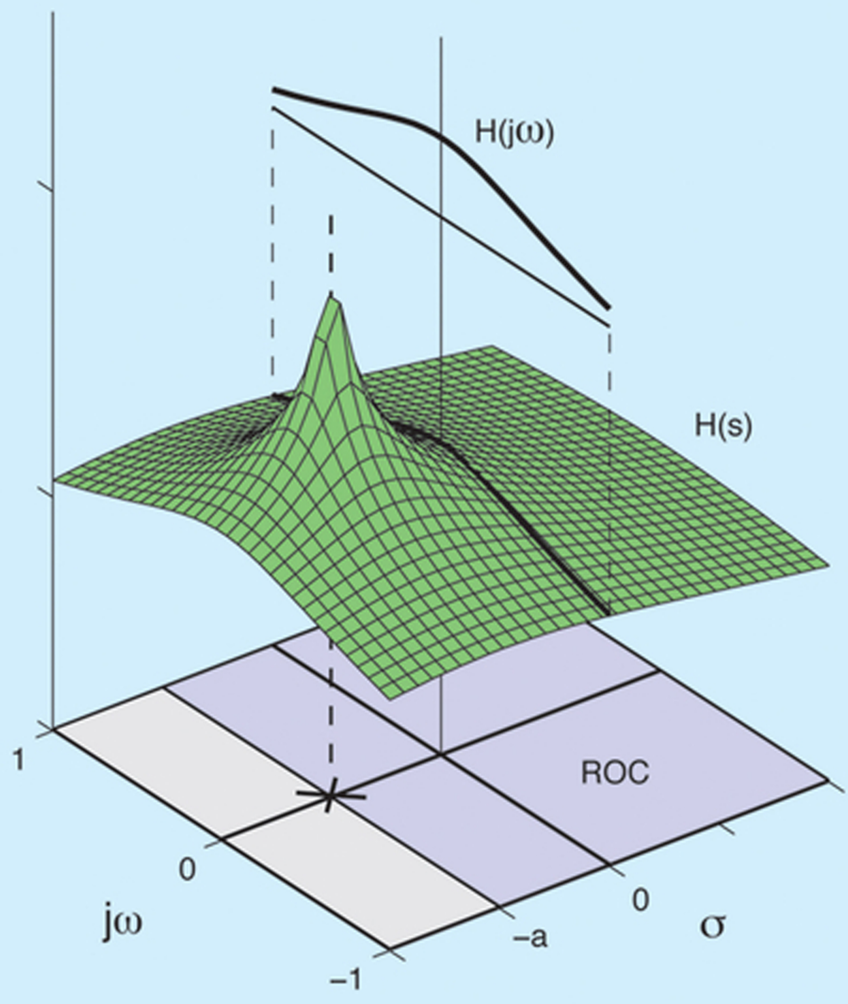
# Practical Signals Theory with MATLAB® Applications

RICHARD J. TERVO



# Practical Signals Theory with MATLAB® Applications

# Practical Signals Theory with MATLAB® Applications

RICHARD J. TERVO

WILEY

**VP & PUBLISHER:** Donald Fowley
**ASSOCIATE PUBLISHER:** Dan Sayre
**EDITORIAL ASSISTANT:** Jessica Knecht
**MARKETING MANAGER:** Chris Ruel
**MARKETING ASSISTANT:** Marissa Carroll
**DESIGNER:** Alson Low
**ASSOCIATE PRODUCTION MANAGER:** Joyce Poh
**PRODUCTION EDITOR:** Jolene Ling
**COVER ILLUSTRATION:** Richard J. Tervo

This book was set in 10/12 NewBaskervilleStd-Roman by MPS Limited, Chennai and printed and bound by Courier Kendallville. The cover was printed by Courier Kendallville.

This book is printed on acid-free paper.



**Library of Congress Cataloging-in-Publication Data**

Tervo, Richard J., 1958–
Practical signals theory with MATLAB® applications / Richard J. Tervo.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-118-11539-8 (pbk.)
1. Signal theory (Telecommunication)—Mathematics. 2. MATLAB. I. Title.
TK5102.83.T47 2014
621.382'23—dc23

2012030207

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

*To Rowena*

# BRIEF CONTENTS

# CONTENTS

The symbol ⊚ indicates advanced content that may be omitted without loss of continuity.

# PREFACE

The title *Practical Signals Theory* incorporates both the words *practical* and *theory* to underscore the reality that engineers use mathematics as a tool for practical ends, often to gain a better understanding of the behavior of the world around them and just as often simply to save time and work. True to this notion, signals theory offers both a means to model complex real-world systems using consistent mathematical methods and a way to avoid tedious manipulations by leveraging the efforts of mathematicians and engineers who have already done it the hard way. Thus, signals theory includes the famous transforms named after *Fourier* and *Laplace* designed to view real systems from advantageous new perspectives. Frequency and phase responses are easily sketched with pencil and ruler following in the footsteps of *Bode*, and modern digital signal processing owes a debt to *Nyquist*. Moreover, in every equation or formula there is a clue that relates to something that is real and that may already be very familiar.

*Practical Signals Theory* was written specifically to present the essential mathematics of signals and systems through an intuitive and graphical approach in which theory and principles emerge naturally from the observed behavior of familiar systems. To this end, every new theorem is introduced by real-world examples, and every new example is chosen to reflect an important element of signals theory. From the first pages, even the most basic mathematical relationships are reexamined in a way that will lend their use to the practical application of signals theory. Many examples are presented using MATLAB, which has become a standard for electrical engineering students around the world.

## Pedagogy

Any presentation of signals theory to an undergraduate audience must confront the inevitable compromise between keeping the subject material accessible to the modern student and maintaining the level of mathematical rigor that is the cornerstone of engineering studies. The philosophical issues surrounding rigor are hardly new[1], although it is perhaps ironic, in this course especially, that many of the distractions now available to students have come about from commercial applications of signals theory.[1] The presentation of material in this text proceeds through a carefully paced progression of concepts using sketches and practical examples to motivate appreciation of the essential elements of signals theory. To that end, the ability to visualize signals and their transforms is developed as an important skill that complements a full appreciation of the underlying mathematics. Indeed, knowing *why* the math works and *how* signals interact through established principles is what distinguishes true understanding from the mere ability to memorize and to manipulate formulas and equations. On the other hand, there are many situations in which a signal is seen on an instrument or in some graphical or numerical output and the important question of *does it make sense?* can only be

[1] Video games, MP3 players, and mobile telephones come to mind.

answered if the expected behavior can be readily reasoned and visualized. To supplement this approach, the use of MATLAB is promoted as a versatile tool to define, to manipulate, to display, and ultimately to better understand the theory of signals and systems. The strengths of this text include:

- The essential mathematics of orthogonal signals, the Fourier series, the Fourier transform (both continuous and discrete), the Laplace transform, and the $z$-transform are thoroughly covered and compared. Examples and applications of different techniques, while focusing on electrical engineering, are drawn from a range of subject areas.
- Introductory sections review fundamental mathematics and important time-domain functions while laying the groundwork for the transform-domain manipulations to follow. Requisite mathematical skills such as the manipulation of complex values are reexamined when first encountered. During a derivation or example calculation, care is taken to include complete intermediate steps to maintain focus and clarity of presentation.
- New concepts in signals and systems are presented intuitively and graphically with real-world examples and by working through the relevant mathematics. Transform domains are presented as different perspectives on the same signal where properties and signal behavior are linked through the underlying mathematics.
- Fourier analysis is defined around a $(t, f)$ notation, which lends itself well to discussions of frequency and avoids awkward rescaling terms. The conventional use of $(t, \omega)$ is employed for Laplace and $z$-transform expositions.
- The appendices include useful reference tables and feature *illustrated transforms* that graphically present transform pairs side by side and that highlight important properties relating the time and transform domains.
- The use of hand-drawn sketches is encouraged when exploring the properties of signals and their manipulation. New and interesting signal properties are revealed, and the underlying theory is developed by testing different ideas on known equations.
- Prior to every calculation, the reader is encouraged to sketch signals by hand and to predict expected results. After a derivation or calculation, the reader is encouraged to check the answer for consistency and to compare to predicted outcomes. This practice serves to build confidence in the theory and in the results obtained.
- The use of MATLAB is presented as a quick and convenient way to rapidly view and manipulate signals and systems. Examples are limited to functions that are available in the MATLAB Student Version.
- Many figures in the text are generated using MATLAB, and the script used to create a figure is often included on the same page.
- MATLAB functions of special interest to signals and systems are used to confirm and to illustrate example exercises. In this way, the behavior of systems modelled as a continuous or discrete transfer function, pole-zero diagram, or state space equation can readily be studied as a Bode diagram, pole-zero diagram, impulse response, or step response.

# Organization

This introductory text covers signals and linear systems theory, including continuous time and discrete time signals, the Fourier transform, the Laplace transform, and the *z*-transform. The sequence follows through continuous time signals and systems, orthogonality, the Fourier series, the Fourier transform, the Laplace transform, discrete time signals including sampling theorem, the DTFT and DFT, and the *z*-transform. A bonus chapter on communications signals is provided as an additional source of practical applications of signals theory and will be of special interest to those students who may not otherwise take a communications systems course as part of their core curriculum.

Each chapter integrates numerous MATLAB examples and illustrations. Particular use is made of the MATLAB system definitions based on transfer function, zero-pole-gain model, or state space model to study the behavior of linear systems using the impulse response, step response, Bode plot, and pole-zero diagram. The ability to model and to examine simple systems with these tools is an important skill that complements and reinforces understanding of the mathematical concepts and manipulations.

## Chapter 1. Introduction to Signals and Systems

Signals and systems and their interaction are developed, beginning with simple and familiar signals and manipulations. Mathematical and graphical concepts are reviewed with emphasis on the skills that will prove most useful to the study of signals and systems. Shifting and scaling and linear combinations of time-domain signals are sketched by hand. The frequency and phase characteristics of a sinusoid are carefully examined, and the elements of a general sinusoid are identified. The impulse function, unit step, and unit rectangle signals are defined, and common elements of system block diagrams are introduced.

## Chapter 2. Classification of Signals

Signals are identified as real or complex, odd or even, periodic or nonperiodic, energy or power, continuous or discrete. Examples of common signals of all types and their definitions in MATLAB are introduced.

## Chapter 3. Linear Systems

The linear time invariant system is defined. Convolution is examined in detail. System impulse response is introduced as well as causality.

## Chapter 4. The Fourier Series

A signal is represented in terms of orthogonal components. The special set of orthogonal sinusoids is introduced, first as the Fourier series and then as the complex Fourier series.

## Chapter 5. The Fourier Transform

The Fourier transform is developed as a limiting case of the Fourier series. The definition of the Fourier transform and its properties follow, with emphasis on relating the time- and frequency-domain characteristics mathematically and graphically.

