $\frac{1+\sqrt{3}j}{2}$
c) $1+j+e^{j\frac{\pi}{2}}$
d) $e^{j\frac{\pi}{3}}+e^{j\pi}+e^{-j\frac{\pi}{3}}$

Problem 2.2: Discovering Roots
Complex numbers expose all the roots of real (and complex) numbers. For example, there should be two square-roots, three cube-roots, etc. of any number. Find the following roots.

a) What are the cube-roots of 27? In other words, what is $27^{\frac{1}{3}}$?
b) What are the fifth roots of 3 ($3^{\frac{1}{5}}$)?
c) What are the fourth roots of one?

Problem 2.3: Cool Exponentials
Simplify the following (cool) expressions.

a) j^j
b) j^{2j}
c) j^{j^j}

Problem 2.4: Complex-valued Signals
Complex numbers and phasors play a very important role in electrical engineering. Solving systems for complex exponentials is much easier than for sinusoids, and linear systems analysis is particularly easy.

a) Find the phasor representation for each, and re-express each as the real and imaginary parts of a complex exponential. What is the frequency (in Hz) of each? In general, are your answers unique? If so, prove it; if not, find an alternative answer for the complex exponential representation.
   i) $3\sin(24t)$
   ii) $\sqrt{2}\cos\left(2\pi 60t+\frac{\pi}{4}\right)$
   iii) $2\cos\left(t+\frac{\pi}{6}\right)+4\sin\left(t-\frac{\pi}{3}\right)$
b) Show that for linear systems having real-valued outputs for real inputs, that when the input is the real part of a complex exponential, the output is the real part of the system's output to the complex exponential (see Figure 2.16).

$$\mathcal{S}\left[\mathrm{Re}\left\{Ae^{j2\pi ft}\right\}\right]=\mathrm{Re}\left\{\mathcal{S}\left[Ae^{j2\pi ft}\right]\right\}$$

Problem 2.5:
For each of the indicated voltages, write it as the real part of a complex exponential ($v(t)=\mathrm{Re}\left[Ve^{st}\right]$). Explicitly indicate the value of the complex amplitude V and the complex frequency s. Represent each complex amplitude as a vector in the V-plane, and indicate the location of the frequencies in the complex s-plane.

a) $v(t)=\cos(5t)$
b) $v(t)=\sin\left(8t+\frac{\pi}{4}\right)$
c) $v(t)=e^{-t}$
d) $v(t)=e^{-3t}\sin\left(4t+\frac{3\pi}{4}\right)$

[10] This content is available online at <http://cnx.org/content/m10348/2.27/>.

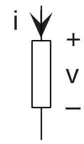


Figure 3.1: The generic circuit element.

Current flows through circuit elements, such as that depicted in Figure 3.1 (Generic Circuit Element), and through conductors, which we indicate by lines in circuit diagrams. For every circuit element we define a voltage and a current. The element has a **v-i** relation defined by the element's physical properties. In defining the **v-i** relation, we have the convention that positive current flows from positive to negative voltage drop. Voltage has units of volts, and both the unit and the quantity are named for Volta[2]. Current has units of amperes, and is named for the French physicist Ampère[3].

Voltages and currents also carry **power**. Again using the convention shown in Figure 3.1 (Generic Circuit Element) for circuit elements, the **instantaneous power** at each moment of time consumed by the element is given by the product of the voltage and current.

$$p(t) = v(t)\, i(t)$$

A positive value for power indicates that at time t the circuit element is **consuming** power; a negative value means it is **producing** power. With voltage expressed in volts and current in amperes, power defined this way has units of **watts**. Just as in all areas of physics and chemistry, power is the rate at which **energy** is consumed or produced. Consequently, energy is the integral of power.

$$E(t) = \int_{-\infty}^{t} p(\alpha)\, d\alpha$$

Again, positive energy corresponds to consumed energy and negative energy corresponds to energy production. Note that a circuit element having a power profile that is both positive and negative over some time interval could consume or produce energy according to the sign of the integral of power. The units of energy are **joules** since a watt equals joules/second.

Exercise 3.1 *(Solution on p. 98.)*

Residential energy bills typically state a home's energy usage in kilowatt-hours. Is this really a unit of energy? If so, how many joules equals one kilowatt-hour?

3.2 Ideal Circuit Elements[4]

The elementary circuit elements—the resistor, capacitor, and inductor— impose **linear** relationships between voltage and current.

[2] http://www.bioanalytical.com/info/calendar/97/volta.htm
[3] http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Ampere.html
[4] This content is available online at <http://cnx.org/content/m0012/2.21/>.

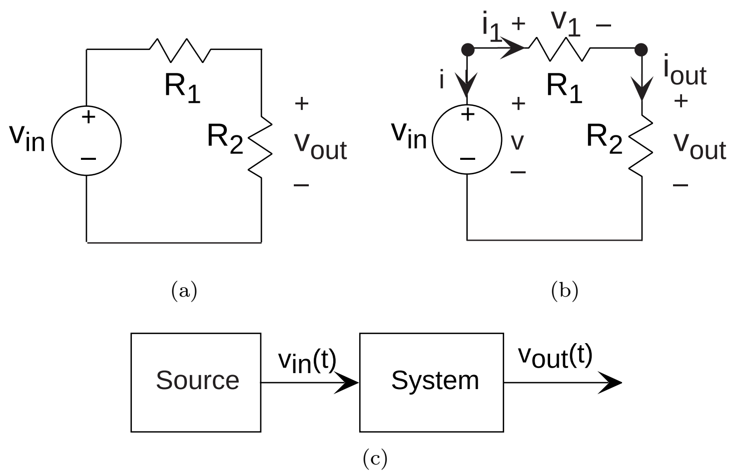


Figure 3.6: The circuit shown in the top two figures is perhaps the simplest circuit that performs a signal processing function. On the bottom is the block diagram that corresponds to the circuit. The input is provided by the voltage source v_{in} and the output is the voltage v_{out} across the resistor label R_2. As shown in the middle, we **analyze** the circuit—understand what it accomplishes—by defining currents and voltages for all circuit elements, and then solving the circuit and element equations.

is intended to make the diagram look pretty. This line simply means that the two elements are connected together. **Kirchhoff's Laws**, one for voltage (Section 3.4.2: Kirchhoff's Voltage Law (KVL)) and one for current (Section 3.4.1: Kirchhoff's Current Law), determine what a connection among circuit elements means. These laws can help us analyze this circuit.

3.4.1 Kirchhoff's Current Law

At every node, the sum of all currents entering a node must equal zero. What this law means physically is that charge cannot accumulate in a node; what goes in must come out. In the example, Figure 3.6, below we have a three-node circuit and thus have three KCL equations.

$$-i - i_1 = 0$$

$$i_1 - i_2 = 0$$

$$i + i_2 = 0$$

Note that the current entering a node is the negative of the current leaving the node.

Given any two of these KCL equations, we can find the other by adding or subtracting them. Thus, one of them is redundant and, in mathematical terms, we can discard any one of them. The convention is to discard the equation for the (unlabeled) node at the bottom of the circuit.

Exercise 3.2 ***(Solution on p. 98.)***

In writing KCL equations, you will find that in an n-node circuit, exactly one of them is always redundant. Can you sketch a proof of why this might be true? Hint: It has to do with the fact that charge won't accumulate in one place on its own.

3.4.2 Kirchhoff's Voltage Law (KVL)

The voltage law says that the sum of voltages around every closed loop in the circuit must equal zero. A closed loop has the obvious definition: Starting at a node, trace a path through the circuit that returns you to the origin node. KVL expresses the fact that electric fields are conservative: The total work performed in moving a test charge around a closed path is zero. The KVL equation for our circuit is

$$v_1 + v_2 - v = 0$$

we obtain the quantity we seek.

$$v_{\text{out}} = \frac{R_2}{R_1 + R_2} v_{\text{in}}$$

Exercise 3.3 ***(Solution on p. 98.)***

Referring back to Figure 3.6, a circuit should serve some useful purpose. What kind of system does our circuit realize and, in terms of element values, what are the system's parameter(s)?

3.5 Power Dissipation in Resistor Circuits[11]

We can find voltages and currents in simple circuits containing resistors and voltage or current sources. We should examine whether these circuits variables obey the Conservation of Power principle: since a circuit is a closed system, it should not dissipate or create energy. For the moment, our approach is to investigate first a resistor circuit's **power** consumption/creation. Later, we will **prove** that because of **KVL** and **KCL** **all** circuits conserve power.

As defined on p. 34, the instantaneous power consumed/created by every circuit element equals the product of its voltage and current. The total power consumed/created by a circuit equals the sum of each element's power.

$$P = \sum_k v_k i_k$$

Recall that each element's current and voltage must obey the convention that positive current is defined to enter the positive-voltage terminal. With this convention, a positive value of $v_k i_k$ corresponds to consumed power, a negative value to created power. Because the total power in a circuit must be zero ($P = 0$), some circuit elements must create power while others consume it.

Consider the simple series circuit should in Section 3.4. In performing our calculations, we defined the current i_{out} to flow through the positive-voltage terminals of both resistors and found it to equal $i_{\text{out}} = \frac{v_{\text{in}}}{R_1+R_2}$. The voltage across the resistor R_2 is the output voltage and we found it to equal $v_{\text{out}} = \frac{R_2}{R_1+R_2} v_{\text{in}}$. Consequently, calculating the power for this resistor yields

$$P_2 = \frac{R_2}{(R_1 + R_2)^2} v_{\text{in}}^2$$

Consequently, this resistor dissipates power because P_2 is positive. This result should not be surprising since we showed (p. 35) that the power consumed by **any** resistor equals either of the following.

$$\frac{v^2}{R} \text{ or } i^2 R \tag{3.3}$$

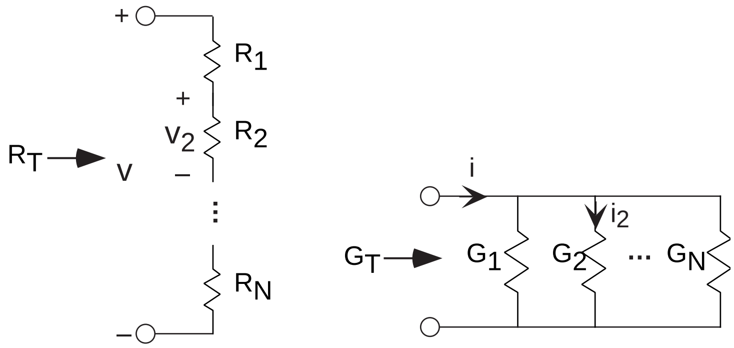
Since resistors are positive-valued, **resistors always dissipate power**. But where does a resistor's power go? By Conversation of Power, the dissipated power must be absorbed somewhere. The answer is not directly predicted by circuit theory, but is by physics. Current flowing through a resistor makes it hot; its power is dissipated by heat.

NOTE: A physical wire has a resistance and hence dissipates power (it gets warm just like a resistor in a circuit). In fact, the resistance of a wire of length L and cross-sectional area A is given by

$$R = \frac{\rho L}{A}$$

The quantity ρ is known as the **resistivity** and presents the resistance of a unit-length unit cross-sectional area material constituting the wire. Resistivity has units of ohm-meters. Most materials

[11] This content is available online at <http://cnx.org/content/m17305/1.5/>.



(a) series combination rule (b) parallel combination rule

Figure 3.16: Series and parallel combination rules. (a) $R_T = \sum_{n=1}^{N} (R_n)$, $v_n = \frac{R_n}{R_T} v$ (b) $G_T = \sum_{n=1}^{N} G_n$, $i_n = \frac{G_n}{G_T} i$

This result is known as an **equivalent circuit**: from the viewpoint of a pair of terminals, a group of resistors functions as a single resistor, the resistance of which can usually be found by applying the parallel and series rules.

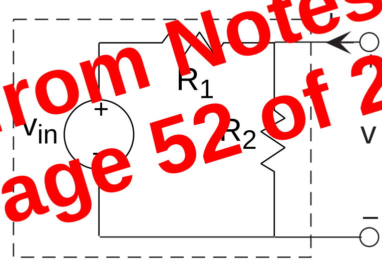


Figure 3.17

This result generalizes to include sources in a very interesting and useful way. Let's consider our simple attenuator circuit (shown in the figure (Figure 3.17)) from the viewpoint of the output terminals. We want to find the **v-i** relation for the output terminal pair, and then find the equivalent circuit for the boxed circuit. To perform this calculation, use the circuit laws and element relations, but do not attach anything to the output terminals. We seek the relation between v and i that describes the kind of element that lurks within the dashed box. The result is

$$v = (R_1 \parallel R_2)\, i + \frac{R_2}{R_1 + R_2} v_{\text{in}} \tag{3.4}$$

If the source were zero, it could be replaced by a short circuit, which would confirm that the circuit does indeed function as a parallel combination of resistors. However, the source's presence means that the circuit is **not** well modeled as a resistor.

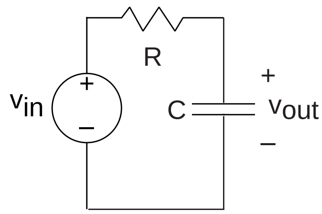


Figure 3.23: A simple RC circuit.

The input-output relation for circuits involving energy storage elements takes the form of an ordinary differential equation, which we must solve to determine what the output voltage is for a given input. In contrast to resistive circuits, where we obtain an **explicit** input-output relation, we now have an **implicit** relation that requires more work to obtain answers.

At this point, we could learn how to solve differential equations. Note first that even finding the differential equation relating an output variable to a source is often very tedious. The parallel and series combination rules that apply to resistors don't directly apply when capacitors and inductors occur. We would have to slog our way through the circuit equations, simplifying them until we finally found the equation that related the source(s) to the output. At the turn of the twentieth century, a method was discovered that not only made finding the differential equation easy, but also simplified the solution process in the most common situation. Although not original with him, Charles Steinmetz[18] presented the key paper describing the **impedance** approach in 1893. It allows circuits containing capacitors and inductors to be solved with the **same** methods we have learned to solved resistor circuits. To use impedances, we must master **complex numbers**. Though the arithmetic of complex numbers is mathematically more complicated than with real numbers, the increased insight into circuit behavior and the ease with which circuits are solved with impedances is well worth the diversion. But more importantly, the impedance concept is central to engineering and physics, having a reach far beyond just circuits.

3.9 The Impedance Concept[19]

Rather than solving the differential equation that arises in circuits containing capacitors and inductors, let's pretend that all sources in the circuit are complex exponentials having the **same** frequency. Although this pretense can only be mathematically true, this fiction will greatly ease solving the circuit no matter what the source really is.