## Chapter 6. Practical Fourier Transforms

The introduction of the convolution theorem opens up the full potential of the Fourier transform in practical applications. The concept of transfer function is introduced, and a discussion of data filtering follows and leads to the inevitable issue of causality in ideal filters. Discussion of the energy in signals introduces Parseval's theorem. The Fourier transform of an impulse train reveals properties that apply to all periodic signals.

## Chapter 7. The Laplace Transform

The Laplace transform is developed as an extension of the Fourier transform to include many signals that do not have a Fourier transform and those which incorporate non-sinusoidal component. First- and second-order systems having differential and integral components are studied. State space models are introduced.

## Chapter 8. Discrete Signals

Sampling theory and the importance of sampling rate, aliasing, and practical sampling issues are covered as well as the nature of discrete signals under the Fourier transform, including an introduction to digital signal processing.

## Chapter 9. The *z*-Transform

The *z*-transform is developed as discrete time version of the Laplace transform and as an extension of the discrete time Fourier transform. FIR and IIR filters are used as applications of the *z*-transform.

## Chapter 10. Introduction to Communications

The frequency spectrum and the concept of carrier frequency are presented along with basic analog modulation types. Radio frequency interference is discussed. The concept of a superheterodyne receiver is studied as an application of the modulation theorem. Amplitude modulation and demodulation are examined as well as an introduction to digital modulation techniques.

In a single-semester course focusing on Fourier theory, the material may be presented in the order 1-2-3-4-5-6-8, thereby omitting the Laplace transform and *z*-transform. For a focus on continuous time signals and systems, the order may be 1-2-3-4-5-6-7, thereby omitting discrete signals and the *z*-transform.

This text was originally written for the students in EE3513 at the University of New Brunswick, where the material was first presented over twenty-five years ago. Questions and feedback in the classroom have contributed immensely to the style and presentation of the material and in the examples given throughout the text. I welcome comments and criticism, and I wish everyone success in the coming semester.

**Richard J. Tervo**

# ACKNOWLEDGMENTS

I am grateful for the support and encouragement received from Wiley during the publication of this new book and I especially recognize those people who have worked behind the scenes to facilitate the entire process. After many years of tentative preparation, this adventure finally took flight when Tiina Ruonamaa recognized a new author and convinced others to buy into the project. Dan Sayre has expertly guided the entire process and during the development stage deftly nudged the content in directions that I would not otherwise have anticipated. I shall miss the midnight emails from Jolene Ling who has led her production team to produce this wonderful printed volume in its final form.

Many years ago I learned a great deal about practical signals analysis from Dr. T. J. Kennett who remains a role model in my academic career. This manuscript was first adopted in the classroom by someone other than myself when Dr. P. A. Parker chose this material for his course and in doing so prompted me to revise and to rewrite my class notes in LaTex form and using MATLAB. Those who remember Toby Tervo will recognize the ways in which he provided support during many long hours working on the final manuscript. Finally, to my colleagues and friends who have endured endless tales about Richard "working on a book", I am happy to announce that it is finally here.

CHAPTER 1

# Introduction to Signals and Systems

In this chapter, basic definitions essential to the study of signals and systems are introduced. While some of these concepts will undoubtedly seem familiar, they are presented here in the context of electrical engineering applications and in terms of specific properties that will prove important in signals analysis. The mathematical operations of shifting, scaling, and multiplication are related to signal delays, time compression, and amplification. Emphasis is placed on employing linear combinations of simple signal types to aid the study of new and interesting functions. A variety of graphical and arithmetic examples are presented. Features of MATLAB used to manipulate and to plot various signals are highlighted.

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Write the definitions of a signal and a system
- Define continuous-time and discrete-time signals
- Explain the mathematical basis of shifting, scaling, and multiplication
- Describe the shifting, scaling, and multiplication operations graphically
- Identify a linear combination of signals
- Describe basic signals in standard form: unit rectangle, unit step, impulse, comb, sinusoid
- Create and sketch new signals as linear combinations of the basic signals
- Identify variations of the basic signals by component
- Apply mathematical operations to describe new signals in terms of the basic signals
- Use MATLAB to define and to plot simple time domain signals

## 1.1 Introduction

The world of electrical engineering includes applications as diverse as digital computers, mobile telephones, communication satellites, and automobile sound systems. All these devices perform some useful function by interacting with the electrical signals that pass through them. The engineers who develop these systems strive to achieve the best performance at the best cost by applying the laws of physics and mathematics to build these everyday products. It follows that engineers require a sound understanding of the fundamental physical behavior of both the signals and the hardware and of the available tools and techniques for modelling and simulation. In the most generalized approach, various types of *signals* might pass through different sorts of *systems*, and this defines the broad and fascinating subject matter of *signals and systems*. This textbook presents the study of signals and systems through illustrations and examples of how the mathematics of signals theory relates to practical and everyday applications, including how signals may be described, how they interact, and how the results of such operations may be effectively applied to the understanding and solution of engineering problems.

### 1.1.1 What Is a Signal?

An electrical signal may be defined as a voltage or current that may vary with time; such a signal can be represented and its behavior studied as a mathematical function of time called, for example, $s(t)$. Any study of the time variations of signals is said to take place in the *time domain*. Time domain signals are those familiar waveforms observed using an oscilloscope, or those that immediately come to mind upon hearing the words *sine wave* or *square wave*. It will be seen that the examination of signals from other perspectives (in particular, the frequency domain) can be of great advantage; indeed, the mastery of manipulating signals from various perspectives represents a major goal of this textbook. It is through the application of mathematical transformations that such different perspectives can be achieved. On the other hand, many of the principles and much of the groundwork for signal analysis can be laid out in the familiar surroundings of the time domain.

### 1.1.2 What Is a System?

Observing the monotonous undulations of the cosine waveform $s(t) = \cos(t)$ or the mathematics of some other function $s(t)$ would not hold anyone's interest for very long. Rather, the utility of such signals depends entirely on what can be done with them and what happens to the signals as they pass through a circuit or over a radio channel or into a computer. Of course, it would be impossible to study the effect of every circuit on every possible input signal, so some abstraction is necessary. Instead, a generalized signal will be studied as it passes through an arbitrary system. Consequently, the signal $s(t)$ may be either a voltage or current, it might be a sine or cosine or some other waveform, and it will be studied as it passes through what could possibly be a transistor amplifier, but which might well be a computer, or perhaps a shortwave radio channel. Simple systems may be described through sketches that illustrate fundamental signal manipulations, which, in turn, are described by mathematical operations. Effectively, these mathematical functions completely describe a system and, from a signals analysis point of view, it does not matter what kind of components or circuitry the system actually contains.

Figure 1.1 models a simple system in which a signal $a(t)$ passes through a system to emerge as the signal $b(t)$. The short form notation for such a system will be $a(t) \rightarrow b(t)$. This notation implies that the input signal $a(t)$ leads to the output signal

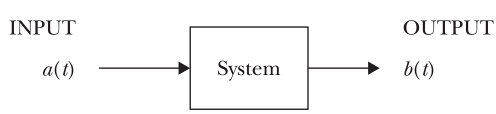



**FIGURE 1.1** System Modelling—The input signal $a(t)$ passes through some system to emerge as the output signal $b(t)$.

$b(t)$ after interacting with the system. Representing voltages and currents, and the interaction of signals in various systems using mathematical models allows a concise and thorough appreciation of signal behavior. Consequently, much of the emphasis of this textbook is in understanding the underlying mathematics and being comfortable with manipulating equations and interpreting their significance in terms of real-world applications. By relating generalized mathematical results to practical systems, the power and convenience of using mathematical descriptions of signals will be emphasized. Through the study of down-to-earth examples, an approach to studying signals will be developed, which can then be applied to any type of signal and to a variety of applications.

In this chapter, a variety of subjects are discussed that will serve to review background material essential to the exploration of signals and systems. But while much of this material may seem like basic mathematical functions and relations, it is presented here from a perspective that lends itself specifically to a study of signals. The study of signals begins with the mathematical abstractions that underlie the manipulation of real signals expressed as a function such as $s(t)$.

As a convention, lowercase letters will be used to define time domain signals. Consequently, $a(t)$, $b(t)$, and $s(t)$ are all time domain signals. This notation will be extended to use the corresponding uppercase letters for related signal representations. Specifically, $A(f)$ will denote the frequency-domain version of the signal $a(t)$. Likewise, in later chapters, the function $A(s)$ will denote the Laplace transform version of $a(t)$ and $A(z)$ will be used to denote the $z$-transform.

## 1.2 Introduction to Signal Manipulation

The study of signals theory begins with a very practical example. Consider a voltage signal within some electronic circuit. As the signal evolves in time, it may be measured and described as the continuous function of time $s(t)$. This observation leads to the definition:

**DEFINITION 1.1 Continuous Time Signal**

*If a signal* s(t) *is a function of time that has a value defined for every* t *on some continuous domain,*[1] *then* s(t) *is a* continuous time signal *or simply a* continuous signal.

The continuous signal $s(t)$ may be of finite or infinite duration and this definition does not exclude a signal $s(t)$ having discontinuities, where $s(t)$ may be *piecewise continuous.* Many elements of signals analysis and manipulation can be represented by relatively simple mathematical operations performed on the function $s(t)$. In a practical sense, these operations describe the simplest of systems, yet they can be combined in a variety of ways to describe the most complex systems.

As an example, the signal $s(t)$ shown in Figure 1.2 is defined for time $t = -1$ to $t = +1$ and is zero otherwise. This time interval[2] may be written as $[-1, 1]$. Upon

[1] The *domain* of a function $s(t)$ is the set of values that $t$ may take on. The *range* of a function $s(t)$ is the set of values that $s(t)$ takes on [5].

[2] An *interval* of real numbers is the set containing all numbers between two given numbers (the endpoints) and one, both, or neither endpoint [5]. A *closed interval* written as $[a, b]$ contains its endpoints $a$ and $b$.

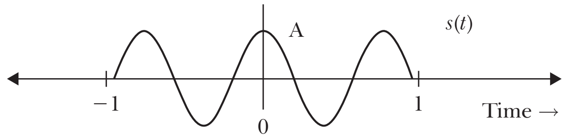



**FIGURE 1.2** The signal $s(t)$ may be manipulated through simple mathematical operations.

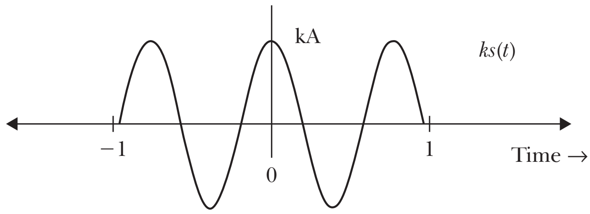



**FIGURE 1.3** The system $s(t) \rightarrow ks(t)$ varies the amplitude of $s(t)$.

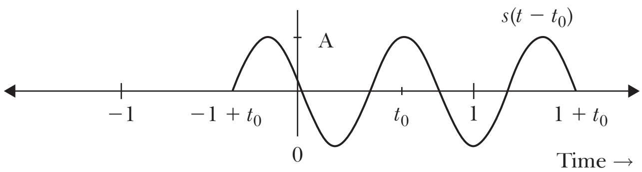



**FIGURE 1.4** The system $s(t) \rightarrow s(t - t_0)$ shifts $s(t)$ in time.

passing through some system, this signal may undergo a transformation, which is reflected both in its appearance and in the function of time describing the output signal. Several such systems are described below, both mathematically and graphically.

**Amplifying** If a signal $s(t)$ is multiplied by a constant $k$, it is increased in amplitude when $k > 1$, and attenuated for $0 < k < 1$, as in Figure 1.3. A simple amplifier or attenuator circuit could be described by the notation $s(t) \rightarrow ks(t)$. Significantly, multiplying by a constant will not affect the overall appearance or shape of the input signal $s(t)$.

**Shifting** A signal $s(t)$ may be shifted along the time axis by a constant nonnegative time $t_0$ using the operation $s(t) \rightarrow s(t - t_0)$, as in Figure 1.4. This shift operation corresponds to a time delay. Note that in the signal $s(t - t_0)$ the original time origin of the graph shifts to $t = t_0$, such that the delayed signal $s(t - t_0)$ appears shifted to the *right* of the origin. Similarly, the signal $s(t + t_0)$ would appear shifted to the *left* of the origin.

Signals can be delayed without being distorted. Just as with ideal amplification, the shape of a signal is not affected by a simple time shift. As an extreme example, consider a signal consisting of recorded music, played back at high volume many years after it was performed in a recording studio; the resulting delayed, amplified signal will ideally be indistinguishable from the original. Combined, the effects of amplifying and shifting describe an *ideal distortionless system*, in which an input signal $s(t)$ is related to an output signal by the expression $s(t) \rightarrow ks(t - t_0)$, where $k$ and $t_0$ are constants.

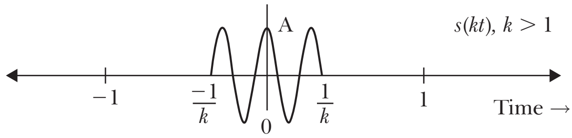



**FIGURE 1.5** The system $s(t) \to s(kt)$ *stretches* or *squeezes* the signal $s(t)$.

**Scaling** A third simple operation that can be performed on a signal $s(t)$ is that of multiplying $t$ by some constant $k$, as in Figure 1.5. The signal $s(t) \to s(kt)$ will be scaled for $k \neq 1$. For $k > 1$, the signal $s(kt)$ will be *squeezed* along the time axis, with respect to $s(t)$ as shown here, while for $0 < k < 1$, $s(kt)$ will be *stretched* in time with respect to $s(t)$.

As implied by the descriptions *stretch* and *squeeze*, the scaling operation acts to distort a signal. For example, a tape recording $s(t)$ will produce the output signal $s(kt)$, where $k$ is a function of the tape speed. Played at slow speed, where $0 < k < 1$, the signal $s(kt)$ is stretched in time and takes longer to play back. Alternatively, when played at high speed where $k > 1$, the signal, is squeezed in time and plays back quickly. The resulting distortion is also evident in the audible frequency changes corresponding to each of these two examples. This effect illustrates the important fact that modifying the time-varying characteristics of a signal has a direct effect on its frequency characteristics.

## 1.2.1 Linear Combination

A linear combination of signals is defined as the signal formed using only the addition of signals and their multiplication by constant values. For example, if $s(t)$ is formed from a linear combination of the signals $a(t)$, $b(t)$, and $c(t)$, then $s(t) = k_a a(t) + k_b b(t) + k_c c(t)$, where $(k_a, k_b, k_c)$ are constants. Conversely, the signal $s(t) = a^2(t)$ is *not* formed from a linear combination of signals because the signal is squared.

## 1.2.2 Addition and Multiplication of Signals

The result of the addition or multiplication of two signals is another signal expressing their sum or product at every point in time. For example, consider two signals $a(t)$ and $b(t)$ as defined below and shown in Figure 1.9.

$$a(t) = \begin{cases} 1 & \text{if } -1/2 \leq t < +1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$b(t) = \begin{cases} t+1 & \text{if } -1 \leq t < +1 \\ 0 & \text{otherwise} \end{cases}$$

A system that mathematically adds two signals together may be called either an *adder* or a *summer*, and will be sketched as shown in Figure 1.6. The output of this adder for the above signals is calculated by considering what the numerical result of the sum would be in each defined region, at every point in time, as shown in Figure 1.9.

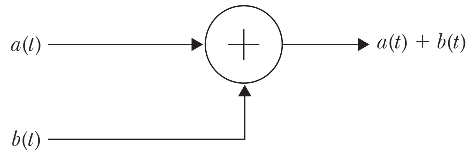



**FIGURE 1.6** Adder (Summer)—The output of this system is the sum of the input signals $a(t)$ and $b(t)$.

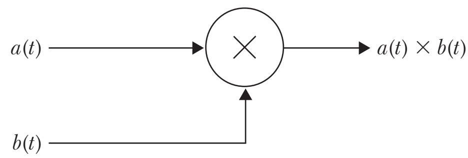



**FIGURE 1.7** Multiplier (Mixer)—The output of this system is the product of the input signals $a(t)$ and $b(t)$.

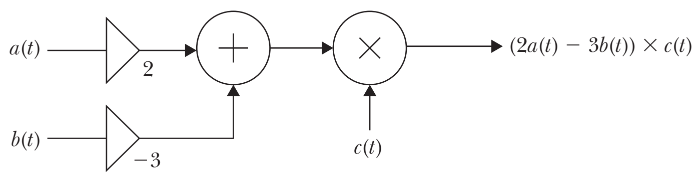



**FIGURE 1.8** System block diagrams can be combined to create larger systems. The output is $[2a(t) - 3b(t)] \times c(t)$.

$$c(t) = a(t) + b(t) = \begin{cases} t+1 & \text{if } -1 \leq t < -1/2 \\ t+2 & \text{if } -1/2 \leq t < +1/2 \\ t+1 & \text{if } +1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases}$$

A system that mathematically multiplies two signals together may be called either a *multiplier* or a *mixer*, and will be sketched as shown in Figure 1.7. The output of this multiplier is calculated by considering what the numerical result of the product would be in each defined region, at every point in time, as shown in Figure 1.9.

$$d(t) = a(t) \times b(t) = \begin{cases} t+1 & \text{if } -1/2 \leq t < +1/2 \\ 0 & \text{otherwise} \end{cases}$$

Multiplication by a constant can be indicated by an *amplifier* block drawn as a triangle as seen in Figure 1.8; the accompanying number represents the gain of the amplifier and may be negative. This system has three inputs and one output as $[2a(t) - 3b(t)] \times c(t)$.

Operations described by system diagrams are performed graphically in Figure 1.9. Note that the addition or multiplication is performed point-by-point on the two operands to produce the output signal. It is often much easier to use graphs to illustrate signals than it is to appreciate functional definitions like the ones shown above. Furthermore, the above results can be sketched and understood by inspection on a graph without the need to specifically compute each point of the result. The graphical approach often simplifies a computation while greatly adding to the understanding of the system under study.

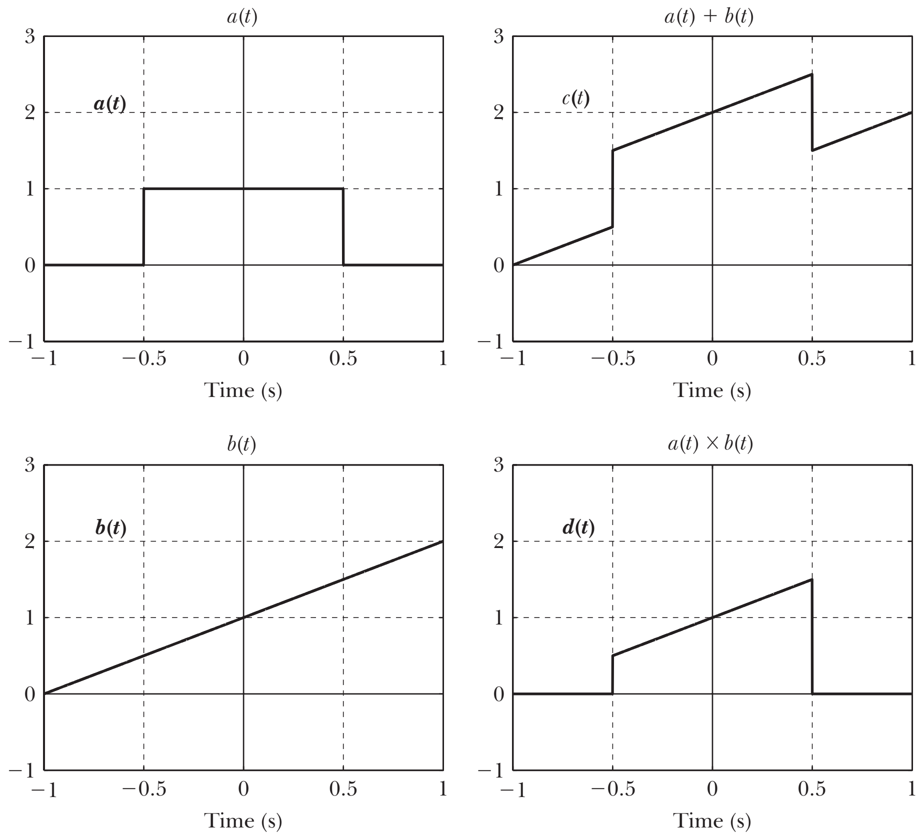



**FIGURE 1.9** Adding and multiplying signals—$c(t) = a(t) + b(t)$; $d(t) = a(t) \times b(t)$.

## 1.2.3 Visualizing Signals—An Important Skill

As the examples above show, many signals may be studied by drawing simple sketches that do not always require a computer package or a calculator to produce. In practice, it is not always possible to work through the equations or to sit down in front of a computer. For example, when using a sophisticated digital instrument to measure an unusual signal, the appearance of the signal on a screen and a knowledge of how its behavior relates to the mathematics may be all there is to guide a confident measurement. Furthermore, the increasing reliance on computers to perform numerical calculations and to produce graphs means that numerical results often will be obtained without seeing the intermediate steps to a calculation.[3] Are those results correct? Ultimately, the same can asked for any results obtained purely by working through the mathematics by hand. Without a basic understanding of the underlying signal theory—without knowing what to expect—there is little chance of being able to interpret an unexpected result or to notice unusual or unreasonable answers. Once this appreciation is achieved, however, the essential mathematics can be approached with confidence, and, importantly, there will always be some means of *checking the answer.* In the following sections, a few basic signals are introduced along with simple mathematical manipulations that will allow them to be used as components in

[3] The use of MATLAB is no exception in this regard.

constructing much more complicated signals. The appreciation of how a signal may be formed from simpler *components* is an important step in successful signals analysis.