For the above example RC circuit (Figure 3.23 (Simple Circuit)), let $v_{\mathrm{in}} = V_{\mathrm{in}} e^{j2\pi ft}$. The complex amplitude V_{in} determines the size of the source and its phase. The critical consequence of assuming that sources have this form is that **all** voltages and currents in the circuit are also complex exponentials, having amplitudes governed by KVL, KCL, and the **v-i** relations and the same frequency as the source. To appreciate why this should be true, let's investigate how each circuit element behaves when either the voltage or current is a complex exponential. For the resistor, $v = Ri$. When $v = Ve^{j2\pi ft}$; then $i = \frac{V}{R}e^{j2\pi ft}$. Thus, if the resistor's voltage is a complex exponential, so is the current, with an amplitude $I = \frac{V}{R}$ (determined by the resistor's **v-i** relation) and a frequency the same as the voltage. Clearly, if the current were assumed to be a complex exponential, so would the voltage. For a capacitor, $i = C\frac{d}{dt}(v)$. Letting the voltage be a complex exponential, we have $i = CVj2\pi fe^{j2\pi ft}$. The amplitude of this complex exponential is $I = CVj2\pi f$. Finally, for the inductor, where $v = L\frac{d}{dt}(i)$, assuming the current to be a complex exponential results in the voltage having the form $v = LIj2\pi fe^{j2\pi ft}$, making its complex amplitude $V = LIj2\pi f$.

The major consequence of assuming complex exponential voltage and currents is that the ratio $Z = \frac{V}{I}$ for each element does not depend on time, but does depend on source frequency. This quantity is known as the element's **impedance**.

---

[18]http://www.invent.org/hall_of_fame/139.html
[19]This content is available online at <http://cnx.org/content/m0024/2.23/>.

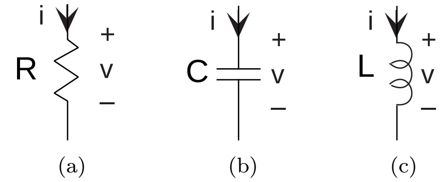


Figure 3.25: (a) Resistor: $Z_R = R$ (b) Capacitor: $Z_C = \frac{1}{j2\pi fC}$ (c) Inductor: $Z_L = j2\pi fL$

The impedance is, in general, a complex-valued, frequency-dependent quantity. For example, the magnitude of the capacitor's impedance is inversely related to frequency, and has a phase of $-\left(\frac{\pi}{2}\right)$. This observation means that if the current is a complex exponential and has constant amplitude, the amplitude of the voltage decreases with frequency.

Let's consider Kirchhoff's circuit laws. When voltages around a loop are all complex exponentials of the same frequency, we have

$$\begin{aligned}\sum_n v_n &= \sum_n V_n e^{j2\pi ft} \\ &= 0\end{aligned} \tag{3.12}$$

which means

$$\sum_n V_n = 0 \tag{3.13}$$

the complex amplitudes of the voltages obey KVL. We can easily imagine that the complex amplitudes of the currents obey KCL.

What we have discovered is that source(s) equaling a complex exponential of the same frequency forces all circuit variables to be complex exponentials of the same frequency. Consequently, the ratio of voltage to current for each element equals the ratio of their complex amplitudes, which depends only on the source's frequency and element values.

This situation occurs because the circuit elements are linear and time-invariant. For example, suppose we had a circuit element where the voltage equaled the square of the current: $v(t) = Ki^2(t)$. If $i(t) = Ie^{j2\pi ft}$, $v(t) = KI^2 e^{j2\pi 2ft}$, meaning that voltage and current no longer had the same frequency and that their ratio was time-dependent.

Because for linear circuit elements the complex amplitude of voltage is proportional to the complex amplitude of current— $V = ZI$ — assuming complex exponential sources means circuit elements behave as if they were resistors, where instead of resistance, we use impedance. **Because complex amplitudes for voltage and current also obey Kirchhoff's laws, we can solve circuits using voltage and current divider and the series and parallel combination rules by considering the elements to be impedances.**

3.10 Time and Frequency Domains[20]

When we find the differential equation relating the source and the output, we are faced with solving the circuit in what is known as the **time domain**. What we emphasize here is that it is often easier to find the output if we use impedances. Because impedances depend only on frequency, we find ourselves in the **frequency domain**. A common error in using impedances is keeping the time-dependent part, the complex exponential, in the fray. The entire point of using impedances is to get rid of time and concentrate on frequency. Only after we find the result in the frequency domain do we go back to the time domain and put things back together again.

To illustrate how the time domain, the frequency domain and impedances fit together, consider the time domain and frequency domain to be two work rooms. Since you can't be two places at the same time, you are faced with solving your circuit problem in one of the two rooms at any point in time. Impedances and complex exponentials are the way you get between the two rooms. Security guards make sure you don't try

[20]This content is available online at <http://cnx.org/content/m10708/2.9/>.

to sneak time domain variables into the frequency domain room and vice versa. Figure 3.26 (Two Rooms) shows how this works.

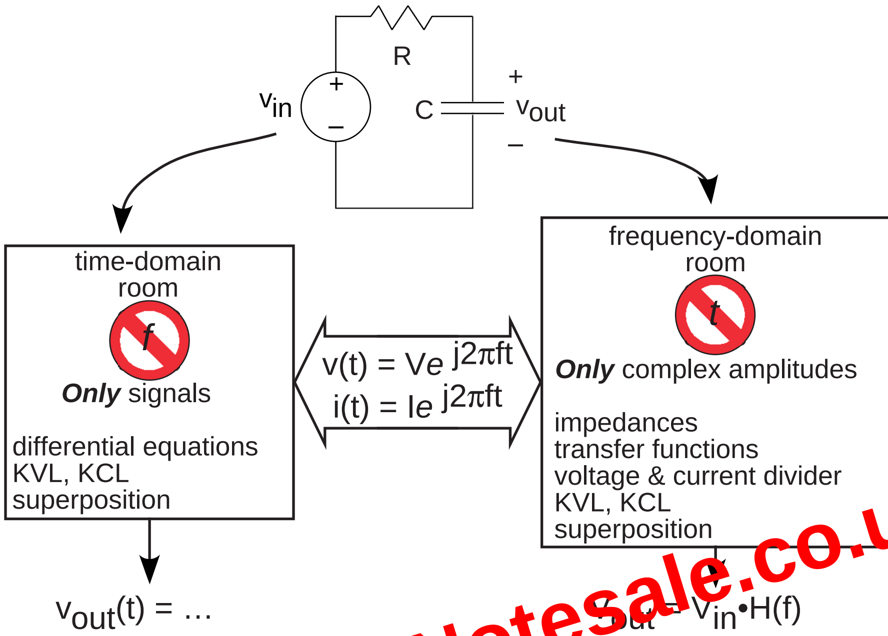


Figure 3.26: The time and frequency domains are linked by assuming signals are complex exponentials. In the time domain, signals can have any form. Passing into the frequency domain "work room," signals are represented entirely by complex amplitudes.

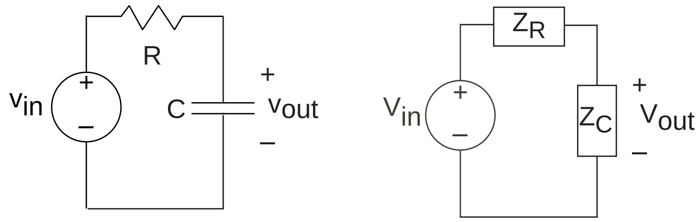
As we unfold the impedance story, we'll see that the powerful use of impedances suggested by Steinmetz[21] greatly simplifies solving circuits, alleviates us from solving differential equations, and suggests a general way of thinking about circuits. Because of the importance of this approach, let's go over how it works.

1. Even though it's not, pretend the source is a complex exponential. We do this because the impedance approach simplifies finding how input and output are related. If it were a voltage source having voltage $v_{in} = p(t)$ (a pulse), still let $v_{in} = V_{in} e^{j2\pi ft}$. We'll learn how to "get the pulse back" later.
2. With a source equaling a complex exponential, **all** variables in a linear circuit will also be complex exponentials having the **same** frequency. The circuit's only remaining "mystery" is what each variable's complex amplitude might be. To find these, we consider the source to be a complex number (V_{in} here) and the elements to be impedances.
3. We can now solve using series and parallel combination rules how the complex amplitude of any variable relates to the sources complex amplitude.

Example 3.3

To illustrate the impedance approach, we refer to the *RC* circuit (Figure 3.27 (Simple Circuits)) below, and we assume that $v_{in} = V_{in} e^{j2\pi ft}$.

[21] http://www.invent.org/hall_of_fame/139.html



(a) (b)

Figure 3.27: (a) A simple *RC* circuit. (b) The impedance counterpart for the *RC* circuit. Note that the source and output voltage are now complex amplitudes.

Using impedances, the complex amplitude of the output voltage V_{out} can be found using voltage divider:

$$V_{\text{out}} = \frac{Z_C}{Z_C + Z_R} V_{\text{in}}$$

$$V_{\text{out}} = \frac{\frac{1}{j2\pi fC}}{\frac{1}{j2\pi fC} + R} V_{\text{in}}$$

$$V_{\text{out}} = \frac{1}{j2\pi fRC + 1} V_{\text{in}}$$

If we refer to the differential equation for this circuit (shown in Circuits with Capacitors and Inductors (Section 3.8) to be $RC\frac{d}{dt}(v_{\text{out}}) + v_{\text{out}} = v_{\text{in}}$), letting the output and input voltages be complex exponentials, we obtain the same relationship between their complex amplitudes. Thus, using impedances is equivalent to using the differential equation and solving it when the source is a complex exponential.

In fact, we can find the differential equation **directly** using impedances. If we cross-multiply the relation between input and output amplitudes,

$$V_{\text{out}}(j2\pi fRC + 1) = V_{\text{in}}$$

and then put the complex exponentials back in, we have

$$RCj2\pi fV_{\text{out}}e^{j2\pi ft} + V_{\text{out}}e^{j2\pi ft} = V_{\text{in}}e^{j2\pi ft}$$

In the process of defining impedances, note that the factor $j2\pi f$ arises from the **derivative** of a complex exponential. We can reverse the impedance process, and revert back to the differential equation.

$$RC\frac{d}{dt}(v_{\text{out}}) + v_{\text{out}} = v_{\text{in}}$$

This is the same equation that was derived much more tediously in Circuits with Capacitors and Inductors (Section 3.8). Finding the differential equation relating output to input is far simpler when we use impedances than with any other technique.

Exercise 3.11 ***(Solution on p. 98.)***

Suppose you had an expression where a complex amplitude was divided by $j2\pi f$. What time-domain operation corresponds to this division?

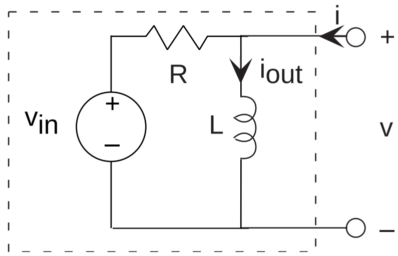


Figure 3.32

---

of linear systems, and therefore superposition applies. In particular, suppose these component signals are complex exponentials, each of which has a frequency different from the others. The transfer function portrays how the circuit affects the amplitude and phase of each component, allowing us to understand how the circuit works on a complicated signal. Those components having a frequency less than the cutoff frequency pass through the circuit with little modification while those having higher frequencies are suppressed. The circuit is said to act as a **filter**, filtering the source signal based on the frequency of each component complex exponential. Because low frequencies pass through the filter, we call it a **lowpass filter** to express more precisely its function.

We have also found the ease of calculating the output for sinusoidal inputs through the use of the transfer function. Once we find the transfer function, we can write the output directly as indicated by the output of a circuit for a sinusoidal input (3.18).

Example 3.5

Let's apply these results to a final example, in which the input is a voltage source and the output is the inductor current. The source voltage equals $V_{in} = 2\cos(2\pi 60t) + 3$. We want the circuit to pass constant (offset) voltage essentially unaltered (save for the fact that the output is a current rather than a voltage) and remove the 60 Hz term. Because the input is the sum of **two** sinusoids–a constant is a zero-frequency cosine–our approach is

1. find the transfer function using impedances;
2. use it to find the output due to each input component;
3. add the results;
4. find element values that accomplish our design criteria.

Because the circuit is a series combination of elements, let's use voltage divider to find the transfer function between V_{in} and V, then use the **v-i** relation of the inductor to find its current.

$$\begin{aligned}\frac{I_{out}}{V_{in}} &= \frac{j2\pi fL}{R + j2\pi fL} \cdot \frac{1}{j2\pi fL} \\ &= \frac{1}{j2\pi fL + R} \\ &= H(f)\end{aligned} \tag{3.19}$$

where

$$\text{voltage divider} = \frac{j2\pi fL}{R + j2\pi fL}$$

and

$$\text{inductor admittance} = \frac{1}{j2\pi fL}$$

[Do the units check?] The form of this transfer function should be familiar; it is a lowpass filter, and it will perform our desired function once we choose element values properly.

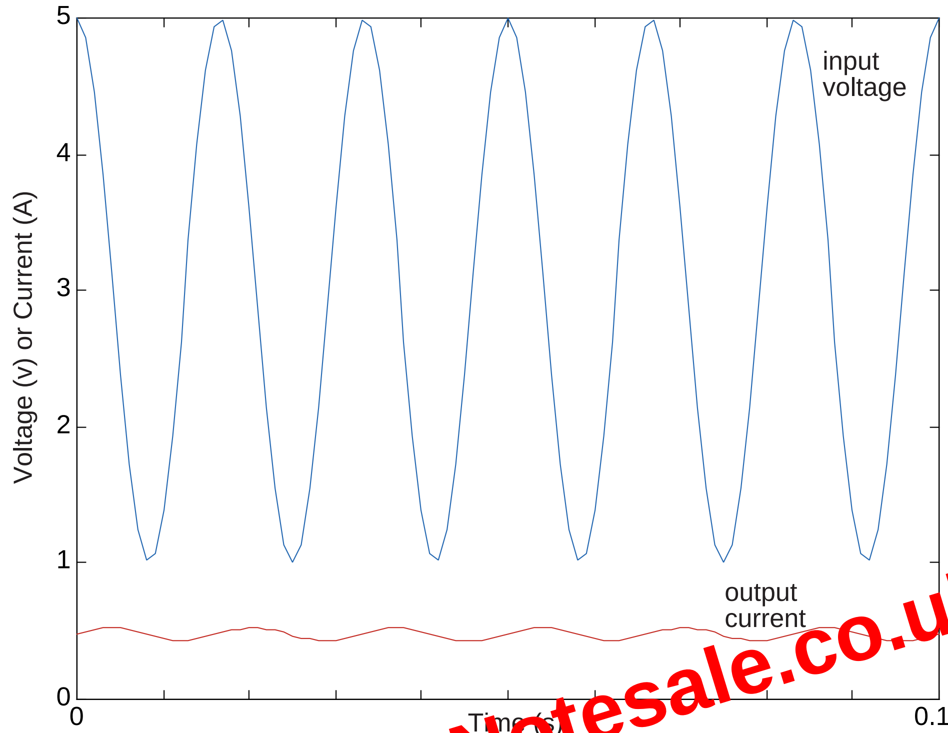


Figure 3.33: Input and output waveforms for the example RL circuit when the element values are $R = 6.28\Omega$ and $L = 100\text{mH}$.

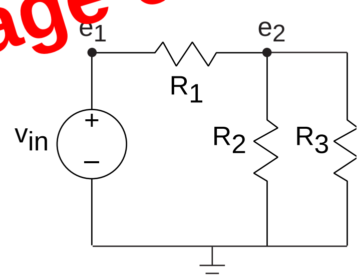


Figure 3.34

A little reflection reveals that when writing the KCL equations for the sum of currents leaving a node, that node's voltage will **always** appear with a plus sign, and all other node voltages with a minus sign. Systematic application of this procedure makes it easy to write node equations and to check them before solving them. Also remember to check units at this point: Every term should have units of current. In our example, solving for the unknown node voltage is easy:

$$e_2 = \frac{R_2 R_3}{R_1 R_2 + R_1 R_3 + R_2 R_3} v_{\text{in}} \tag{3.23}$$

Have we really solved the circuit with the node method? Along the way, we have used KVL, KCL, and the **v-i** relations. Previously, we indicated that the set of equations resulting from applying these laws is necessary and sufficient. This result guarantees that the node method can be used to "solve" **any** circuit. One fallout of this result is that we must be able to find any circuit variable given the node voltages and

- If we select the values of R_F and R so that $(GR \gg R_F)$, this factor will no longer depend on the op-amp's inherent gain, and it will equal $-\frac{1}{R_F}$.

Under these conditions, we obtain the classic input-output relationship for the op-amp-based inverting amplifier.

$$v_{\text{out}} = -\frac{R_F}{R} v_{\text{in}} \tag{3.30}$$

Consequently, the gain provided by our circuit is entirely determined by our choice of the feedback resistor R_F and the input resistor R. It is always negative, and can be less than one or greater than one in magnitude. It cannot exceed the op-amp's inherent gain and should not produce such large outputs that distortion results (remember the power supply!). Interestingly, note that this relationship does not depend on the load resistance. This effect occurs because we use load resistances large compared to the op-amp's output resistance. Thus observation means that, if careful, we can place op-amp circuits in cascade, **without** incurring the effect of succeeding circuits changing the behavior (transfer function) of previous ones; see this problem (Problem 3.43).

3.19.2 Active Filters

As long as design requirements are met, the input-output relation for the inverting amplifier also applies when the feedback and input circuit elements are impedances (resistors, capacitors, and inductors).

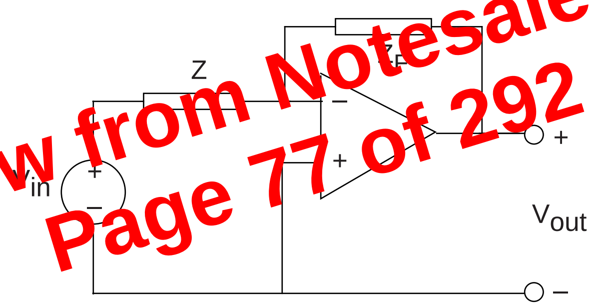


Figure 3.45: $\frac{V_{\text{out}}}{V_{\text{in}}} = -\frac{Z_F}{Z}$

Example 3.7
Let's design an op-amp circuit that functions as a lowpass filter. We want the transfer function between the output and input voltage to be

$$H(f) = \frac{K}{1 + \frac{jf}{f_c}}$$

where K equals the passband gain and f_c is the cutoff frequency. Let's assume that the inversion (negative gain) does not matter. With the transfer function of the above op-amp circuit in mind, let's consider some choices.

- $Z_F = K$, $Z = 1 + \frac{jf}{f_c}$. This choice means the feedback impedance is a resistor and that the input impedance is a series combination of an inductor and a resistor. In circuit design, we try to avoid inductors because they are physically bulkier than capacitors.
- $Z_F = \frac{1}{1+\frac{jf}{f_c}}$, $Z = \frac{1}{K}$. Consider the reciprocal of the feedback impedance (its admittance): ${Z_F}^{-1} = 1 + \frac{jf}{f_c}$. Since this admittance is a sum of admittances, this expression suggests

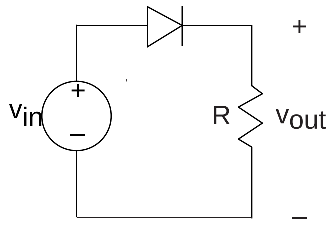


Figure 3.50

---

its application, and KVL is a statement about voltage drops around a closed path **regardless** of whether the elements are linear or not. Thus, for this simple circuit we have

$$\frac{v_{\text{out}}}{R} = I_0 \cdot \left(e^{\frac{q}{kT}(v_{\text{in}} - v_{\text{out}})} - 1\right) \tag{3.33}$$

This equation **cannot** be solved in closed form. We must understand what is going on from basic principles, using computational and graphical aids. As an approximation, when v_{in} is positive, current flows through the diode so long as the voltage v_{out} is smaller than v_{in} (so the diode is forward biased). If the source is negative or v_{out} "tries" to be bigger than v_{in}, the diode is reverse-biased, and the reverse-bias current flows through the diode. Thus, at this level of analysis, positive input voltages result in positive output voltages with negative ones resulting in $v_{\text{out}} = -(RI_0)$.

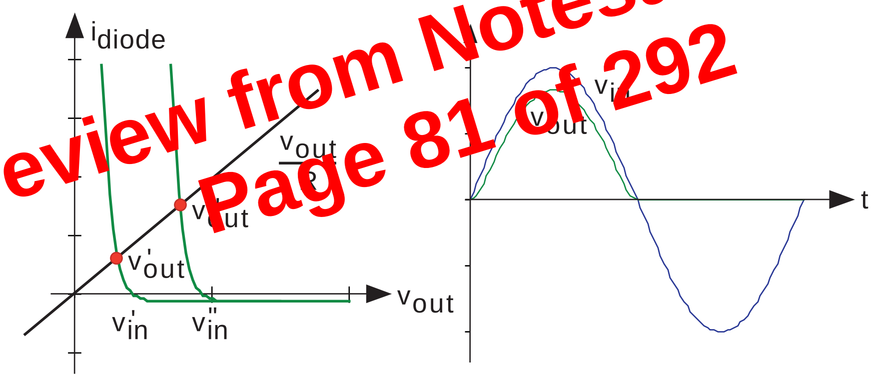
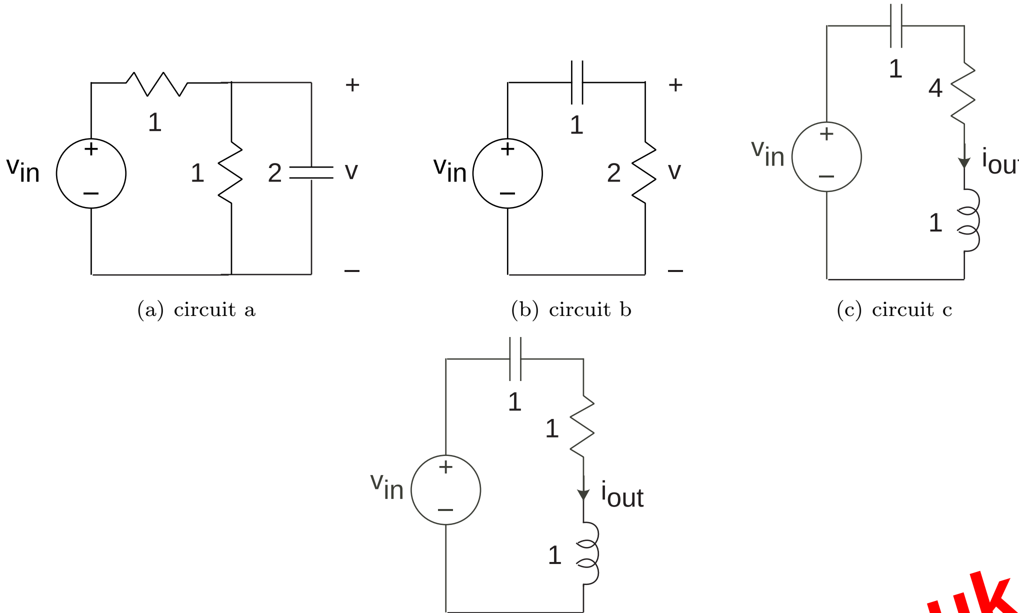


Figure 3.51

We need to detail the exponential nonlinearity to determine how the circuit distorts the input voltage waveform. We can of course numerically solve Figure 3.50 (diode circuit) to determine the output voltage when the input is a sinusoid. To learn more, let's express this equation graphically. We plot each term as a function of v_{out} for various values of the input voltage v_{in}; where they intersect gives us the output voltage. The left side, the current through the output resistor, does not vary itself with v_{in}, and thus we have a fixed straight line. As for the right side, which expresses the diode's **v-i** relation, the point at which the curve crosses the v_{out} axis gives us the value of v_{in}. Clearly, the two curves will always intersect just once for any value of v_{in}, and for positive v_{in} the intersection occurs at a value for v_{out} **smaller** than v_{in}. This reduction is smaller if the straight line has a shallower slope, which corresponds to using a bigger output resistor. For negative v_{in}, the diode is reverse-biased and the output voltage equals $-(RI_0)$.

What utility might this simple circuit have? The diode's nonlinearity cannot be escaped here, and the clearly evident distortion must have some practical application if the circuit were to be useful. This circuit, known as a **half-wave rectifier**, is present in virtually every AM radio **twice** and each serves very different functions! We'll learn what functions later.



(a) circuit a (b) circuit b (c) circuit c

(d) circuit d

Figure 3.61

Problem 3.13: Transfer Functions
Find the transfer function relating the complex amplitudes of the indicated variable and the source shown in Figure 3.61. Plot the magnitude and phase of the transfer function.

Problem 3.14: Using Impedances
Find the differential equation relating the indicated variable to the source(s) using impedances for each circuit shown in Figure 3.62.

Problem 3.15: Measurement Chaos
The following simple circuit (Figure 3.63) was constructed but the signal measurements were made haphazardly. When the source was $\sin(2\pi f_0 t)$, the current $i(t)$ equaled $\frac{\sqrt{2}}{3}\sin\left(2\pi f_0 t + \frac{\pi}{4}\right)$ and the voltage $v_2(t) = \frac{1}{3}\sin(2\pi f_0 t)$.

a) What is the voltage $v_1(t)$?
b) Find the impedances Z_1 and Z_2.
c) Construct these impedances from elementary circuit elements.

Problem 3.16: Transfer Functions
In the following circuit (Figure 3.64), the voltage source equals $v_{\text{in}}(t) = 10\sin\left(\frac{t}{2}\right)$.

a) Find the transfer function between the source and the indicated output voltage.
b) For the given source, find the output voltage.

Problem 3.17: A Simple Circuit
You are given this simple circuit (Figure 3.65).

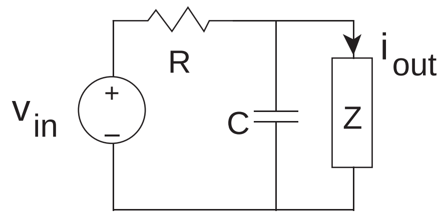


Figure 3.71

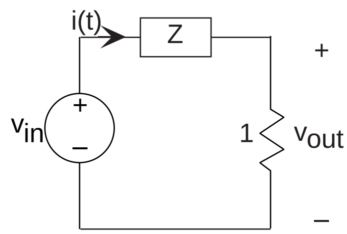


Figure 3.72

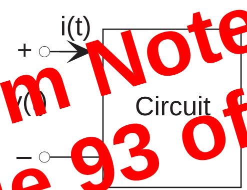


Figure 3.73

Problem 3.29: A Testing Circuit
The simple circuit here (Figure 3.72) was given on a test.

When the voltage source is $\sqrt{5}\sin(t)$, the current $i(t) = \sqrt{2}\cos\left(t - \arctan(2) - \frac{\pi}{4}\right)$.

a) What is voltage $v_{out}(t)$?
b) What is the impedance Z at the frequency of the source?

Problem 3.30: Black-Box Circuit
You are given a circuit (Figure 3.73) that has two terminals for attaching circuit elements.

When you attach a voltage source equaling $\sin(t)$ to the terminals, the current through the source equals $4\sin\left(t + \frac{\pi}{4}\right) - 2\sin(4t)$. When no source is attached (open-circuited terminals), the voltage across the terminals has the form $A\sin(4t + \phi)$.

a) What will the terminal current be when you replace the source by a short circuit?
b) If you were to build a circuit that was identical (from the viewpoint of the terminals) to the given one, what would your circuit be?
c) For your circuit, what are A and ϕ?

The two integrals are very similar, one equaling the negative of the other. The final expression becomes

$$\begin{aligned} b_k &= -\frac{2}{j2\pi k}\left((-1)^k - 1\right) \\ &= \begin{cases} \frac{2}{j\pi k} \text{ if } k \text{ odd} \\ 0 \text{ if } k \text{ even} \end{cases} \end{aligned} \tag{4.5}$$

Thus, the complex Fourier series for the square wave is

$$\operatorname{sq}(t) = \sum_{k \in \{\dots,-3,-1,1,3,\dots\}} \frac{2}{j\pi k} e^{+j\frac{2\pi kt}{T}} \tag{4.6}$$

Consequently, the square wave equals a sum of complex exponentials, but only those having frequencies equal to odd multiples of the fundamental frequency $\frac{1}{T}$. The coefficients decay slowly as the frequency index k increases. This index corresponds to the k-th harmonic of the signal's period.

A signal's Fourier series spectrum c_k has interesting properties.

Property 4.1:
If $s(t)$ is real, $c_k = c_{-k}^*$ (real-valued periodic signals have conjugate-symmetric spectra).

This result follows from the integral that calculates the c_k from the signal. Furthermore, this result means that $\operatorname{Re}[c_k] = \operatorname{Re}[c_{-k}]$: The real part of the Fourier coefficients for real-valued signals is even. Similarly, $\operatorname{Im}[c_k] = -(\operatorname{Im}[c_{-k}])$: The imaginary parts of the Fourier coefficients have odd symmetry. Consequently, if you are given the Fourier coefficients for positive indices and zero and are told the signal is real-valued, you can find the negative-indexed coefficients, hence the entire spectrum. This kind of symmetry, $c_k = c_{-k}^*$, is known as **conjugate symmetry**.

Property 4.2:
If $s(-t) = s(t)$, which says the signal has even symmetry about the origin, $c_{-k} = c_k$.