### 1.2.4 Introduction to Signal Manipulation Using MATLAB

MATLAB is a powerful tool that may be readily used to view signals and their manipulations with simple command line instructions. Once a signal or a technique has been defined for use in MATLAB, it is immediately tempting to try variations, which can shed much light on the behavior of signals and systems.

In this section, the signals and graphs in Figure 1.9 are replicated. The rectangle and straight line signals are first defined, along with their sum and product on the time interval $[-1, +1]$ or $-1 \leq t \leq +1$. This example can be followed step by step in the MATLAB command window, or it may be copied as a sequence of commands to be executed all at once. The use of such a *script* or *m-file* is to be recommended when an operation is to be used repeatedly or tried with different parameters. Help with any MATLAB function such as `plot()` is available on the command line as `help plot`. A summary of MATLAB commands may be found in Appendix D.

**Defining Signals** A time axis called $t$ is defined stretching from $t = -1$ to $t = +1$ s with increments every 0.001 s.

```
t = -1 : 0.001 : 1; % define a time axis every 0.001 s
```

If the final semicolon is omitted, all 2001 points in the resulting vector will be listed. Comments in MATLAB follow the % character. *The effective use of comments can greatly increase the readability of a sequence of commands.* Next, functions $a(t)$ and $b(t)$ are defined on the same time interval $t$. Let $a(t)$ be a simple rectangle centered at the origin using the built-in MATLAB function `rectpuls(t)`.[4] Let $b(t)$ be a straight line with slope = 1 and intercept = 1 using the formula $b = t + 1$. The sum $c(t)$ and product $d(t)$ can then be computed directly.

```
a = rectpuls(t);   % define a(t) = a rectangle
b = t + 1;         % define b(t) = a unit slope line
c = a + b;         % create c(t) = the sum
d = a .* b;        % create d(t) = the product
```

It is necessary to use a dot as `d = a .* b` to multiply the two signals $a(t)$ and $b(t)$ point by point.[5]

**Basic Plotting Commands** Finally, the results may be shown graphically with the `plot()` command. Although MATLAB automatically scales the axes to fit, the horizontal and vertical axes may be set to run from $-1$ to $+1$ and $-1$ to $+3$, respectively, using the `axis()` command. Further commands serve to enhance the appearance of the graph by adding grid lines and a title, and by labelling the time axis. Let only the signal $c(t)$ from Figure 1.9 be displayed.

---

[4] In MATLAB releases prior to 2011a, the command `rectpuls` returns a *logical* result. The command `double()` can be used to convert a variable to *double* format as required.

[5] Otherwise the simple `d = a * b` specifies an invalid matrix multiplication.

```
figure(1);                  % start figure 1
plot(t,c);                  % plot the sum
axis([-1 1 -1 3 ]);         % change (x,y) limits
grid on;                    % include grid lines
title ('a(t) + b(t)');      % add a title
xlabel ('time (sec)');      % label the time axis
```

**Multiple Plots on One Figure** All four signals may be shown on one figure as in Figure 1.9 using the `subplot` command, where `subplot(2,2,x)` selects graph $x$ from four graphs in 2 rows and 2 columns.

```
figure(2);           % start figure 2
clf;                 % clear any old plots
subplot(2,2,1);      % plot 1 of 4
plot(t,a);           % plot signal A
subplot(2,2,2);      % plot 2 of 4
plot(t,c);           % plot signal C
subplot(2,2,3);      % plot 3 of 4
plot(t,b);           % plot signal B
subplot(2,2,4);      % plot 4 of 4
plot(t,d);           % plot signal D
```

The use of multiple plots is helpful when comparing related results, and each subplot may be embellished as shown above to complete the four-plot figure.

As each of the following basic signals is introduced, the corresponding MATLAB code is presented.

## 1.3 A Few Useful Signals

In this section, some of the most fundamental and useful signals are introduced. By applying the manipulations of shifting and scaling and the addition or multiplication operations, many other signals can be formed. The study of a certain signal can be greatly simplified if it can be recognized as being composed of some basic signal types.

### 1.3.1 The Unit Rectangle rect(*t*)

One of the most versatile signals of interest is a rectangular pulse called the *unit rectangle*:

**DEFINITION 1.2 Unit Rectangle**

*Let rect*(t) *be a function of* t *such that:*

$$\text{rect}(t) = \begin{cases} 1 & \text{if } |t| \leq \dfrac{1}{2} \\[2ex] 0 & \text{if } |t| > \dfrac{1}{2} \end{cases}$$

*then rect*(t) *will be called the* unit rectangle.

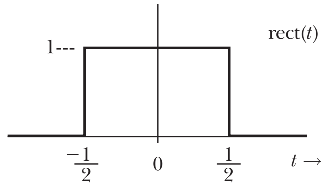



**FIGURE 1.10** The unit rectangle rect($t$) has width = 1, height = 1, and area = 1.

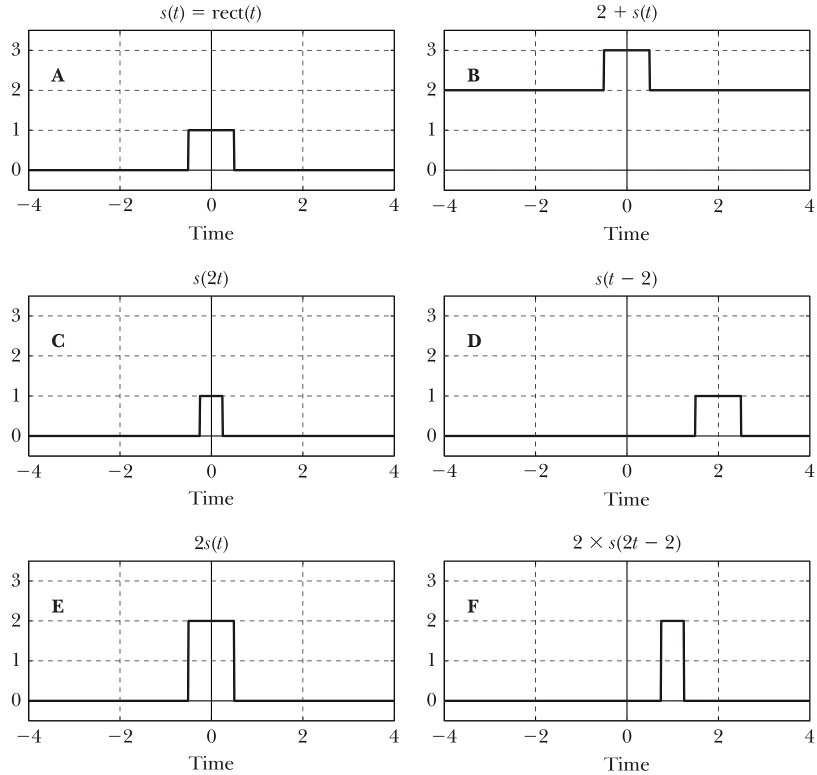



**FIGURE 1.11** Mathematical variations manipulate a signal $s(t)$ = rect($t$).

As seen in Figure 1.10, the signal rect($t$) has unit amplitude, unit width, and unit area. The unit area is shown by the integral:

$$\int_{-\infty}^{+\infty} \text{rect}(t)\, dt = 1$$

This simple signal may represent a switch being turned on, then off. Alternatively, it could be a digital pulse within a computer, or it may be viewed as a *window* about the origin ($t = 0$) of the graph intended to exclude signals outside of the time interval $[-\frac{1}{2}, +\frac{1}{2}]$. For example, the signal of Figure 1.2 could be seen as a cosine multiplied by a rectangle. This same time interval is also expressed using the absolute value as $|t| \leq \frac{1}{2}$. Various definitions of rect($t$) vary in how its value at the discontinuities is treated, and rect$(+\frac{1}{2})$ or rect$(-\frac{1}{2})$ may be seen as 1 or $\frac{1}{2}$ or 0. In MATLAB, the unit rectangle is defined as the function `rectpuls(t)` that spans the time interval $[-\frac{1}{2}, +\frac{1}{2}]$ such that the value `rectpuls(-0.5)` returns 1, while `rectpuls(0.5)` returns 0.

In Figure 1.11, the unit rectangle is manipulated mathematically by employing the constant 2 in various ways to accomplish amplifying, scaling, and shifting, and the

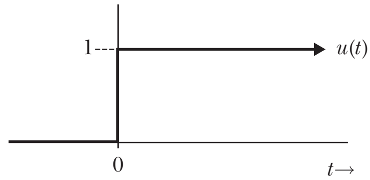



**FIGURE 1.12** The unit step function $u(t)$ represents the onset of a signal at $t = 0$.

addition of a DC-offset. In the last example (F) several operations are achieved at once. In particular, note that $2s(2t - 2)$ does not shift the rectangle by 2 seconds; rather, since $s(2t - 2) = s(2(t - 1))$ the narrow rectangle is shifted by only one second.

## 1.3.2 The Unit Step *u(t)*

Turning on a circuit that was previously off, or closing a switch to allow a signal to suddenly appear, produces a sudden *step* in voltage or current. The unit step function written as $u(t)$, serves to describe such events mathematically. The function $u(t)$ is also known as the *Heaviside function*. One special application of the step function $u(t)$ in systems analysis is in establishing the overall behavior of a given system. By using $u(t)$ as an input signal, the resulting output or *step response* uniquely characterizes the system.

The unit step function $u(t)$ is defined as *a signal that is zero for all times prior to a certain instant and unity for all values of time following* [2], or:

**DEFINITION 1.3 Unit Step**

*Let* u(t) *be a function of* t *such that:*

$$u(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$$

*then* u(t) *will be called the* unit step.

Any signal $s(t)$ can be *switched on* at time $t = 0$ by considering the product $s(t) \times u(t)$, as shown in Figure 1.13.

$$s(t) \times u(t) = \begin{cases} 0 & \text{if } t < 0 \\ s(t) & \text{if } t \geq 0 \end{cases}$$

A switch closure at any other time $T$ can be accomplished by shifting the unit step function. For example, the signal $u(t - 5)$ represents a unit step beginning at $t = +5$ s.

$$s(t) \times u(t-5) = \begin{cases} 0 & \text{if } t < 5 \\ s(t) & \text{if } t \geq 5 \end{cases}$$

Unit step functions can be combined to model the effect of turning a signal ON, then OFF after a time. For example, the unit rectangle $\text{rect}(t)$ can be defined in terms of step functions as $\text{rect}(t) = u(t + \frac{1}{2}) - u(t - \frac{1}{2})$.

In MATLAB, the unit step $u(t)$ can be accomplished by using the conditional expression `(t >= 0)`, which returns *true* (1) when $t \geq 0$ and *false* (0) otherwise.