Given the previous property for real-valued signals, the Fourier coefficients of even signals are real-valued. A real-valued Fourier expansion amounts to an expansion in terms of only cosines, which is the simplest example of an even signal.

Property 4.3:
If $s(-t) = -s(t)$, which says the signal has odd symmetry, $c_{-k} = -c_k$.

Therefore, the Fourier coefficients are purely imaginary. The square wave is a great example of an odd-symmetric signal.

Property 4.4:
The spectral coefficients for a periodic signal delayed by τ, $s(t-\tau)$, are $c_k e^{-\frac{j2\pi k\tau}{T}}$, where c_k denotes the spectrum of $s(t)$. Delaying a signal by τ seconds results in a spectrum having a **linear phase shift** of $-\frac{2\pi k\tau}{T}$ in comparison to the spectrum of the un-delayed signal. Note that the spectral magnitude is unaffected. Showing this property is easy.
Proof:

$$\begin{aligned} \frac{1}{T}\int_0^T s(t-\tau) e^{-j\frac{2\pi kt}{T}}\,dt &= \frac{1}{T}\int_{-\tau}^{T-\tau} s(t) e^{-j\frac{2\pi k(t+\tau)}{T}}\,dt \\ &= \frac{1}{T} e^{-j\frac{2\pi k\tau}{T}} \int_{-\tau}^{T-\tau} s(t) e^{-j\frac{2\pi kt}{T}}\,dt \end{aligned} \tag{4.7}$$

Note that the range of integration extends over a period of the integrand. Consequently, it should not matter how we integrate over a period, which means that $\int_{-\tau}^{T-\tau} (\cdot)\,dt = \int_0^T (\cdot)\,dt$, and we have our result.

in the fundamental. Find an expression for the total harmonic distortion for any periodic signal. Is this calculation most easily performed in the time or frequency domain?

4.5 Fourier Series Approximation of Signals[8]

It is interesting to consider the sequence of signals that we obtain as we incorporate more terms into the Fourier series approximation of the half-wave rectified sine wave (Example 4.2). Define $s_K(t)$ to be the signal containing $K+1$ Fourier terms.

$$s_K(t) = a_0 + \sum_{k=1}^{K} a_k \cos\left(\frac{2\pi kt}{T}\right) + \sum_{k=1}^{K} b_k \sin\left(\frac{2\pi kt}{T}\right) \quad (4.23)$$

Figure 4.5 (Fourier Series spectrum of a half-wave rectified sine wave) shows how this sequence of signals portrays the signal more accurately as more terms are added.

We need to assess quantitatively the accuracy of the Fourier series approximation so that we can judge how rapidly the series approaches the signal. When we use a $K+1$-term series, the error—the difference between the signal and the $K+1$-term series—corresponds to the unused terms from the series.

$$\epsilon_K(t) = \sum_{k=K+1}^{\infty} a_k \cos\left(\frac{2\pi kt}{T}\right) + \sum_{k=K+1}^{\infty} b_k \sin\left(\frac{2\pi kt}{T}\right) \quad (4.24)$$

To find the rms error, we must square this expression and integrate it over a period. Again, the integral of most cross-terms is zero, leaving

$$\mathrm{rms}(\epsilon_K) = \sqrt{\frac{1}{2} \sum_{k=K+1}^{\infty} \left(a_k^2 + {b_k}^2\right)} \quad (4.25)$$

Figure 4.6 (Approximation error for a half-wave rectified sinusoid) shows how the error in the Fourier series for the half-wave rectified sinusoid decreases as more terms are incorporated. In particular, the use of four terms, as shown in the bottom plot of Figure 4.5 (Fourier Series spectrum of a half-wave rectified sine wave), has a rms error (relative to the rms value of the signal) of about 3%. The Fourier series in this case converges quickly to the signal.

We can look at Figure 4.7 (Power spectrum and approximation error for a square wave) to see the power spectrum and the rms approximation error for the square wave. Because the Fourier coefficients decay more slowly here than for the half-wave rectified sinusoid, the rms error is not decreasing quickly. Said another way, the square-wave's spectrum contains more power at higher frequencies than does the half-wave-rectified sinusoid. This difference between the two Fourier series results because the half-wave rectified sinusoid's Fourier coefficients are proportional to $\frac{1}{k^2}$ while those of the square wave are proportional to $\frac{1}{k}$. If fact, after 99 terms of the square wave's approximation, the error is bigger than 10 terms of the approximation for the half-wave rectified sinusoid. Mathematicians have shown that no signal has an rms approximation error that decays more slowly than it does for the square wave.

Exercise 4.8 *(Solution on p. 143.)*

Calculate the harmonic distortion for the square wave.

More than just decaying slowly, Fourier series approximation shown in Figure 4.8 (Fourier series approximation of a square wave) exhibits interesting behavior. Although the square wave's Fourier series requires more terms for a given representation accuracy, when comparing plots it is not clear that the two are equal. Does the Fourier series really equal the square wave at **all** values of t? In particular, at each step-change in the square wave, the Fourier series exhibits a peak followed by rapid oscillations. As more terms are added

[8] This content is available online at <http://cnx.org/content/m10687/2.9/>.

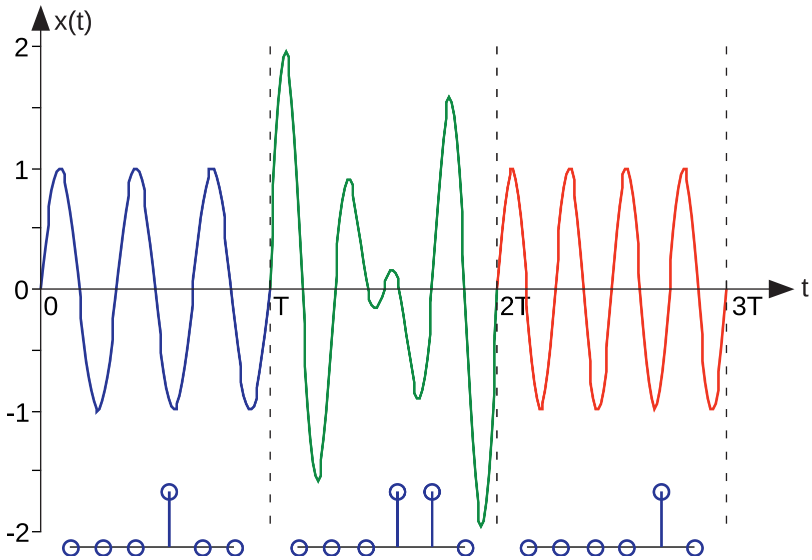


Figure 4.9: The encoding of signals via the Fourier spectrum is shown over three "periods." In this example, only the third and fourth harmonics are used, as shown by the spectral magnitudes corresponding to each T-second interval plotted below the waveforms. Can you determine the phase of the harmonics from the waveform?

---

a communications system. This approach represents a simplification of how modern modems represent text that they transmit over telephone lines.

4.7 Filtering Periodic Signals[12]

The Fourier series representation of a periodic signal makes it easy to determine how a linear, time-invariant filter reshapes such signals **in general**. The fundamental property of a linear system is that its input-output relation obeys superposition: $L(a_1 s_1(t) + a_2 s_2(t)) = a_1 L(s_1(t)) + a_2 L(s_2(t))$. Because the Fourier series represents a periodic signal as a linear combination of complex exponentials, we can exploit the superposition property. Furthermore, we found for linear circuits that their output to a complex exponential input is just the frequency response evaluated at the signal's frequency times the complex exponential. Said mathematically, if $x(t) = e^{j\frac{2\pi kt}{T}}$, then the output $y(t) = H\left(\frac{k}{T}\right) e^{j\frac{2\pi kt}{T}}$ because $f = \frac{k}{T}$. Thus, if $x(t)$ is periodic thereby having a Fourier series, a linear circuit's output to this signal will be the superposition of the output to each component.

$$y(t) = \sum_{k=-\infty}^{\infty} c_k H\left(\frac{k}{T}\right) e^{j\frac{2\pi kt}{T}} \tag{4.27}$$

Thus, the output has a Fourier series, which means that it too is periodic. Its Fourier coefficients equal $c_k H\left(\frac{k}{T}\right)$. **To obtain the spectrum of the output, we simply multiply the input spectrum by the frequency response**. The circuit modifies the magnitude and phase of each Fourier coefficient. Note especially that while the Fourier coefficients do not depend on the signal's period, the circuit's transfer function does depend on frequency, which means that the circuit's output will differ as the period varies.

Example 4.3

The periodic pulse signal shown on the left above serves as the input to a RC-circuit that has the transfer function (calculated elsewhere (Figure 3.31: Magnitude and phase of the transfer function))

$$H(f) = \frac{1}{1 + j2\pi fRC} \tag{4.28}$$

---

[12]This content is available online at <http://cnx.org/content/m0044/2.10/>.

$s(t)$	$S(f)$
$e^{-at}u(t)$	$\frac{1}{j2\pi f+a}$
$e^{-a\vert t\vert}$	$\frac{2a}{4\pi^2f^2+a^2}$
$p(t)=\begin{cases}1 \text{ if } \vert t\vert<\frac{\Delta}{2}\\ 0 \text{ if } \vert t\vert>\frac{\Delta}{2}\end{cases}$	$\frac{\sin(\pi f\Delta)}{\pi f}$
$\frac{\sin(2\pi Wt)}{\pi t}$	$S(f)=\begin{cases}1 \text{ if } \vert f\vert<W\\ 0 \text{ if } \vert f\vert>W\end{cases}$

Table 4.1

	Time-Domain	**Frequency Domain**
Linearity	$a_1s_1(t)+a_2s_2(t)$	$a_1S_1(f)+a_2S_2(f)$
Conjugate Symmetry	$s(t)\in\mathbb{R}$	$S(f)=S(-f)^*$
Even Symmetry	$s(t)=s(-t)$	$S(f)=S(-f)$
Odd Symmetry	$s(t)=-(s(-t))$	$S(f)=-(S(-f))$
Scale Change	$s(at)$	$\frac{1}{\vert a\vert}S\left(\frac{f}{a}\right)$
Time Delay	$s(t-\tau)$	$e^{-j2\pi f\tau}S(f)$
Complex Modulation	$e^{j2\pi f_0t}s(t)$	$S(f-f_0)$
Amplitude Modulation by Cosine	$s(t)\cos(2\pi f_0t)$	$\frac{S(f-f_0)+S(f+f_0)}{2}$
Amplitude Modulation by Sine	$s(t)\sin(2\pi f_0t)$	$\frac{S(f-f_0)-S(f+f_0)}{2j}$
Differentiation	$\frac{d}{dt}s(t)$	$j2\pi fS(f)$
Integration	$\int_{-\infty}^{t}s(\alpha)\,d\alpha$	$\frac{1}{j2\pi f}S(f)$ if $S(0)=0$
Multiplication by t	$ts(t)$	$\frac{1}{-j2\pi}\frac{d}{df}S(f)$
Area	$\int_{-\infty}^{\infty}s(t)\,dt$	$S(0)$
Value at Origin	$s(0)$	$\int_{-\infty}^{\infty}S(f)\,df$
Parseval's Theorem	$\int_{-\infty}^{\infty}\vert s(t)\vert^2dt$	$\int_{-\infty}^{\infty}\vert S(f)\vert^2df$

Table 4.2

Example 4.5
In communications, a very important operation on a signal $s(t)$ is to **amplitude modulate** it. Using this operation more as an example rather than elaborating the communications aspects here, we want to compute the Fourier transform — the spectrum — of

$$(1+s(t))\cos(2\pi f_ct)$$

Thus,

$$(1+s(t))\cos(2\pi f_ct)=\cos(2\pi f_ct)+s(t)\cos(2\pi f_ct)$$

f) $s(t)$ given by the depicted waveform (Figure 4.18).

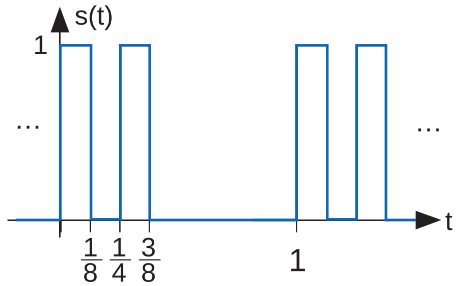


Figure 4.18

Problem 4.2: Fourier Series

Find the Fourier series representation for the following periodic signals (Figure 4.19). For the third signal, find the complex Fourier series for the triangle wave **without** performing the usual Fourier integrals. Hint: How is this signal related to one for which you already have the series?

Problem 4.3: Phase Distortion

We can learn about phase distortion by returning to circuits and investigate the following circuit (Figure 4.20).

a) Find this filter's transfer function.
b) Find the magnitude and phase of this transfer function. How would you characterize this circuit?
c) Let $v_{in}(t)$ be a square-wave of period T. What is the Fourier series for the output voltage?
d) Use Matlab to find the output's waveform for the cases $T = 0.01$ and $T = 2$. What value of T delineates the two kinds of results you found? The software in `fourier2.m` might be useful.
e) Instead of the depicted circuit, the square wave is passed through a system that delays its input, which applies a linear phase shift to the signal's spectrum. Let the delay τ be $\frac{T}{4}$. Use the transfer function of a delay to compute using Matlab the Fourier series of the output. Show that the square wave is indeed delayed.

Problem 4.4: Approximating Periodic Signals

Often, we want to approximate a reference signal by a somewhat simpler signal. To assess the quality of an approximation, the most frequently used error measure is the mean-squared error. For a periodic signal $s(t)$,

$$\epsilon^2 = \frac{1}{T}\int_0^T (s(t) - \tilde{s}(t))^2 dt$$

where $s(t)$ is the reference signal and $\tilde{s}(t)$ its approximation. One convenient way of finding approximations for periodic signals is to truncate their Fourier series.

$$\tilde{s}(t) = \sum_{k=-K}^{K} c_k e^{j\frac{2\pi k}{T}t}$$

The point of this problem is to analyze whether this approach is the best (*i.e.*, always minimizes the mean-squared error).

c) Find the Fourier series for the depicted signal (Figure 4.21). Use Matlab to find the truncated approximation and best approximation involving two terms. Plot the mean-squared error as a function of K for both approximations.

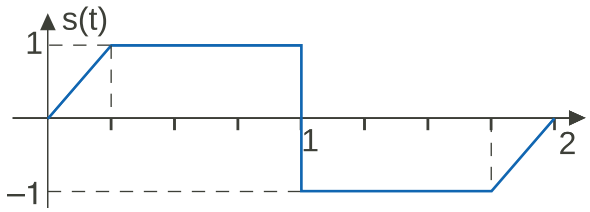


Figure 4.21

Problem 4.5: Long, Hot Days
The daily temperature is a consequence of several effects, one of them being the sun's heating. If this were the dominant effect, then daily temperatures would be proportional to the number of daylight hours. The plot (Figure 4.22) shows that the average daily high temperature does **not** behave that way.

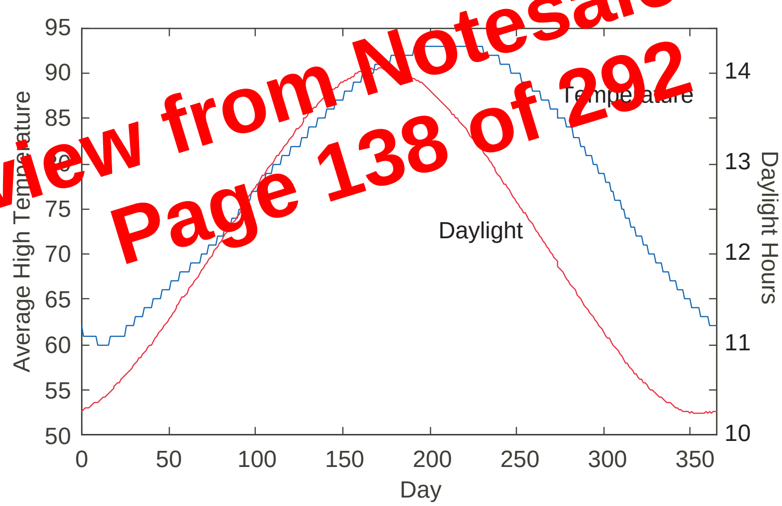


Figure 4.22

In this problem, we want to understand the temperature component of our environment using Fourier series and linear system theory. The file `temperature.mat` contains these data (daylight hours in the first row, corresponding average daily highs in the second) for Houston, Texas.

a) Let the length of day serve as the sole input to a system having an output equal to the average daily temperature. Examining the plots of input and output, would you say that the system is linear or not? How did you reach you conclusion?
b) Find the first five terms ($c_0, \dots, c_4$) of the complex Fourier series for each signal.
c) What is the harmonic distortion in the two signals? Exclude c_0 from this calculation.
d) Because the harmonic distortion is small, let's concentrate only on the first harmonic. What is the phase shift between input and output signals?

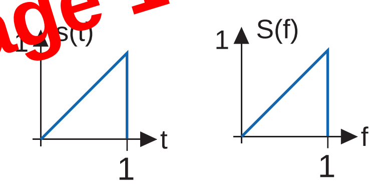
e) Find the transfer function of the simplest possible linear model that would describe the data. Characterize and interpret the structure of this model. In particular, give a physical explanation for the phase shift.
f) Predict what the output would be if the model had no phase shift. Would days be hotter? If so, by how much?

Problem 4.6: Fourier Transform Pairs
Find the Fourier or inverse Fourier transform of the following.

a) $x(t) = e^{-a|t|}$
b) $x(t) = te^{-at}u(t)$
c) $X(f) = \begin{cases} 1 & |f| < W \\ 0 & |f| > W \end{cases}$
d) $x(t) = e^{-at}\cos(2\pi f_0 t)\,u(t)$

Problem 4.7: Duality in Fourier Transforms
"Duality" means that the Fourier transform and the inverse Fourier transform are very similar. Consequently, the waveform $s(t)$ in the time domain and the spectrum $s(f)$ have a Fourier transform and an inverse Fourier transform, respectively, that are very similar.

a) Calculate the Fourier transform of the signal shown below (Figure 4.23(a)).
b) Calculate the inverse Fourier transform of the spectrum shown below (Figure 4.23(b)).
c) How are these answers related? What is the general relationship between the Fourier transform of $s(t)$ and the inverse transform of $s(f)$?



(a) (b)

Figure 4.23

Problem 4.8: Spectra of Pulse Sequences
Pulse sequences occur often in digital communication and in other fields as well. What are their spectral properties?

a) Calculate the Fourier transform of the single pulse shown below (Figure 4.24(a)).
b) Calculate the Fourier transform of the two-pulse sequence shown below (Figure 4.24(b)).
c) Calculate the Fourier transform for the **ten**-pulse sequence shown in below (Figure 4.24(c)). You should look for a general expression that holds for sequences of any length.
d) Using Matlab, plot the magnitudes of the three spectra. Describe how the spectra change as the number of repeated pulses increases.

Chapter 5

Digital Signal Processing

5.1 Introduction to Digital Signal Processing[1]

Not only do we have analog signals — signals that are real- or complex-valued functions of a continuous variable such as time or space — we can define **digital** ones as well. Digital signals are **sequences**, functions defined only for the integers. We thus use the notation $s(n)$ to denote a discrete-time one-dimensional signal such as a digital music recording and $s(m,n)$ for a discrete-"time" two-dimensional signal like a photo taken with a digital camera. Sequences are fundamentally different than continuous-time signals. For example, continuity has no meaning for sequences.

Despite such fundamental differences, the theory underlying digital signal processing mirrors that for analog signals: Fourier transforms, linear filtering, and linear systems parallel what previous chapters described. These similarities make it easy to understand the definitions and why we need them, but the similarities should not be construed as "analog wannabes." We will discover that digital signal processing is **not** an approximation to analog processing. We must explicitly worry about the fidelity of converting analog signals into digital ones. The music stored on CDs, the speech sent over digital cellular telephones, and the video carried by digital television all evidence that analog signals can be accurately converted to digital ones and back again.

The key reason why digital signal processing systems have a technological advantage today is the **computer**: computations, like the Fourier transform, can be performed quickly enough to be calculated as the signal is produced,[2] and programmability means that the signal processing system can be easily changed. This flexibility has obvious appeal, and has been widely accepted in the marketplace. Programmability means that we can perform signal processing operations impossible with analog systems (circuits). We will also discover that digital systems enjoy an **algorithmic** advantage that contributes to rapid processing speeds: Computations can be restructured in non-obvious ways to speed the processing. This flexibility comes at a price, a consequence of how computers work. How do computers perform signal processing?

5.2 Introduction to Computer Organization[3]

5.2.1 Computer Architecture

To understand digital signal processing systems, we must understand a little about how computers compute. The modern definition of a **computer** is an electronic device that performs calculations on data, presenting the results to humans or other computers in a variety of (hopefully useful) ways.

[1] This content is available online at <http://cnx.org/content/m10781/2.3/>.

[2] Taking a systems viewpoint for the moment, a system that produces its output as rapidly as the input arises is said to be a **real-time** system. All analog systems operate in real time; digital ones that depend on a computer to perform system computations may or may not work in real time. Clearly, we need real-time signal processing systems. Only recently have computers become fast enough to meet real-time requirements while performing non-trivial signal processing.

[3] This content is available online at <http://cnx.org/content/m10263/2.28/>.

5.5.5 Unit Step

The **unit step** in discrete-time is well-defined at the origin, as opposed to the situation with analog signals.

$$u(n) = \begin{cases} 1 & n \geq 0 \\ 0 & n < 0 \end{cases} \quad (5.16)$$

5.5.6 Symbolic Signals

An interesting aspect of discrete-time signals is that their values do not need to be real numbers. We do have real-valued discrete-time signals like the sinusoid, but we also have signals that denote the sequence of characters typed on the keyboard. Such characters certainly aren't real numbers, and as a collection of possible signal values, they have little mathematical structure other than that they are members of a set. More formally, each element of the **symbolic-valued** signal $s(n)$ takes on one of the values $\{a_1, \dots, a_K\}$ which comprise the **alphabet** A. This technical terminology does not mean we restrict symbols to being members of the English or Greek alphabet. They could represent keyboard characters, bytes (8-bit quantities), integers that convey daily temperature. Whether controlled by software or not, discrete-time systems are ultimately constructed from digital circuits, which consist **entirely** of analog circuit elements. Furthermore, the transmission and reception of discrete-time signals, like e-mail, is accomplished with analog signals and systems. Understanding how discrete-time and analog signals and systems intertwine is perhaps the main goal of this course.

5.5.7 Discrete-Time Systems

Discrete-time systems can act on discrete-time signals in ways similar to those found in analog signals and systems. Because of the role of software in discrete-time systems, many more different systems can be envisioned and "constructed" with programs than can be with analog signals. In fact, a special class of analog signals can be converted into discrete-time signals, processed with software, and converted back into an analog signal, all without the incursion of error. For such signals, systems can be easily produced in software, with equivalent analog realizations difficult, if not impossible, to design.

5.6 Discrete-Time Fourier Transform (DTFT)[15]

The Fourier transform of the discrete-time signal $s(n)$ is defined to be

$$S\left(e^{j2\pi f}\right) = \sum_{n=-\infty}^{\infty} s(n)\, e^{-j2\pi f n} \quad (5.17)$$

Frequency here has no units. As should be expected, this definition is linear, with the transform of a sum of signals equaling the sum of their transforms. Real-valued signals have conjugate-symmetric spectra: $S\left(e^{-j2\pi f}\right) = S^*\left(e^{j2\pi f}\right)$.

Exercise 5.11 ***(Solution on p. 193.)***

A special property of the discrete-time Fourier transform is that it is periodic with period one: $S\left(e^{j2\pi(f+1)}\right) = S\left(e^{j2\pi f}\right)$. Derive this property from the definition of the DTFT.

Because of this periodicity, we need only plot the spectrum over one period to understand completely the spectrum's structure; typically, we plot the spectrum over the frequency range $\left[-\frac{1}{2}, \frac{1}{2}\right]$. When the signal is real-valued, we can further simplify our plotting chores by showing the spectrum only over $\left[0, \frac{1}{2}\right]$; the spectrum at negative frequencies can be derived from positive-frequency spectral values.

When we obtain the discrete-time signal via sampling an analog signal, the Nyquist frequency (p. 151) corresponds to the discrete-time frequency $\frac{1}{2}$. To show this, note that a sinusoid having a frequency equal to the Nyquist frequency $\frac{1}{2T_s}$ has a sampled waveform that equals

$$\cos\left(2\pi \frac{1}{2Ts} nTs\right) = \cos(\pi n) = (-1)^n$$

[15] This content is available online at <http://cnx.org/content/m10247/2.31/>.

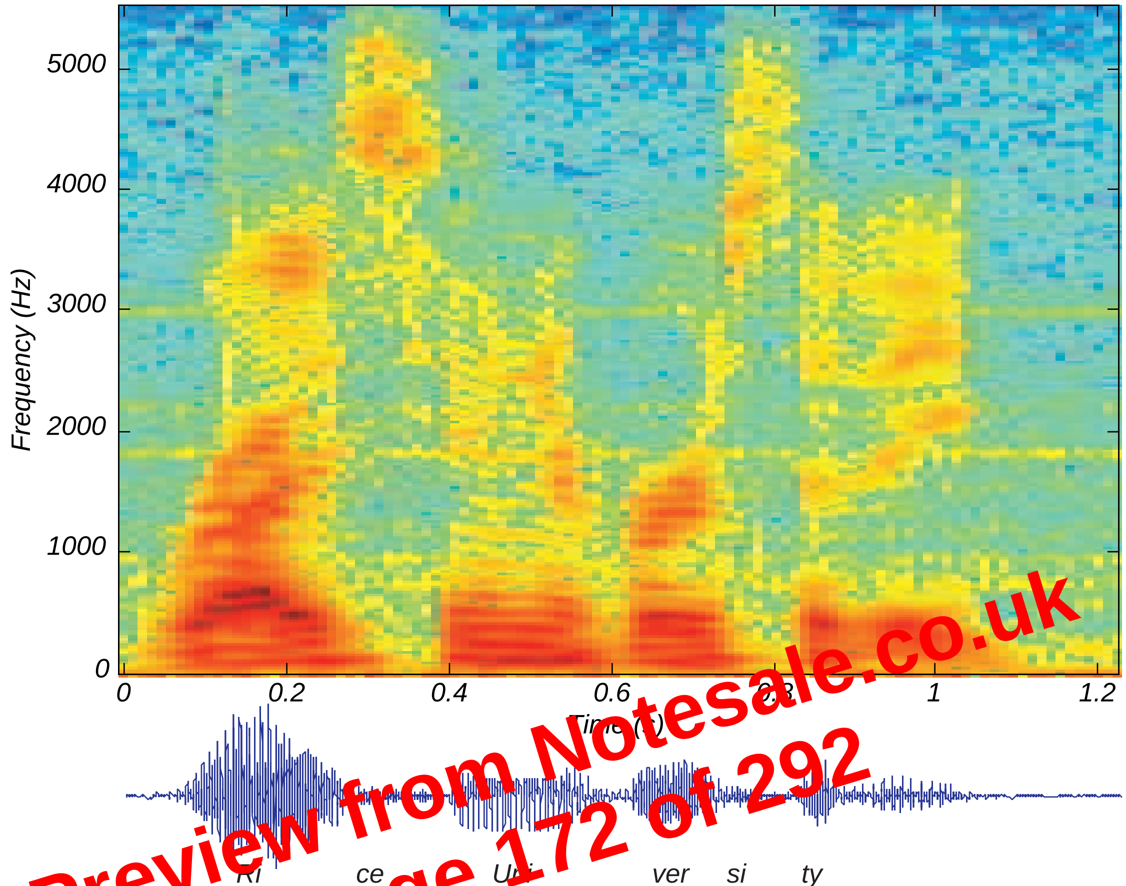


Figure 5.14

in bps (bits per second)? Assuming the computer storage is organized in terms of bytes (8-bit quantities), how many bytes of computer memory does the speech consume?

The resulting discrete-time signal, shown in the bottom of Figure 5.14 (speech spectrogram), clearly changes its character with time. To display these spectral changes, the long signal was sectioned into **frames**: comparatively short, contiguous groups of samples. Conceptually, a Fourier transform of each frame is calculated using the FFT. Each frame is not so long that significant signal variations are retained within a frame, but not so short that we lose the signal's spectral character. Roughly speaking, the speech signal's spectrum is evaluated over successive time segments and stacked side by side so that the x-axis corresponds to time and the y-axis frequency, with color indicating the spectral amplitude.