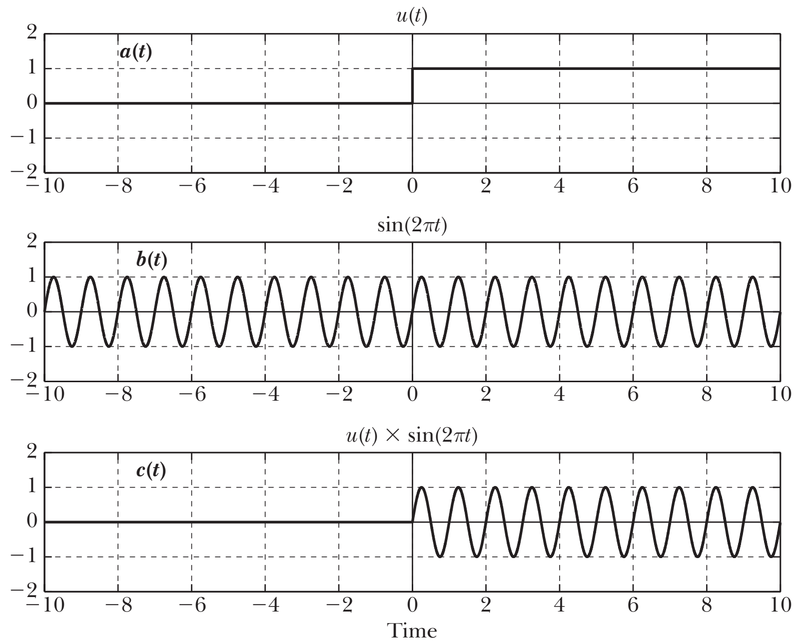



**FIGURE 1.13** The unit step can be used to model the onset of a signal $b(t)$ at a specific time: $c(t) = a(t) \times b(t)$.

Such logical results can be converted to numeric using the command `double()` as required. For example:

```
t = -3:5

t = -3   -2   -1   0   1   2   3   4   5

u = double(t> = 0)

u = 0    0    0    1    1   1   1   1   1
```

Like the unit rectangle, definitions of $u(t)$ vary in the value at the discontinuity. In MATLAB, the unit step is implemented directly as `heaviside(t)`, which returns the value 0.5 when $t = 0$.

## 1.3.3 Reflection about $t = 0$

Given a signal $s(t)$, the signal $s(-t)$ is *time reversed* and would appear reflected about the origin $t = 0$. In Figure 1.14, the unit step as $u(-t)$ may serve to represent a signal that is switched off at time $t = 0$.

## 1.3.4 The Exponential $e^{xt}$

The exponential signal $e^{xt}$ plays many different roles in signals theory depending on whether $x$ is positive or negative, real or complex, and the signal is often multiplied

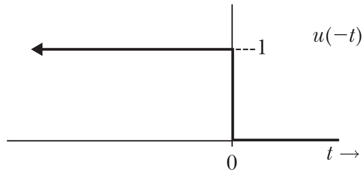



**FIGURE 1.14** The time reversed unit step function $u(-t)$ represents a signal that switches off at $t = 0$.

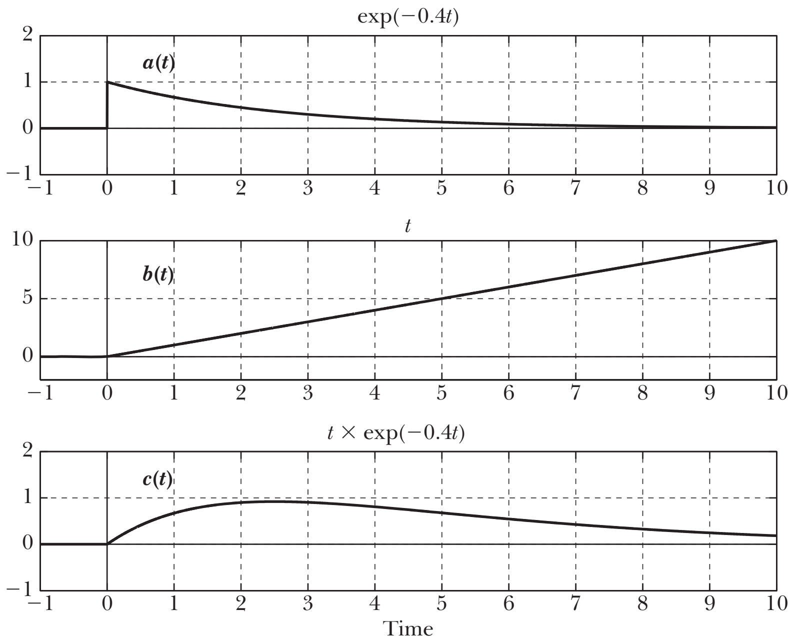



**FIGURE 1.15** The exponential signal $a(t) = e^{xt}$ can be used to limit the excursions of a signal $b(t)$ as $t$ increases by multiplying to give $c(t) = a(t) \times b(t)$. To be effective, the value $x$ must be chosen to overcome the growing tendency of the signal $b(t)$. Here, $x = -0.4$. These example signals are assumed to exist only for $t \geq 0$.

by the unit step to exist only for $t \geq 0$. In MATLAB the exponential function $e^x$ is implemented as `exp(x)`.

Real values of $x < 0$ lead to a function that diminishes over time and may serve to limit the excursions of another signal that grows over time. In Figure 1.15, the signal $b(t)$ is a straight line with unit slope that grows without bounds as $t$ increases. When multiplied by the exponential signal $a(t) = e^{xt}$ with $x = -0.4$, the resulting signal $c(t) = a(t) \times b(t)$ tends to zero as $t$ grows. All signals in this figure are assumed to exist only for $t > 0$. To be successful, the signal to be limited in this way must grow more slowly than does $e^{xt}$, something that can be adjusted by changing $x$ and that can be successful for most engineering signals of interest. In particular, note that the signal $b(t)$ has an integral that goes to infinity, while the new signal product $c(t)$ is integrable.

For complex values of $x$ as $s(t) = e^{j2\pi ft}$, the exponential serves to define real and imaginary cosines and sines. The complex exponential is fundamental to study of signals using the Fourier series and Fourier transform and together with the real exponential defines the Laplace transform of Chapter 7 and the *z*-transform of Chapter 9.

## 1.3.5 The Unit Impulse $\delta(t)$

The impulse can be described as *a pulse that begins and ends within a time so short that it may be regarded mathematically as infinitesimal although the area remains finite* [2]. The *unit impulse* or *delta function* is written using the lowercase Greek letter delta, as $\delta(t)$ and is defined as having unit area and a value of zero everywhere except at the origin.

**DEFINITION 1.4 Unit Impulse**

*Let* $\delta(t)$ *be a function of* t *such that:*

$$\delta(t) = 0 \quad \text{for } t \neq 0$$

*and*

$$\int_{-\infty}^{+\infty} \delta(t)\, dt = 1$$

*then* $\delta(t)$ *will be called the* unit impulse.

Like the unit step, the *unit impulse* is frequently employed in signals analysis, and the two signals are closely related as the unit impulse may be recognized as the time derivative of the unit step function, where:

$$\delta(t) = \frac{d}{dt}u(t) \tag{1.1}$$

Consider the time derivative (slope) of the unit step function from Figure 1.16. The slope is zero everywhere except at the origin ($t = 0$) where its value goes to infinity. Conversely, the time integral (area) of the impulse must be the unit step function and for $t < 0$ the area of the impulse is zero, while for $t > 0$ the area equals 1. In conclusion, the signal $\delta(t)$ is infinitely narrow, infinitely high, and has area equal one. The practical impossibility of such a signal does not detract from its utility as a tool of signals analysis.[3]

The special nature of the unit impulse may be explored by using the model shown in Figure 1.17. In this figure, the impulse is shown realistically having *width* = $2a$ and *area* = 1, such that the corresponding integral has a step with *slope* = $1/2a$. As $a$ is made smaller, the step increases in slope, the pulse gets narrower and higher, and in the limit as $a \to 0$, both the unit step function $u(t)$ and the unit impulse $\delta(t)$ are achieved.

In systems analysis, the function $\delta(t)$ is typically used to establish the overall behavior of a given system. It will be shown that by using $\delta(t)$ as an input signal, the output *impulse response* uniquely characterizes the system. If the behavior of a linear system is known for either a step input or an impulse input, then recognizing these relationships will allow its behavior for a step function to be computed, and vice-versa. It will be shown that if either the step response or the impulse response of a linear system is known, the system behavior can be readily predicted for *any* input signal $s(t)$.

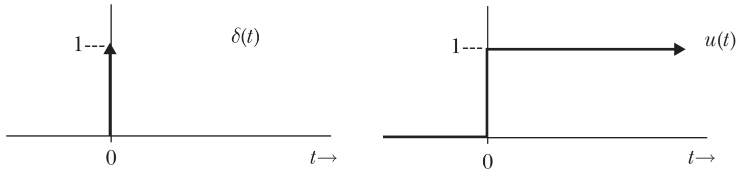



**FIGURE 1.16** The unit impulse $\delta(t)$ is the time derivative of the unit step $u(t)$.

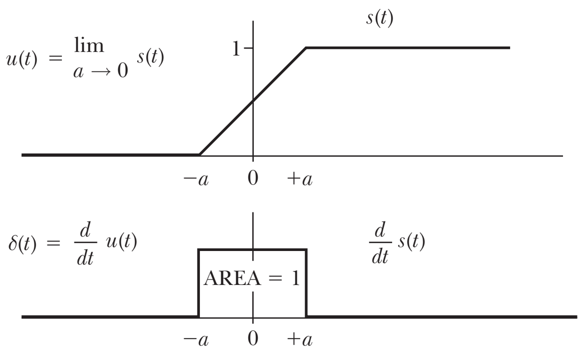



**FIGURE 1.17** As $a \rightarrow 0$, the ramp tends to a step function $u(t)$ and its derivative narrows to $\delta(t)$ while retaining unit area.

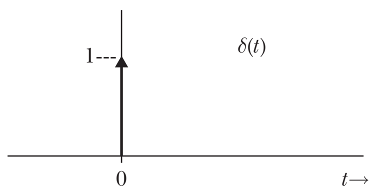



**FIGURE 1.18** The unit impulse $\delta(t)$ has area = 1, shown as a unit length arrow at $t = 0$.

In practical terms, the sudden and very brief appearance of an impulse signal could be likened to a single hand clap or a hammer blow or (electrically) a spark or a stroke of lighting.

**Sketching the Unit Impulse** The impulse function $\delta(t)$ is generally drawn as an arrow one unit long pointing upward from the origin as in Figure 1.18, but this sketch must be carefully interpreted in light of the definition of $\delta(t)$.

Although a unit-height spike at $t = 0$ serves to illustrate the $\delta(t)$ function graphically, it is important to remember that although $\delta(t)$ has unit area, the value of $\delta(0)$ tends to infinity. Consequently, the unit impulse is drawn as a vertical arrow (pointing to infinity) *with a height corresponding to its area.* In other words, the height reflects the unit area of the impulse and the arrow, its infinite amplitude. It follows that the signal $A\delta(t)$ has area $A$ and can be drawn as an arrow of height $A$. Variations on the impulse function are shown in Figure 1.19.

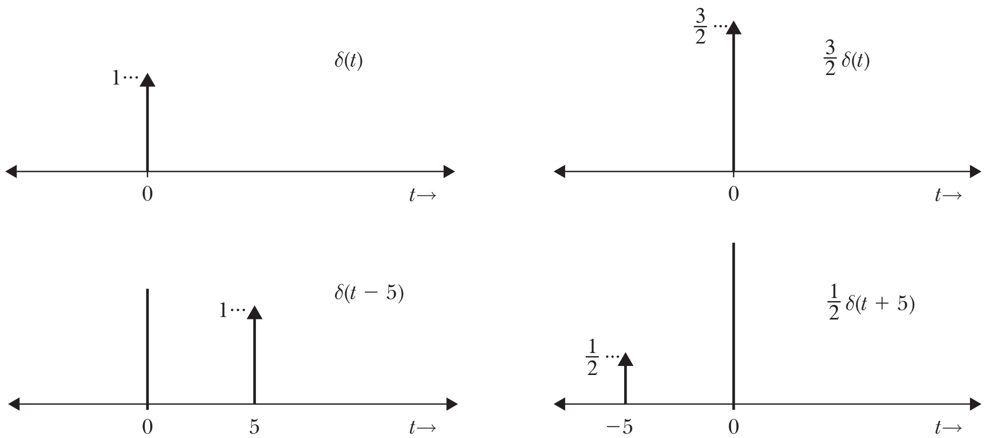



**FIGURE 1.19** The impulse function $\delta(t)$ is sketched as a vertical arrow going to infinity, drawn with a height reflecting its area.

**The Sifting Property of $\delta(t)$** One of the most used properties of $\delta(t)$ is revealed when integrating a signal $s(t)$ multiplied by $\delta(t)$:

$$\int_{-\infty}^{+\infty} s(t)\,\delta(t)\,dt = s(0) \tag{1.2}$$

since the product must be zero for all $t \neq 0$, yet $\delta(t)$ has unit area.

This result may be extended to give the general result shown below, where $\delta(t-k)$ is a shifted impulse as in the examples of Figure 1.19.

$$\int_{-\infty}^{+\infty} s(t)\,\delta(t-k)\,dt = s(k) \tag{1.3}$$

since the product must be zero for all $t \neq k$, yet $\delta(t-k)$ has unit area. Essentially, the impulse function is moved across $s(t)$ to extract the value of $s(t)$ at time $t = k$.

Even within very complicated expressions, the integral of a product with $\delta(t)$ can often be obtained directly. For example, the following unnecessarily complicated integral includes a delta function:

$$\int_{-\infty}^{+\infty} 3\left[\frac{t}{\pi}\right]^2 \cos^4(t)\,\delta(t-\pi)\,dt \tag{1.4}$$

By inspection, the product term is zero except at $t = \pi$, and the overall result is $3 \times (\pi/\pi)^2 \times \cos^4(\pi) = 3$. Note that unless the limits of integration include the position where the impulse is located, the integral is simply zero.

**Sampling Function** The sifting property of the impulse can be usefully applied to extracting the value of a signal at a specific time $t_0$. In Figure 1.20 the impulse $\delta(t-3)$ is shifted 3 seconds to the right. When this impulse is multiplied by the signal $s(t)$, the impulse area takes on the corresponding value of $s(t)$ at time $t = 3$ and the product is zero everywhere else. The new $s(3)$ is sketched as the height of the impulse arrow. The new area of the impulse is a *sample* of $s(t)$ taken at exactly $t = 3$.

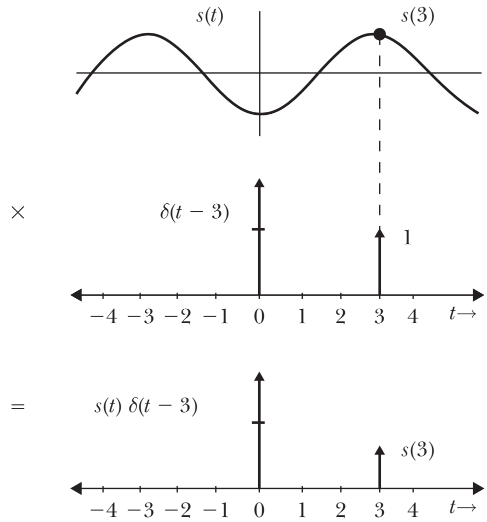



**FIGURE 1.20** Sampling Function—The value of $s(t)$ at $t = 3$ is found by multiplying $s(t)$ by the shifted impulse $\delta(t - 3)$. The new area of the impulse is $s(3)$.

## 1.4 The Sinusoidal Signal

The sinusoidal signal is fundamentally important in the description and analysis of signals and systems. The sinusoid is usually defined in terms of the sine function as:

**DEFINITION 1.5 Sinusoid**

*Any signal* s(t) *of the form*

$$s(t) = A\sin(2\pi f_0 t + \Phi)$$

*with constants* $(A, f_0, \Phi)$ *is called a* sinusoid.

The term *sinusoid* will be employed generically to mean all manner of sine or cosine signals as may be generated from this definition. Using the sinusoid as a starting point, a general description of many other signals can be developed. Where it is desirable that a signal be based specifically on a sine function, the above definition will be called a *general sine* function. Similarly, a *general cosine* may be defined as:

**DEFINITION 1.6 General Cosine**

*Any signal* s(t) *of the form*

$$s(t) = A\cos(2\pi f_0 t + \Phi)$$

*with constants* $(A, f_0, \Phi)$ *will be called a* general cosine.

In the definition $s(t) = A\cos(2\pi f_0 t + \Phi)$, three distinct parameters may be identified as shown in Figure 1.21. The constant terms ($A$ = amplitude, $f_0$ = frequency, $\Phi$ = phase) uniquely describe this cosine, where phase is written using the uppercase Greek letter *phi* ($\Phi$). For example, in the signal $s(t) = \cos(t)$, the values

$$s(t) = \underset{\text{AMPLITUDE}}{\underset{\uparrow}{A}} \cos(2\pi \underset{\text{FREQUENCY}}{\underset{\uparrow}{f_0}} t + \underset{\text{PHASE}}{\underset{\uparrow}{\Phi}})$$

**FIGURE 1.21** General Cosine—The parameters $(A, f_0, \Phi)$ uniquely define the waveform for all time.

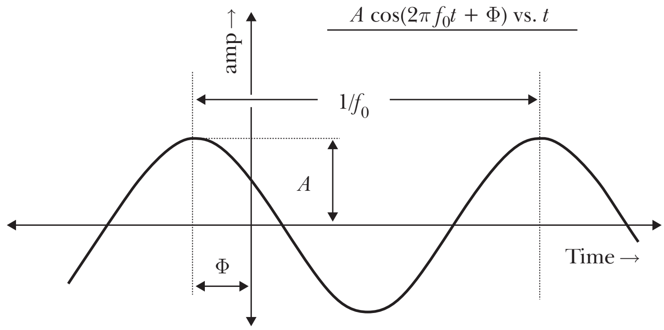



**FIGURE 1.22** General Cosine—The defining parameters $(A, f_0, \Phi)$ can readily be identified in this time domain graph.

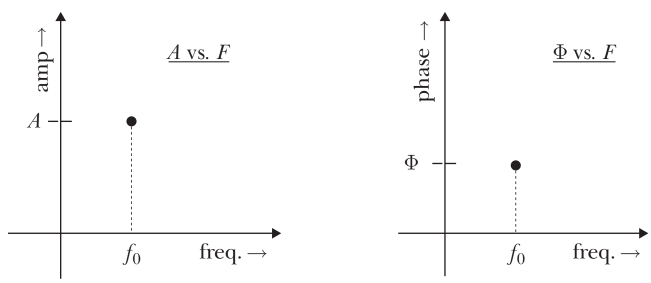



**FIGURE 1.23** General Cosine—The defining parameters $(A, f_0, \Phi)$ can readily be identified in these graphs sketched as a function of frequency.

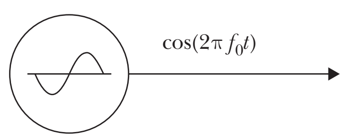



**FIGURE 1.24** Cosine Generator—An oscillator operating at $f_0$ Hz generates the signal $\cos(2\pi f_0 t)$.

$A = 1$, $f_0 = 1/2\pi$, and $\Phi = 0$ can be identified by inspection. Conversely, it is only necessary to specify this set of three constants $(A, f_0, \Phi)$ to completely define this waveform for all time. It follows that all three parameters can be identified in, and determined from, the time domain (Amplitude vs. Time) graph of the signal shown in Figure 1.22.

Determination of the phase $(\Phi)$ depends on having a reference point for the origin $(t = 0)$, and in any case the phase does not affect the overall appearance or amplitude of the cosine waveform. Now, while these three parameters fully describe the waveform, the time domain sketch is not the only way to represent them graphically. For example, it is possible to plot Amplitude vs. Frequency $(A, f)$ on one graph,

$$s(t) = \underset{\text{AMPLITUDE}}{\overset{\uparrow}{A}} \cos(\underset{\text{FREQUENCY}}{\overset{\uparrow}{\omega_0}} t + \underset{\text{PHASE}}{\overset{\uparrow}{\Phi}})$$

**FIGURE 1.25** General Cosine (alternate form)—The parameters $(A, \omega_0, \Phi)$ uniquely define the waveform for all time, where $\omega_0 = 2\pi f_0$. Compare to Figure 1.21.

and Phase vs. Frequency $(\Phi, f)$ on another, showing components at $f_0$ in both cases. Such a graph will be called the *frequency-domain* representation of $s(t)$. Because the same constants $(A, f_0, \Phi)$ can be identified, the two graphs in Figure 1.23 convey exactly the same information as the time domain cosine waveform in Figure 1.22.

## 1.4.1 The One-Sided Cosine Graph

An important assumption has been made in the frequency graphs of Figure 1.23, in that the parameters shown specifically correspond to those found in the general cosine equation $s(t) = A\cos(2\pi f_0 t + \Phi)$. The same graph would represent a different signal if it were defined in terms of the general sine. Consequently, this particular amplitude vs. frequency graph will be called here a *one-sided cosine graph* of the signal $s(t)$. The term *one-sided* refers to the fact that all the points lie in the region for which $f_0 > 0$. In this way, each point on the one-sided cosine representation of $s(t)$ illustrates *the component of* $\cos(2\pi f_0 t)$ *found at frequency* $f_0$. Consequently, each component is sketched as a single point $(f_0, A)$ on the one graph, and at $(f_0, \Phi)$ on the other. The dashed vertical lines are included for clarity.