An important detail emerges when we examine each framed signal (Figure 5.15 (Spectrogram Hanning vs. Rectangular)). At the frame's edges, the signal may change very abruptly, a feature not present in the original signal. A transform of such a segment reveals a curious oscillation in the spectrum, an artifact directly related to this sharp amplitude change. A better way to frame signals for spectrograms is to apply a **window**: Shape the signal values within a frame so that the signal decays gracefully as it nears the edges. This shaping is accomplished by multiplying the framed signal by the sequence $w(n)$. In sectioning the signal, we essentially applied a rectangular window: $w(n) = 1$, $0 \leq n \leq N-1$. A much more graceful window is the **Hanning window**; it has the cosine shape $w(n) = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N}\right)\right)$. As shown in Figure 5.15 (Spectrogram Hanning vs. Rectangular), this shaping greatly reduces spurious oscillations in each frame's spectrum. Considering the spectrum of the Hanning windowed frame, we find that the oscillations resulting from applying the rectangular window obscured a formant (the one located at a little more than half the

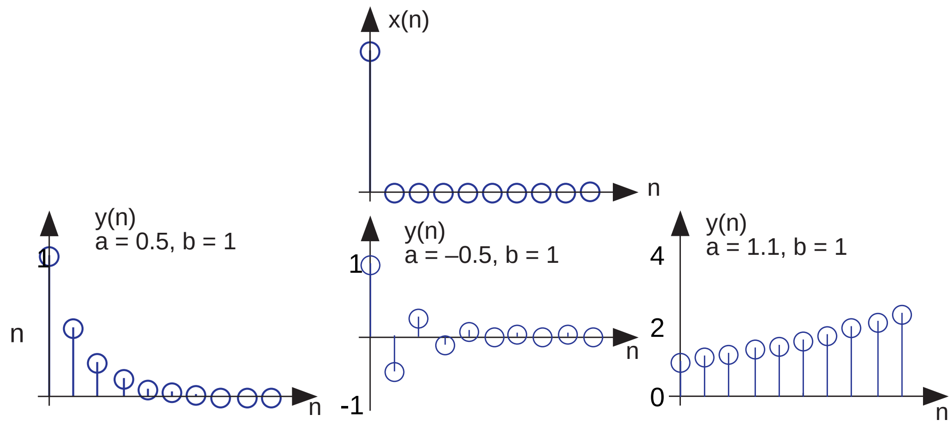


Figure 5.18: The input to the simple example system, a unit sample, is shown at the top, with the outputs for several system parameter values shown below.

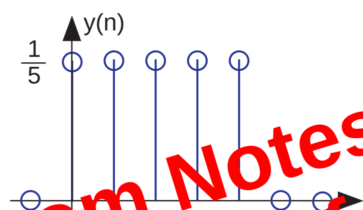


Figure 5.19: The plot shows the unit-sample response of a length-5 boxcar filter.

Positive values of a are used in population models to describe how population size increases over time. Here, n might correspond to generation. The difference equation says that the number in the next generation is some multiple of the previous one. If this multiple is less than one, the population becomes extinct; if greater than one, the population flourishes. The same difference equation also describes the effect of compound interest on deposits. Here, n indexes the times at which compounding occurs (daily, monthly, etc.), a equals the compound interest rate plus one, and $b = 1$ (the bank provides no gain). In signal processing applications, we typically require that the output remain bounded for any input. For our example, that means that we restrict $|a| < 1$ and choose values for it and the gain according to the application.

Exercise 5.24 *(Solution on p. 194.)*

Note that the difference equation (5.42),

$$y(n) = a_1 y(n-1) + \cdots + a_p y(n-p) + b_0 x(n) + b_1 x(n-1) + \cdots + b_q x(n-q)$$

does not involve terms like $y(n+1)$ or $x(n+1)$ on the equation's right side. Can such terms also be included? Why or why not?

Example 5.4

A somewhat different system has no "a" coefficients. Consider the difference equation

$$y(n) = \frac{1}{q}\left(x(n) + \cdots + x(n-q+1)\right) \quad (5.46)$$

Because this system's output depends only on current and previous input values, we need not be concerned with initial conditions. When the input is a unit-sample, the output equals $\frac{1}{q}$ for

Example 5.6
The length-q boxcar filter (difference equation found in a previous example (Example 5.4)) has the frequency response

$$H\left(e^{j2\pi f}\right) = \frac{1}{q}\sum_{m=0}^{q-1} e^{-j2\pi f m} \tag{5.51}$$

This expression amounts to the Fourier transform of the boxcar signal (Figure 5.19). There we found that this frequency response has a magnitude equal to the absolute value of $\text{dsinc}\left(\pi f\right)$; see the length-10 filter's frequency response (Figure 5.11: Spectrum of length-ten pulse). We see that boxcar filters–length-q signal averagers–have a lowpass behavior, having a cutoff frequency of $\frac{1}{q}$.

Exercise 5.25 ***(Solution on p. 194.)***
Suppose we multiply the boxcar filter's coefficients by a sinusoid: $b_m = \frac{1}{q}\cos\left(2\pi f_0 m\right)$ Use Fourier transform properties to determine the transfer function. How would you characterize this system: Does it act like a filter? If so, what kind of filter and how do you control its characteristics with the filter's coefficients?

These examples illustrate the point that systems described (and implemented) by difference equations serve as filters for discrete-time signals. The filter's **order** is given by the number p of denominator coefficients in the transfer function (if the system is IIR) or by the number q of numerator coefficients if the filter is FIR. When a system's transfer function has both terms, the system is usually IIR, and its order equals p regardless of q. By selecting the coefficients and filter type, filters having virtually any frequency response desired can be designed. This design flexibility can't be found in analog systems. In the next section, we detail how analog signals can be filtered by computers, offering a much greater range of filtering possibilities than is possible with circuits.

5.14 Filtering in the Frequency Domain[25]

Because we are interested in actual computations rather than analytic calculations, we must consider the details of the discrete Fourier transform. To compute the length-N DFT, we assume that the signal has a duration less than or equal to N. Because frequency responses have an explicit frequency-domain specification (5.47) in terms of filter coefficients, we don't have a direct handle on which signal has a Fourier transform equaling a given frequency response. Finding this signal is quite easy. First of all, note that the discrete-time Fourier transform of a unit sample equals one for all frequencies. Because the input and output of linear, shift-invariant systems are related to each other by $Y\left(e^{j2\pi f}\right) = H\left(e^{j2\pi f}\right)X\left(e^{j2\pi f}\right)$, **a unit-sample input, which has $X\left(e^{j2\pi f}\right) = 1$, results in the output's Fourier transform equaling the system's transfer function**.

Exercise 5.26 ***(Solution on p. 194.)***
This statement is a very important result. Derive it yourself.

In the time-domain, the output for a unit-sample input is known as the system's **unit-sample response**, and is denoted by $h\left(n\right)$. Combining the frequency-domain and time-domain interpretations of a linear, shift-invariant system's unit-sample response, we have that $h\left(n\right)$ and the transfer function are Fourier transform pairs **in terms of the discrete-time Fourier transform**.

$$\left(h\left(n\right) \leftrightarrow H\left(e^{j2\pi f}\right)\right) \tag{5.52}$$

Returning to the issue of how to use the DFT to perform filtering, we can analytically specify the frequency response, and derive the corresponding length-N DFT by sampling the frequency response.

$$H\left(k\right) = H\left(e^{j2\pi k/N}\right) \quad , \quad k = \{0, \ldots, N-1\} \tag{5.53}$$

Computing the inverse DFT yields a length-N signal **no matter what the actual duration of the unit-sample response might be**. If the unit-sample response has a duration less than or equal to N (it's a FIR

[25] This content is available online at <http://cnx.org/content/m10257/2.17/>.

filter), computing the inverse DFT of the sampled frequency response indeed yields the unit-sample response. If, however, the duration exceeds N, errors are encountered. The nature of these errors is easily explained by appealing to the Sampling Theorem. By sampling in the frequency domain, we have the potential for aliasing in the time domain (sampling in one domain, be it time or frequency, can result in aliasing in the other) unless we sample fast enough. Here, the duration of the unit-sample response determines the minimal sampling rate that prevents aliasing. For FIR systems — they by definition have finite-duration unit sample responses — the number of required DFT samples equals the unit-sample response's duration: $N \geq q$.

Exercise 5.27 *(Solution on p. 194.)*

Derive the minimal DFT length for a length-q unit-sample response using the Sampling Theorem. Because sampling in the frequency domain causes repetitions of the unit-sample response in the time domain, sketch the time-domain result for various choices of the DFT length N.

Exercise 5.28 *(Solution on p. 194.)*

Express the unit-sample response of a FIR filter in terms of difference equation coefficients. Note that the corresponding question for IIR filters is far more difficult to answer: Consider the example (Example 5.5).

For IIR systems, we cannot use the DFT to find the system's unit-sample response: aliasing of the unit-sample response will **always** occur. Consequently, we can only implement an IIR filter accurately in the time domain with the system's difference equation. **Frequency-domain implementations are restricted to FIR filters.**

Another issue arises in frequency-domain filtering that is related to time-domain aliasing, this time when we consider the output. Assume we have an input signal having duration N_x that we pass through a FIR filter having a length-$q+1$ unit-sample response. What is the duration of the output signal? The difference equation for this filter is

$$y(n) = b_0 x(n) + \cdots + b_q x(n-q) \tag{5.54}$$

This equation says that the output depends on current and past input values, with the input value q samples previous defining the extent of the filter's **memory** of past input values. For example, the output at index N_x depends on $x(N_x)$ (which equals zero), $x(N_x - 1)$, through $x(N_x - q)$. Thus, the output returns to zero only after the last input value passes through the filter's memory. As the input signal's last value occurs at index $N_x - 1$, the last nonzero output value occurs when $n - q = N_x - 1$ or $n = q + N_x - 1$. Thus, the output signal's duration equals $q + N_x$.

Exercise 5.29 *(Solution on p. 194.)*

In words, we express this result as "The output's duration equals the input's duration plus the filter's duration minus one." Demonstrate the accuracy of this statement.

The main theme of this result is that a filter's output extends longer than either its input or its unit-sample response. Thus, to avoid aliasing when we use DFTs, the dominant factor is not the duration of input or of the unit-sample response, but of the output. Thus, the number of values at which we must evaluate the frequency response's DFT must be at least $q + N_x$ and we must compute the same length DFT of the input. To accommodate a shorter signal than DFT length, we simply **zero-pad** the input: Ensure that for indices extending beyond the signal's duration that the signal is zero. Frequency-domain filtering, diagrammed in Figure 5.20, is accomplished by storing the filter's frequency response as the DFT $H(k)$, computing the input's DFT $X(k)$, multiplying them to create the output's DFT $Y(k) = H(k)X(k)$, and computing the inverse DFT of the result to yield $y(n)$.

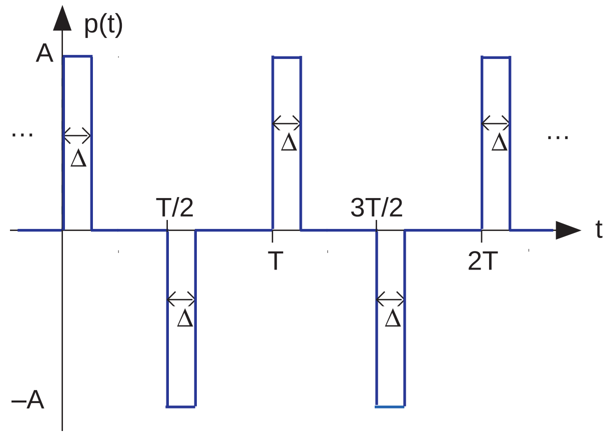


Figure 5.25

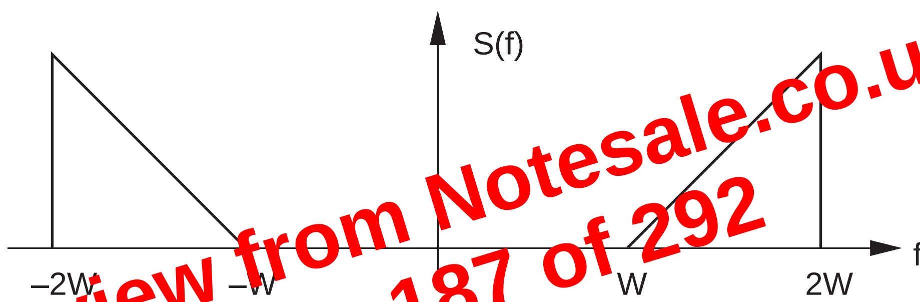


Figure 5.27

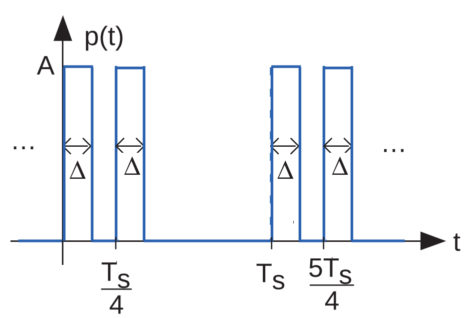


Figure 5.26

a) Find the Fourier spectrum of this signal.
b) Will this scheme work? If so, how should T_S be related to the signal's bandwidth? If not, why not?

Problem 5.4: Bandpass Sampling
The signal $s(t)$ has the indicated spectrum.

a) What is the minimum sampling rate for this signal suggested by the Sampling Theorem?

Problem 5.34: Signal Compression

Because of the slowness of the Internet, lossy signal compression becomes important if you want signals to be received quickly. An enterprising ELEC 241 student has proposed a scheme based on frequency-domain processing. First of all, he would section the signal into length-N blocks, and compute its N-point DFT. He then would discard (zero the spectrum) at **half** of the frequencies, quantize them to b-bits, and send these over the network. The receiver would assemble the transmitted spectrum and compute the inverse DFT, thus reconstituting an N-point block.

a) At what frequencies should the spectrum be zeroed to minimize the error in this lossy compression scheme?
b) The nominal way to represent a signal digitally is to use simple b-bit quantization of the time-domain waveform. How long should a section be in the proposed scheme so that the required number of bits/sample is smaller than that nominally required?
c) Assuming that effective compression can be achieved, would the proposed scheme yield satisfactory results?

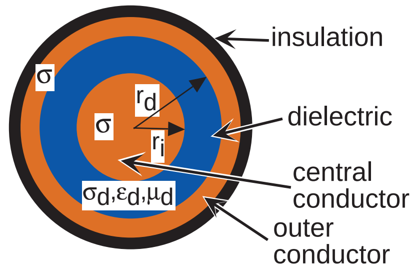


Figure 6.1: Coaxial cable consists of one conductor wrapped around the central conductor. This type of cable supports broader bandwidth signals than twisted pair, and finds use in cable television and Ethernet.

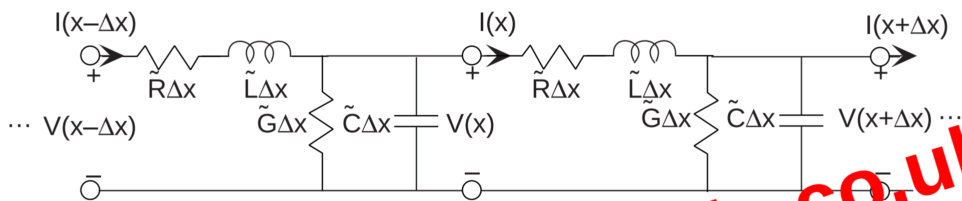


Figure 6.2: The so-called distributed parameter model for two-wire cables has the depicted circuit model structure. Element values depend on geometry and the properties of materials used to construct the transmission line.