Similar amplitude vs. frequency graphs could be defined as required. For example, the *one-sided sine graph* might illustrate *the component of* $A\sin(2\pi f_0 t + \Phi)$ *found at frequency* $f_0$. In every case, the representation of signals as frequency-based terms depends on some underlying definition that must be specified. Similarly, other forms of the *general cosine* are possible. The frequency term $f_0$ could be replaced by $1/T_0$, where $T_0$ is the period (seconds). It is commonplace in the analysis of specific systems[6] and applications and in many textbooks to use the lowercase Greek letter *omega* as $\omega = 2\pi f$, where the angular frequency $\omega$ has units of

### EXAMPLE 1.1 (Cosine Parameters)

*Determine the amplitude, frequency, and phase of the signal* $g(t) = \cos(t + \pi/4)$.

**Solution:**
From the general form of cosine $s(t) = A\cos(2\pi f_0 t + \Phi)$ it can be seen that for this $g(t)$:

| | |
|---|---|
| Amplitude: | $A = 1$ |
| Frequency: | $f_0 = 1/2\pi$ Hz |
| Phase: | $\Phi = \pi/4$ rad (shifted towards the left) |
| Period: | $T_0 = 1/f_0 = 2\pi$ s |
| Omega: | $\omega_0 = 2\pi f_0 = 1$ rad/s |

[6] Such as the Laplace transform of Chapter 7 and the *z*-transform of Chapter 9.

The result can readily be confirmed using MATLAB, as shown below. Note that `pi` $= \pi$ is pre-defined in MATLAB.

```
t = -10:0.01:10;   % define time interval = 20s
s = cos(t+pi/4);   % define s(t)
plot(t,s);         % plot s(t)
grid on;
```

The amplitude, period, and phase found above can be identified in Figure 1.26. The positive phase change shifts the cosine waveform towards the left.

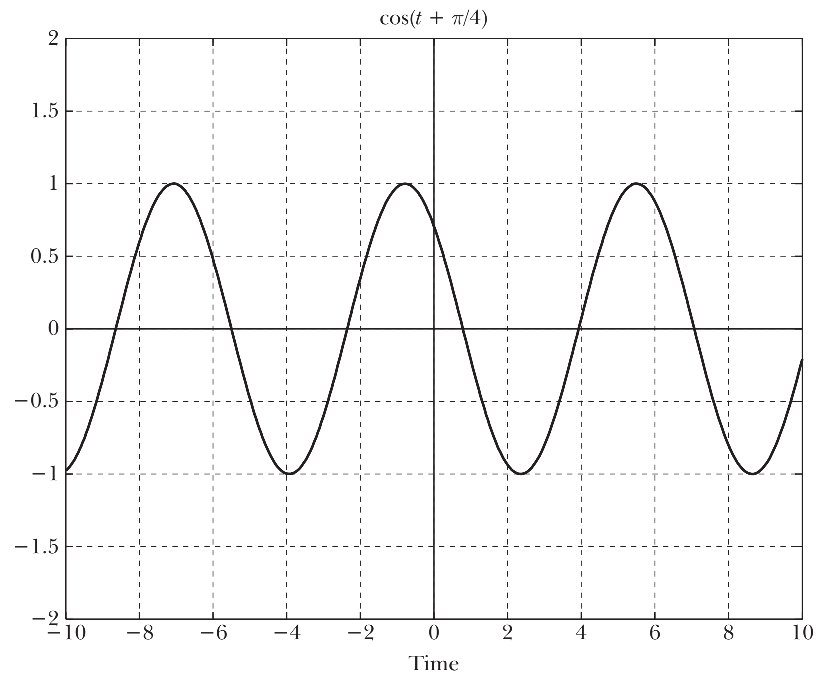



**FIGURE 1.26** The cosine $\cos(t + \pi/4)$ is shifted left by $\pi/4$ s. Period: $T = 2\pi$ s.

rad/s and the general cosine becomes $s(t) = A\cos(\omega_0 t + \Phi)$ as seen in Figure 1.25. Whatever the form used, the same parameters for amplitude, frequency, and phase can always be identified in the mathematical expression of a sinusoidal signal.

## 1.4.2 Phase Change—$\Phi$

The phase term $\Phi$ (the Greek letter *phi*) indicates the origin or *starting point* of the cosine waveform. When $\Phi = 0$, the signal $s(t) = A\cos(2\pi f_0 t)$ has a maximum value at $t = 0$. When the phase is non-zero, the cosine appears shifted along the time axis as was seen in Figure 1.26. In effect, changing the phase is equivalent to shifting in time, although the units of phase (radians) and of time (seconds) imply that the numerical value of the shift will be different in each case. Like a time change, a phase change does not affect the amplitude or the frequency of a waveform. If a cosine is shifted in time or phase, the Amplitude vs. Frequency graph would not be affected and only the

Phase vs. Frequency plot would change. The exact relationship between a phase change and a time shift is of fundamental importance in the design of systems.

## 1.5 Phase Change vs. Time Shift

Consider the signal $s(t) = A\cos(2\pi f_0 t)$. This signal may be delayed by a time $t_0$ (or shifted along the time axis a distance $t_0$ seconds) by letting $t = (t - t_0)$ as follows:

$$s(t - t_0) = A\cos(2\pi f_0(t - t_0)) = A\cos(2\pi f_0 t - 2\pi f_0 t_0) = A\cos(2\pi f_0 t + \Phi),$$
$$\text{where } \Phi = -2\pi f_0 t_0$$

Since $t_0$ is a constant, it is possible to isolate the phase term $\Phi = -2\pi f_0 t_0$ in the above equation. This result shows that when a waveform is delayed by a time $t_0$, the corresponding phase shift also depends on the frequency $f_0$, as $\Phi = -2\pi f_0 t_0$. In other words, *when shifted the same amount in time, different frequency cosines will not be shifted by the same number of radians.*

In Figure 1.27, a new origin sketched at $t = 1$ shifts two cosine waveforms by exactly one second. Although both have the same time shift, the upper signal $a(t) = \cos(\pi t/2)$ is effectively shifted $\pi/2$ rad, while the signal $b(t) = \cos(3\pi t/2)$ undergoes a phase shift of $3\pi/2$ rad.

In general, for a given frequency the phase change ($\Phi$) equivalent to a time delay of $t_0$ is described by the linear function of frequency $\Phi(f) = -2\pi f_0 t_0$ as shown in Figure 1.28. If the frequency is specified using $\omega = 2\pi f$ rad/s, then the same straight line is $\Phi(\omega) = -\omega t_0$ as shown.

This straight-line phase vs. frequency relationship is characteristic of a delay, or a simple shift in time. In dealing with circuits and signals, time delays are unavoidable. For example, an electrical signal takes some time to pass through an amplifier circuit, and a transmitted radio signal takes time to propagate through the air or via a satellite. Consequently, the above straight line graph showing *Phase Change Corresponding to a Time Delay* will be encountered frequently.

In summary, a time shift does not affect the overall shape of a waveform in the time domain, and its effect is not visible at all in an amplitude vs. frequency graph. Furthermore, a change in phase as a function of frequency corresponding to the

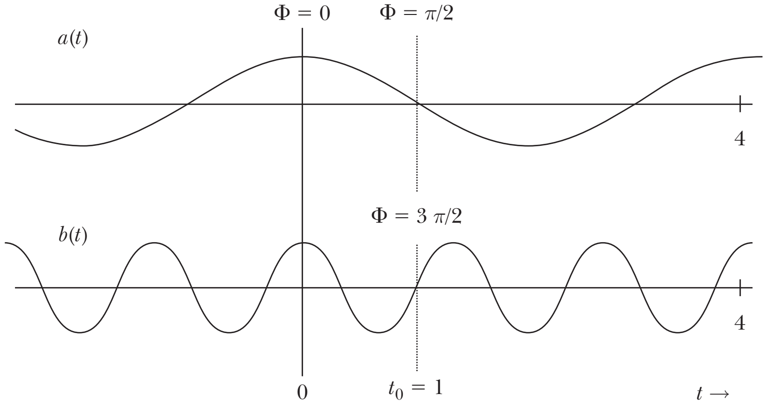



**FIGURE 1.27** Phase Change Corresponding to a Time Delay of $t_0 = 1$ s. When waveforms $a(t)$ and $b(t)$ are both shifted by $t_0 = 1$ s, the phase change $\Phi$ rad in each depends on frequency.

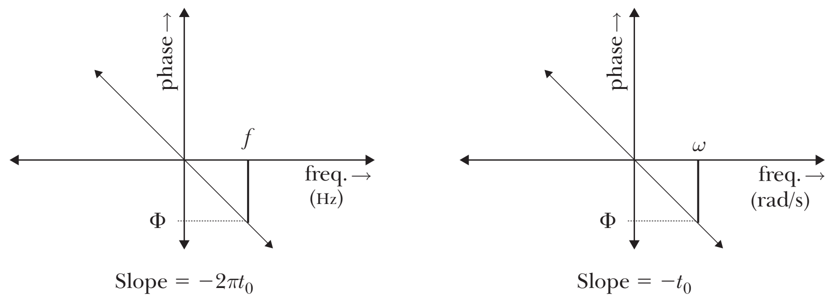



**FIGURE 1.28** Phase Change Corresponding to a Time Delay of $t_0$ s. This straight-line phase change as a function of frequency characterizes a simple time delay. The two graphs show frequency as $f$ Hz or $\omega = 2\pi f$ rad/s.

above graph amounts to a simple shift in time. Note that the slope of this line equals $-2\pi t_0$; logically, an unshifted waveform ($t_0 = 0$) has no phase change (the slope of the line goes to zero).

## 1.5.1 Sine vs. Cosine

The difference between a sine and cosine waveform is a phase shift of $\pi/2$ rad:

$$\sin(2\pi f_0 t) = \cos(2\pi f_0 t - \pi/2)$$

Consequently, the general sinusoid expressed as a cosine with a phase shift incorporates both sine and cosine signals. Furthermore, a one-sided cosine graph may be used for either sines or cosines, if the appropriate phase differences are recorded on an accompanying phase vs. frequency graph. Since phase cannot be illustrated using the one-sided amplitude vs. frequency graph alone, there is no way to distinguish sine from cosine directly on this graph; a phase change or time shift in the cosine signal would not be apparent. On the other hand, a sine cannot be distinguished from a cosine either by eye, when viewed on an oscilloscope, or by ear, when played through headphones. In fact, the difference between a sine and cosine at the same frequency only becomes apparent when one waveform can observed relative to another, or relative to some known origin.

The effect of varying the parameters that define cosine signals can be studied by combining a constant factor into the defining equation. Table 1.1 shows several manipulations of $s(t) = A\cos(2\pi f_0 t)$, using only the constant value 2 in various ways. All these manipulations would affect the time domain graph of the waveform. Each would have a corresponding effect on the amplitude vs. frequency and phase representations of the signal.

## 1.5.2 Combining Signals: The Gated Sine Wave

Linear combinations of familiar signals will define new and different signals that can be analyzed by reference to their known components. This exercise uses a rectangle to turn a sine wave on for exactly 3 seconds beginning at time $t = 0$ as shown in Figure 1.29. The new *gated sine* will have properties related to both the rectangle and the sine waveform.