Consequently, most wireline channels today essentially consist of pairs of conducting wires (see Figure 6.1: Coaxial Cable Cross-section), and the transmitter applies a message-related voltage across the pair. How these pairs of wires are physically configured greatly affects their transmission characteristics. One example is **twisted pair**, wherein the wires are wrapped about each other. Telephone cables are one example of a twisted pair channel. Another is **coaxial cable**, where a concentric conductor surrounds a central wire with a dielectric material in between. Coaxial cable, fondly called "co-ax" by engineers, is what Ethernet uses as its channel. In either case, wireline channels form a dedicated circuit between transmitter and receiver. As we shall find subsequently, several transmissions can share the circuit by amplitude modulation techniques; commercial cable TV is an example. These information-carrying circuits are designed so that interference from nearby electromagnetic sources is minimized. Thus, by the time signals arrive at the receiver, they are relatively interference- and noise-free.

Both twisted pair and co-ax are examples of **transmission lines**, which all have the circuit model shown in Figure 6.2 (Circuit Model for a Transmission Line) for an infinitesimally small length. This circuit model arises from solving Maxwell's equations for the particular transmission line geometry. The series resistance comes from the conductor used in the wires and from the conductor's geometry. The inductance and the capacitance derive from transmission line geometry, and the parallel conductance from the medium between the wire pair. Note that all the circuit elements have values expressed by the product of a constant times a length; this notation represents that element values here have per-unit-length units. For example, the series resistance $\widetilde{R}$ has units of ohms/meter. For coaxial cable, the element values depend on the inner conductor's radius r_i, the outer radius of the dielectric r_d, the conductivity of the conductors σ, and the conductivity where G is the gravitational constant and M the earth's mass. Calculations yield $R = 42200\text{km}$, which corresponds to an altitude of 35700km. This altitude greatly exceeds that of the ionosphere, requiring satellite transmitters to use frequencies that pass through it. Of great importance in satellite communications is the transmission delay. The time for electromagnetic fields to propagate to a geosynchronous satellite and return is 0.24 s, a significant delay.

Exercise 6.6 ***(Solution on p. 255.)***

In addition to delay, the propagation attenuation encountered in satellite communication far exceeds what occurs in ionospheric-mirror based communication. Calculate the attenuation incurred by radiation going to the satellite (one-way loss) with that encountered by Marconi (total going up and down). Note that the attenuation calculation in the ionospheric case, assuming the ionosphere acts like a perfect mirror, is not a straightforward application of the propagation loss formula (p. 201).

6.8 Noise and Interference[8]

We have mentioned that communications are, to varying degrees, subject to interference and noise. It's time to be more precise about what these quantities are and how they differ.

Interference represents man-made signals. Telephone lines are subject to power-line interference (in the United States a distorted 60 Hz sinusoid). Cellular telephone channels are subject to adjacent-cell phone conversations using the same signal frequency. The problem with such interference is that it occupies the same frequency band as the desired communication signal, and has a similar structure.

Exercise 6.7 ***(Solution on p. 255.)***

Suppose interference occupied a different frequency band; how would the receiver remove it?

We use the notation $i(t)$ to represent interference. Because interference has man-made structure, we can write an explicit expression for it that may contain some unknown aspects (how large it is, for example).

Noise signals have little structure and arise from both human and natural sources. Satellite channels are subject to deep space noise arising from electromagnetic radiation pervasive in the galaxy. Thermal noise plagues **all** electronic circuits that contain resistors. Thus, in receiving small amplitude signals, receiver amplifiers will most certainly add noise as they boost the signal's amplitude. All channels are subject to noise, and we need a way of describing such signals despite the fact we can't write a formula for the noise signal like we can for interference. The most widely used noise model is **white noise**. It is defined entirely by its frequency-domain characteristics.

- White noise has constant power at all frequencies.
- At each frequency, the phase of the noise spectrum is totally uncertain: It can be any value in between 0 and 2π, and its value at any frequency is unrelated to the phase at any other frequency.
- When noise signals arising from two different sources add, the resultant noise signal has a power equal to the sum of the component powers.

Because of the emphasis here on frequency-domain power, we are led to define the **power spectrum**. Because of Parseval's Theorem[9], we define the power spectrum $P_s(f)$ of a non-noise signal $s(t)$ to be the magnitude-squared of its Fourier transform.

$$P_s(f) \equiv |S(f)|^2 \tag{6.21}$$

Integrating the power spectrum over any range of frequencies equals the power the signal contains in that band. Because signals **must** have negative frequency components that mirror positive frequency ones, we routinely calculate the power in a spectral band as the integral over positive frequencies multiplied by two.

$$\text{Power in } [f_1, f_2] = 2 \int_{f_1}^{f_2} P_s(f)\, df \tag{6.22}$$

[8] This content is available online at <http://cnx.org/content/m0515/2.17/>.
[9] "Parseval's Theorem," (1) <http://cnx.org/content/m0047/latest/#parseval>

$$\text{SNR} = \frac{2\alpha^2 \int_0^{\infty} P_x(f)\,df}{N_0(f_u - f_l)} \quad (6.26)$$

In most cases, the interference and noise powers do not vary for a given receiver. Variations in signal-to-interference and signal-to-noise ratios arise from the attenuation because of transmitter-to-receiver distance variations.

6.10 Baseband Communication[11]

We use analog communication techniques for analog message signals, like music, speech, and television. Transmission and reception of analog signals using analog results in an inherently noisy received signal (assuming the channel adds noise, which it almost certainly does).

The simplest form of analog communication is **baseband communication**.

> Point of Interest: We use analog communication techniques for analog message signals, like music, speech, and television. Transmission and reception of analog signals using analog results in an inherently noisy received signal (assuming the channel adds noise, which it almost certainly does).

Here, the transmitted signal equals the message times a transmitter gain.

$$x(t) = Gm(t) \quad (6.27)$$

An example, which is somewhat out of date, is the wireline telephone system. You don't use baseband communication in wireless systems simply because low-frequency signals do not radiate well. The receiver in a baseband system can't do much more than filter the received signal to remove out-of-band noise (interference is small in wireline channels). Assuming the signal occupies a bandwidth of W Hz (the signal's spectrum extends from zero to W), the receiver applies a lowpass filter having the same bandwidth, as shown in Figure 6.5.

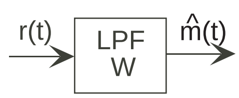


Figure 6.5: The receiver for baseband communication systems is quite simple: a lowpass filter having the same bandwidth as the signal.

We use the **signal-to-noise ratio** of the receiver's output $\widehat{m}(t)$ to evaluate any analog-message communication system. Assume that the channel introduces an attenuation α and white noise of spectral height $\frac{N_0}{2}$. The filter does not affect the signal component—we assume its gain is unity—but does filter the noise, removing frequency components above W Hz. In the filter's output, the received signal power equals $\alpha^2 G^2 \text{power}(m)$ and the noise power $N_0 W$, which gives a signal-to-noise ratio of

$$\text{SNR}_{\text{baseband}} = \frac{\alpha^2 G^2 \text{power}(m)}{N_0 W} \quad (6.28)$$

The signal term power (m) will be proportional to the bandwidth W; thus, in baseband communication the signal-to-noise ratio varies only with transmitter gain and channel attenuation and noise level.

[11] This content is available online at <http://cnx.org/content/m0517/2.19/>.

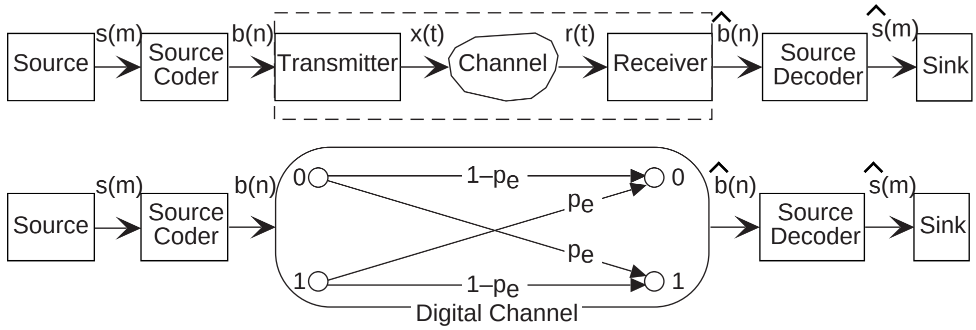


Figure 6.17: The steps in transmitting digital information are shown in the upper system, the Fundamental Model of Communication. The symbolic-valued signal $s(m)$ forms the message, and it is encoded into a bit sequence $b(n)$. The indices differ because more than one bit/symbol is usually required to represent the message by a bitstream. Each bit is represented by an analog signal, transmitted through the (unfriendly) channel, and received by a matched-filter receiver. From the received bitstream $\hat{b}(n)$ the received symbolic-valued signal $\hat{s}(m)$ is derived. The lower block diagram shows an equivalent system wherein the analog portions are combined and modeled by a transition diagram, which shows how each transmitted bit could be received. For example, transmitting a `0` results in the reception of a `1` with probability p_e (an error) or a `0` with probability $1 - p_e$ (no error).

(Section 6.13), Signal Sets[25], BPSK Signal Set[26], Transmission Bandwidth[27], Frequency Shift Keying (Section 6.15), Digital Communication Receivers (Section 6.16), Factors in Receiver Error (Section 6.17), Digital Communication System Properties[28], and Error Probability[29], can be lumped into a single system known as the digital channel.

Digital channels are described by **transition diagrams**, which indicate the output alphabet symbols that result for each possible transmitted symbol and the probabilities of the various reception possibilities. The probabilities on transitions coming from the same symbol must sum to one. For the matched-filter receiver and the signal sets we have seen, the depicted transition diagram, known as a **binary symmetric channel**, captures how transmitted bits are received. The probability of error p_e is the sole parameter of the digital channel, and it encapsulates signal set choice, channel properties, and the matched-filter receiver. With this simple but entirely accurate model, we can concentrate on how bits are received.

6.20 Entropy[30]

Communication theory has been formulated best for symbolic-valued signals. Claude Shannon published in 1948 *The Mathematical Theory of Communication*, which became the cornerstone of digital communication. He showed the power of **probabilistic models** for symbolic-valued signals, which allowed him to quantify the information present in a signal. In the simplest signal model, each symbol can occur at index n with a probability $\Pr[a_k]$, $k = \{1, \ldots, K\}$. What this model says is that for each signal value a K-sided coin is flipped (note that the coin need not be fair). For this model to make sense, the probabilities must be numbers between zero and one and must sum to one.

$$0 \leq \Pr[a_k] \leq 1 \tag{6.48}$$

$$\sum_{k=1}^{K} \Pr[a_k] = 1 \tag{6.49}$$

[25]"Signal Sets" <http://cnx.org/content/m0542/latest/>
[26]"BPSK signal set" <http://cnx.org/content/m0543/latest/>
[27]"Transmission Bandwidth" <http://cnx.org/content/m0544/latest/>
[28]"Digital Communication System Properties" <http://cnx.org/content/m0547/latest/>
[29]"Error Probability" <http://cnx.org/content/m0548/latest/>
[30]This content is available online at <http://cnx.org/content/m0070/2.13/>.

module on the Source Coding Theorem (Section 6.21) we find that using a so-called **fixed rate** source coder, one that produces a fixed number of bits/symbol, may not be the most efficient way of encoding symbols into bits. What is not discussed there is a procedure for designing an efficient source coder: one **guaranteed** to produce the fewest bits/symbol on the average. That source coder is not unique, and one approach that does achieve that limit is the **Huffman source coding algorithm**.

POINT OF INTEREST: In the early years of information theory, the race was on to be the first to find a **provably** maximally efficient source coding algorithm. The race was won by then MIT graduate student David Huffman in 1954, who worked on the problem as a project in his information theory course. We're pretty sure he received an "A."

- Create a vertical table for the symbols, the best ordering being in decreasing order of probability.
- Form a binary tree to the right of the table. A binary tree always has two branches at each node. Build the tree by merging the two lowest probability symbols at each level, making the probability of the node equal to the sum of the merged nodes' probabilities. If more than two nodes/symbols share the lowest probability at a given level, pick any two; your choice won't affect $\overline{B}(A)$.
- At each node, label each of the emanating branches with a binary number. The bit sequence obtained from passing from the tree's root to the symbol is its Huffman code.

Example 6.3
The simple four-symbol alphabet used in the Entropy (Example 6.1) and Source Coding (Example 6.2) modules has a four-symbol alphabet with the following probabilities,

$$\Pr[a_0] = \frac{1}{2} \quad \Pr[a_1] = \frac{1}{4} \quad \Pr[a_2] = \frac{1}{8} \quad \Pr[a_3] = \frac{1}{8}$$

and an entropy of 1.75 bits (Example 6.1). This alphabet has the Huffman coding tree shown in Figure 6.18.

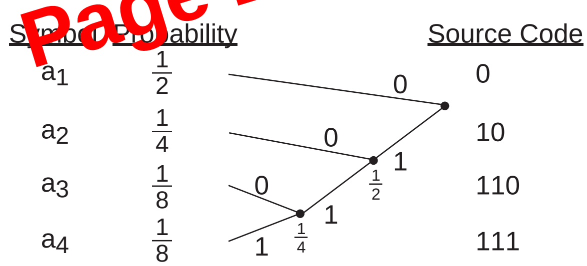


Figure 6.18: We form a Huffman code for a four-letter alphabet having the indicated probabilities of occurrence. The binary tree created by the algorithm extends to the right, with the root node (the one at which the tree begins) defining the codewords. The bit sequence obtained by traversing the tree from the root to the symbol defines that symbol's binary code.

The code thus obtained is not unique as we could have labeled the branches coming out of each node differently. The average number of bits required to represent this alphabet equals 1.75 bits, which is the Shannon entropy limit for this source alphabet. If we had the symbolic-valued signal $s(m) = \{a_2, a_3, a_1, a_4, a_1, a_2, \dots\}$, our Huffman code would produce the bitstream $b(n) = 101100111010\dots$.

If the alphabet probabilities were different, clearly a different tree, and therefore different code, could well result. Furthermore, we may not be able to achieve the entropy limit. If our symbols had the probabilities $\Pr[a_1] = \frac{1}{2}$, $\Pr[a_2] = \frac{1}{4}$, $\Pr[a_3] = \frac{1}{5}$, and $\Pr[a_4] = \frac{1}{20}$, the average number of bits/symbol resulting from the Huffman coding algorithm would equal 1.75 bits. However, the

	%	**Morse Code**	**Huffman Code**
A	6.22	.-	1011
B	1.32	-...	010100
C	3.11	-.-.	10101
D	2.97	-..	01011
E	10.53	.	001
F	1.68	..-.	110001
G	1.65	–.	110000
H	3.63	....	11001
I	6.14	..	1001
J	0.06	.—	01010111011
K	0.31	-.-	01010110
L	3.07	.-..	10100
M	2.48	–	00011
N	5.73	-.	0100
O	6.06	—	1000
P	1.87	.–.	00000
Q	0.10	–.-	0101011100
R	5.87	.-.	0111
S	5.81	...	0110
T	7.68	-	1101
U	2.27	..-	00010
V	0.70	...-	0101010
W	1.13	.–	000011
X	0.25	-..-	010101111
Y	1.07	-.—	000010
Z	0.06	–..	0101011101011

Figure 6.19: Morse and Huffman Codes for American-Roman Alphabet. The % column indicates the average probability (expressed in percent) of the letter occurring in English. The entropy $H(A)$ of the this source is 4.14 bits. The average Morse codeword length is 2.5 symbols. Adding one more symbol for the letter separator and converting to bits yields an average codeword length of 5.56 bits. The average Huffman codeword length is 4.35 bits.

---

would tap the message using a telegraph key to another operator, who would relay the message on to the next operator, presumably getting the message closer to its destination. In short, the telegraph relied on a **network** not unlike the basics of modern computer networks. To say it presaged modern communications would be an understatement. It was also far ahead of some needed technologies, namely the Source Coding Theorem. The Morse code, shown in Figure 6.19, was not a prefix code. To separate codes for each letter, Morse code required that a space—a pause—be inserted between each letter. In information theory, that space counts as another code letter, which means that the Morse code encoded text with a three-letter source code: dots, dashes and space. The resulting source code is not within a bit of entropy, and is grossly inefficient (about 25%). Figure 6.19 shows a Huffman code for English text, which as we know **is** efficient.

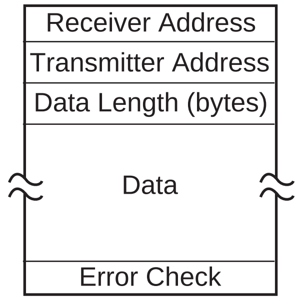


Figure 6.26: Long messages, such as files, are broken into separate packets, then transmitted over computer networks. A packet, like a letter, contains the destination address, the return address (transmitter address), and the data. The data includes the message part and a sequence number identifying its order in the transmitted message.

network model by subdividing messages into smaller chunks called **packets**(Figure 6.26). The rationale for the network enforcing smaller transmissions was that large file transfers would consume network resources all along the route, and, because of the long transmission time, a communication failure might require retransmission of the entire file. By creating packets, each of which has its own address and is routed independently of others, the network can better manage congestion. The analogy is that the postal service, rather than sending a long letter in the envelope you provide, opens the envelope, places each page in a separate envelope, and using the address on your envelope, addresses each page's envelope accordingly, and mails them separately. The network does need to make sure packet sequence (page numbering) is maintained, and the network exit point must reassemble the original message accordingly.

Communications networks are now categorized according to whether they use packets or not. A system like the telephone network is said to be **circuit switched**: The network establishes a **fixed** route that lasts the entire duration of the message. Circuit switching has the advantage that once the route is determined, the users can use the capacity provided them however they like. Its main disadvantage is that the users may not use their capacity efficiently, clogging network links and nodes along the way. **Packet-switched** networks continuously monitor network utilization, and route messages accordingly. Thus, messages can, on the average, be delivered efficiently, but the network cannot guarantee a specific amount of capacity to the users.

6.35 Network architectures and interconnection[46]

The network structure—its architecture (Figure 6.25)—typifies what are known as **wide area networks**(WANs). The nodes, and users for that matter, are spread geographically over long distances. "Long" has no precise definition, and is intended to suggest that the communication links vary widely. The Internet is certainly the largest WAN, spanning the entire earth and beyond. **Local area networks**, LANs, employ a single communication link and special routing. Perhaps the best known LAN is Ethernet[47]. LANs connect to other LANs and to wide area networks through special nodes known as **gateways**(Figure 6.27). In the Internet, a computer's address consists of a four byte sequence, which is known as its **IP address** (Internet Protocol address). An example address is **128.42.4.32**: each byte is separated by a period. The first two bytes specify the computer's **domain**(here Rice University). Computers are also addressed by a more human-readable form: a sequence of alphabetic abbreviations representing institution, type of institution, and computer name. A given computer has both names (**128.42.4.32** is the same as **soma.rice.edu**). Data transmission on the Internet requires the numerical form. So-called **name servers** translate between alphabetic and numerical forms, and the transmitting computer requests this translation before the message is sent to the network.

[46] This content is available online at <http://cnx.org/content/m0077/2.10/>.
[47] "Ethernet" <http://cnx.org/content/m0078/latest/>

each with individual specifications, the most distinguishing of which is capacity: 10 Mbps and 100 Mbps. If the minimum transmission time is such that the beginning of the packet has not propagated the full length of the Ethernet before the end-of-transmission, it is possible that two computers will begin transmission at the same time and, by the time their transmissions cease, the other's packet will not have propagated to the other. In this case, computers in-between the two will sense a collision, which renders both computer's transmissions senseless to them, without the two transmitting computers knowing a collision has occurred at all! For Ethernet to succeed, we must have the minimum packet transmission time exceed **twice** the voltage propagation time: $\frac{P_{\min}}{C} > \frac{2L}{c}$ or

$$P_{\min} > \frac{2LC}{c} \tag{6.63}$$

Thus, for the 10 Mbps Ethernet having a 1 km maximum length specification, the minimum packet size is 200 bits.

Exercise 6.38 *(Solution on p. 259.)*

The 100 Mbps Ethernet was designed more recently than the 10 Mbps alternative. To maintain the same minimum packet size as the earlier, slower version, what should its length specification be? Why should the minimum packet size remain the same?

6.37 Communication Protocols[49]

The complexity of information transmission in a computer network—reliable transmission of bits across a channel, routing, and directing information to the correct destination within the destination computers operating system—demands an overarching concept of how to organize information delivery. No unique set of rules satisfies the various constraints communication channels and network organization place on information transmission. For example, random access issues in Ethernet are not present in wide-area networks such as the Internet. A **protocol** is a set of rules that governs how information is delivered. For example, to use the telephone network, the protocol is to pick up the phone, listen for a dial tone, dial a number having a specific number of digits, wait for the phone to ring, and say hello. In radio, the station uses amplitude or frequency modulation with a specific carrier frequency and transmission bandwidth, and you know to turn on the radio and tune in the station. In technical terms, no one protocol or set of protocols can be used for any communication situation. Be that as it may, communication engineers have found that a common thread runs through the **organization** of the various protocols. This grand design of information transmission organization runs through all modern networks today.

What has been defined as a networking standard is a layered, hierarchical protocol organization. As shown in Figure 6.30 (Protocol Picture), protocols are organized by function and level of detail. Segregation of information transmission, manipulation, and interpretation into these categories directly affects how communication systems are organized, and what role(s) software systems fulfill. Although not thought about in this way in earlier times, this organizational structure governs the way communication engineers think about all communication systems, from radio to the Internet.

Exercise 6.39 *(Solution on p. 259.)*

How do the various aspects of establishing and maintaining a telephone conversation fit into this layered protocol organization?

We now explicitly state whether we are working in the physical layer (signal set design, for example), the data link layer (source and channel coding), or any other layer. IP abbreviates Internet protocol, and governs gateways (how information is transmitted between networks having different internal organizations). TCP (transmission control protocol) governs how packets are transmitted through a wide-area network such as the Internet. Telnet is a protocol that concerns how a person at one computer logs on to another computer across a network. A moderately high level protocol such as telnet, is not concerned with what data links (wireline or wireless) might have been used by the network or how packets are routed. Rather, it establishes connections between computers and directs each byte (presumed to represent a typed character) to the

[49]This content is available online at <http://cnx.org/content/m0080/2.19/>.

Solution to Exercise 6.21 (p. 219)
Equally likely symbols each have a probability of $\frac{1}{K}$. Thus, $H(A) = -\left(\sum_k \left(\frac{1}{K}\log_2\left(\frac{1}{K}\right)\right)\right) = \log_2 K$. To prove that this is the maximum-entropy probability assignment, we must explicitly take into account that probabilities sum to one. Focus on a particular symbol, say the first. $\Pr[a_0]$ appears **twice** in the entropy formula: the terms $\Pr[a_0]\log_2(\Pr[a_0])$ and $(1-\Pr[a_0]+\cdots+\Pr[a_{K-2}])\log_2(1-\Pr[a_0]+\cdots+\Pr[a_{K-2}])$. The derivative with respect to this probability (and all the others) must be zero. The derivative equals $\log_2(\Pr[a_0]) - \log_2(1-\Pr[a_0]+\cdots+\Pr[a_{K-2}])$, and all other derivatives have the same form (just substitute your letter's index). Thus, each probability must equal the others, and we are done. For the minimum entropy answer, one term is $1\log_2 1 = 0$, and the others are $0\log_2 0$, which we define to be zero also. The minimum value of entropy is zero.

Solution to Exercise 6.22 (p. 222)
The Huffman coding tree for the second set of probabilities is **identical** to that for the first (Figure 6.18 (Huffman Coding Tree)). The average code length is $\frac{1}{2}1+\frac{1}{4}2+\frac{1}{5}3+\frac{1}{20}3 = 1.75$ bits. The entropy calculation is straightforward: $H(A) = -\left(\frac{1}{2}\log\left(\frac{1}{2}\right)+\frac{1}{4}\log\left(\frac{1}{4}\right)+\frac{1}{5}\log\left(\frac{1}{5}\right)+\frac{1}{20}\log\left(\frac{1}{20}\right)\right)$, which equals 1.68 bits.

Solution to Exercise 6.23 (p. 222)
$T = \frac{1}{\overline{B}(A)R}$.

Solution to Exercise 6.24 (p. 222)
Because no codeword begins with another's codeword, the first codeword encountered in a bit stream must be the right one. Note that we must start at the beginning of the bit stream; jumping into the middle does not guarantee perfect decoding. The end of one codeword and the beginning of another could be a codeword, and we would get lost.

Solution to Exercise 6.25 (p. 222)
Consider the bitstream …0110111… taken from the bitstream 0|10|110|110|111|…. We would decode the initial part incorrectly, then would synchronize. If we had a fixed-length code (say 00,01,10,11), the situation is **much** worse. Jumping into the middle leads to no synchronization at all!

Solution to Exercise 6.26 (p. 224)
This question is equivalent to $3p_e(1-p_e)+p_e^2 \le 1$ or $2p_e^2+(-3)p_e+1 \ge 0$. Because this is an upward-going parabola, we need only check where its roots are. Using the quadratic formula, we find that they are located at $\frac{1}{2}$ and 1. Consequently in the range $0 \le p_e \le \frac{1}{2}$ the error rate produced by coding is smaller.

Solution to Exercise 6.27 (p. 226)
With no coding, the average bit-error probability p_e is given by the probability of error equation (6.47): $p_e = Q\left(\sqrt{\frac{2\alpha^2 E_b}{N_0}}\right)$. With a threefold repetition code, the bit-error probability is given by $3{p_e'}^2(1-p_e')+{p_e'}^3$, where $p_e' = Q\left(\sqrt{\frac{2\alpha^2 E_b}{3N_0}}\right)$. Plotting this reveals that the increase in bit-error probability out of the channel because of the energy reduction is not compensated by the repetition coding.

Chapter 7

Appendix

7.1 Decibels[1]

The decibel scale expresses amplitudes and power values **logarithmically**. The definitions for these differ, but are consistent with each other.

$$\begin{aligned} \text{power}(s, \text{ in decibels}) &= 10\log_{10}\left(\frac{\text{power}(s)}{\text{power}(s_0)}\right) \\ \text{amplitude}(s, \text{ in decibels}) &= 20\log_{10}\left(\frac{\text{amplitude}(s)}{\text{amplitude}(s_0)}\right) \end{aligned} \quad (7.1)$$

Here $\text{power}(s_0)$ and $\text{amplitude}(s_0)$ represent a **reference** power and amplitude, respectively. Quantifying power or amplitude in decibels essentially means that we are comparing quantities to a standard or that we want to express how they changed. You will hear statements like "The signal went down by 3 dB" and "The filter's gain in the stopband is −60" (Decibels is abbreviated dB.).

Exercise 7.1 ***(Solution on p. 265.)***

The prefix "deci" implies a tenth; a decibel is a tenth of a Bel. Who is this measure named for?

The consistency of these two definitions arises because power is proportional to the square of amplitude:

$$\left(\text{power}(s) \propto \text{amplitude}^2(s)\right) \quad (7.2)$$

Plugging this expression into the definition for decibels, we find that

$$\begin{aligned} 10\log_{10}\left(\frac{\text{power}(s)}{\text{power}(s_0)}\right) &= 10\log_{10}\left(\frac{\text{amplitude}^2(s)}{\text{amplitude}^2(s_0)}\right) \\ &= 20\log_{10}\left(\frac{\text{amplitude}(s)}{\text{amplitude}(s_0)}\right) \end{aligned} \quad (7.3)$$

Because of this consistency, **stating relative change in terms of decibels is unambiguous**. A factor of 10 increase in amplitude corresponds to a 20 dB increase in both amplitude and power!

The accompanying table provides "nice" decibel values. Converting decibel values back and forth is fun, and tests your ability to think of decibel values as sums and/or differences of the well-known values and of ratios as products and/or quotients. This conversion rests on the logarithmic nature of the decibel scale. For example, to find the decibel value for $\sqrt{2}$, we halve the decibel value for 2; 26 dB equals $10 + 10 + 6$ dB that corresponds to a ratio of $10 \times 10 \times 4 = 400$. Decibel quantities add; ratio values multiply.

One reason decibels are used so much is the frequency-domain input-output relation for linear systems: $Y(f) = X(f)H(f)$. Because the transfer function multiplies the input signal's spectrum, to find the output

[1] This content is available online at <http://cnx.org/content/m0082/2.16/>.

Figure 7.2

Module: "Digital Communication System Properties"
By: Don Johnson
URL: http://cnx.org/content/m10282/2.9/
Page: 217
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Digital Channels"
By: Don Johnson
URL: http://cnx.org/content/m0102/2.14/
Pages: 217-218
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Entropy"
By: Don Johnson
URL: http://cnx.org/content/m0070/2.13/
Pages: 218-219
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Source Coding Theorem"
By: Don Johnson
URL: http://cnx.org/content/m0091/2.13/
Pages: 219-220
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Compression and the Huffman Code"
By: Don Johnson
URL: http://cnx.org/content/m0092/2.19/
Pages: 220-222
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Subtleties of Source Coding"
Used here as: "Subtleties of Coding"
By: Don Johnson
URL: http://cnx.org/content/m0093/2.16/
Pages: 222-223
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0

Module: "Channel Coding"
By: Don Johnson
URL: http://cnx.org/content/m10782/2.5/
Page: 224
Copyright: Don Johnson
License: http://creativecommons.org/licenses/by/1.0