A rectangle 3 seconds wide and centered at 1.5 seconds is derived by shifting and stretching the unit rectangle. The expression $\text{rect}(t/3)$ gives a rectangle 3 s wide,

**TABLE 1.1**

**Mathematical Manipulation of $A\cos(2\pi f_0 t)$**

| Manipulation | Effect |
|---|---|
| $A\cos(2\pi f_0 t)$ | original signal |
| $2\,A\cos(2\pi f_0 t)$ | doubled in amplitude |
| $2 + A\cos(2\pi f_0 t)$ | added a constant (DC offset) |
| $A\cos(4\pi f_0 t)$ | doubled in freq. (compressed in time) |
| $A\cos(2\pi f_0 (t+2))$ | shifted left by 2 s (= $4\pi f_0$ rad) |
| $A\cos(2\pi f_0 (t-2))$ | shifted right by 2 s (= $4\pi f_0$ rad) |
| $A\cos(2\pi f_0 t - 2)$ | shifted right by 2 rad (= $1/\pi f_0$ s) |

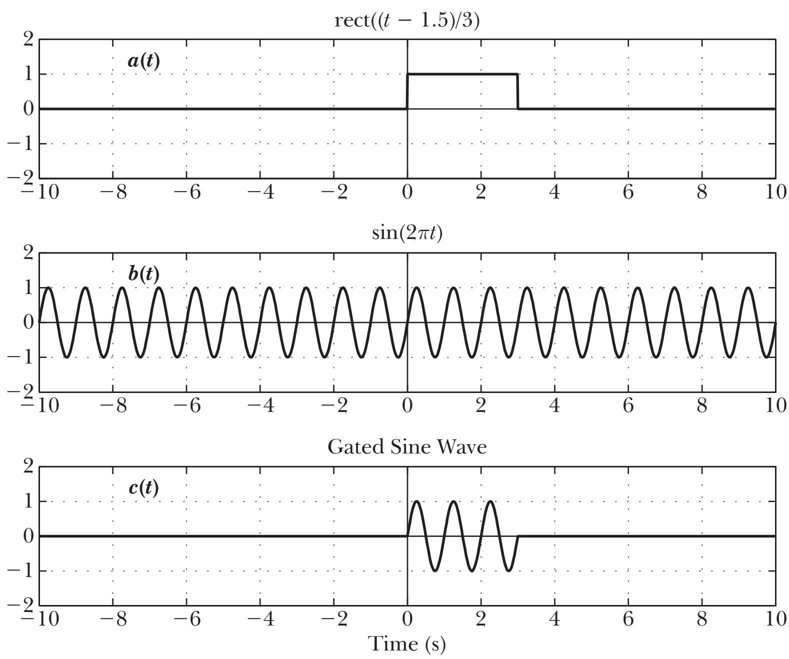



**FIGURE 1.29** Gated Sine Wave—The on/off sine pulse $c(t)$ may be formed as the product of a shifted, scaled unit rectangle $a(t)$ and a sine $b(t)$.

while $\text{rect}(t - 1.5)$ is a rectangle shifted left 1.5 s. Combined, the desired rectangle is given by $\text{rect}((t - 1.5)/3)$ and *not* $\text{rect}(t/3 - 1.5)$.

```
t = -10:0.01:10;                          % define a time axis
rect = double(rectpuls((1/3)*(t-1.5)));   % shifted rectangle
sine = sin(2*pi*t);                       % T=1 s, A=1
gsine = rect .* sine;                     % create gated sine wave
plot(t,gsine);                            % plot gated sine wave
axis([-10 10 -2 2 ]);                     % zoom-in to show detail
title('Gated Sine Wave');                 % add a title
xlabel('Time (s)');                       % label horizontal axis
```

Conversely, if the above 3-second-long sinusoidal burst was being encountered for the first time, it would be valuable to recognize that it may be considered as the combination of two well-known signals. In general, expressing a signal by components greatly simplifies the analysis task.

## 1.5.3 Combining Signals: A Dial Tone Generator

The *dial tone* heard on a telephone in North America is composed of two cosines added together, one at 350 Hz and the other at 440 Hz. This operation is shown in Figure 1.30 and the resulting signal could be written as:

$$\text{dialtone}(t) = \cos(700\pi t) + \cos(880\pi t)$$

Figure 1.31 shows the two component cosines and the resulting dial tone waveform. Unlike the gated sinewave example above, the identity of the (two) components

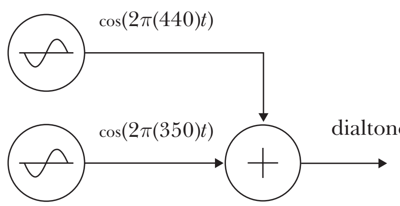



**FIGURE 1.30** Telephone Dial Tone Generator—The familiar sound of a dial tone is the sum of cosines at 350 Hz and 440 Hz.

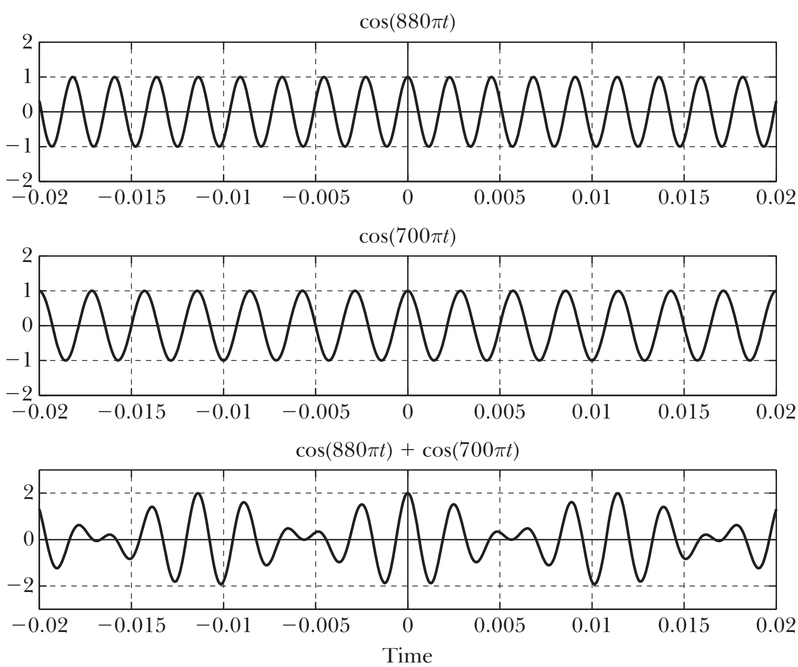



**FIGURE 1.31** Dial tone—The dial tone sound is formed as the sum of two cosines.

that add to give this dial tone are not immediately obvious from its appearance. Identifying the unknown components of an arbitrary signal is an important way of simplifying its study.

# 1.6 Useful Hints and Help with MATLAB

This section summarizes some MATLAB commands that are useful in creating neat and meaningful signal plots. More information on any command or function in MATLAB can be found by typing `help name` with the name of the command or function.

## 1.6.1 Annotating Graphs

The simplest graph of a MATLAB array $s$ is obtained using `plot(s)`, which creates a graph automatically scaled in the $y$-axis to fully fit the values in $s$, and the horizontal axis will be labelled with the full set of indices spanned by $s$. The option `plot(t,s)` uses values from $t$ to fill the $x$-axis; both $s$ and $t$ must have the same number of elements. Some of the most-used functions to enhance a plot are listed below.

While `plot(s)` connects points with a line, another useful graph type with similar options is `stem(s)`, which connects a vertical line to each point from the $x$-axis. By default, a small circle is also drawn on each point. The option `stem(s,marker,'none')` omits the circles.

1. **Plot Window** By default, the first plot window will be named *Figure 1* and further calls to the `plot()` function will replace the current graph with a new plot.
   - **a.** `figure(N);` Use figure $N$ for future calls to plot.
   - **b.** `clf;` Clear the current figure.
   - **c.** `subplot(A,B,C);` Use plot $C$ from among $A \times B$ subplots.
   - **d.** `hold on;` Further calls to `plot()` will *not* delete the current plot.
2. **Axis Labels**
   - **a.** `title('text');` Adds a text title to a plot.
   - **b.** `xlabel('text');` Adds a label to the $x$-axis.
   - **c.** `ylabel('text');` Adds a label to the $y$-axis.
3. **Plot Markup**
   - **a.** `text(x,y,'text');` Inserts text within a plot at $(x, y)$.
   - **b.** `line([x1,y1],[x2,y2]);` Draws a line within a plot.
4. **Plot Adjustment**
   - **a.** `axis([x1, x2, y1, y2]);` Set the horizontal and vertical limits.
   - **b.** `set(gca,'XTick',0:15);` Set $x$-axis tick marks 0:15.
   - **c.** `set(gca,'XTickLabel',0:15);` Label tick marks as above.
   - **d.** `grid on;` Turn on plot grid lines.

# 1.7 Conclusions

Engineers benefit from modeling physical systems and representing signals as mathematical functions of time and frequency. Many signals are most familiar as time domain functions, and many other signals can be constructed by combining familiar signal types. In this chapter, some basic signal types have been introduced as well as the mathematical manipulations that may be used to express various signals in terms of others. These concepts will be expanded upon in later chapters as more sophisticated mathematical tools are introduced and explored.

# End-of-Chapter Exercises

**1.1** Sketch the signal $a(t)$ defined below. Determine graphically the value $a(0)$ and the area under $a(t)$. Confirm your results with direct calculation.

$$a(t) = \begin{cases} 2 - 2t & \text{if } |t| < 1 \\ 0 & \text{otherwise} \end{cases}$$

**1.2** Sketch the signal $a(t)$ defined below. Determine graphically the value $a(0)$ and the area under $a(t)$. Confirm your results with direct calculation.

$$a(t) = \begin{cases} 2 - t & \text{if } |t| < 1 \\ 1 & \text{otherwise} \end{cases}$$

**1.3** Sketch the signal $a(t) = 2\,\text{rect}(2(t-2))$. Determine graphically the value $a(0)$ and the area under $a(t)$. Confirm your results with direct calculation.

**1.4** Sketch the signal $a(t) = 2\delta(t-3)$. Determine graphically the area under $a(t)$. What is the value of $a(0)$? What is the value of $a(3)$?

**1.5** The signal $s(t) = \text{rect}(t/2)$ is delayed by 2 seconds. State whether a sketch of the signal is shifted to the right or to the left along the time axis. Give an expression for the new signal in terms of $s(t)$.

**1.6** Carefully sketch each of the signals in Table 1.1 as both amplitude vs. time, and as amplitude vs. frequency, using $A = 10$, and $f_0 = 1/2\pi$.

**1.7** Refer to Figure 1.32 and express the signal $b(t)$ in terms of the signal $a(t)$ in two different ways.

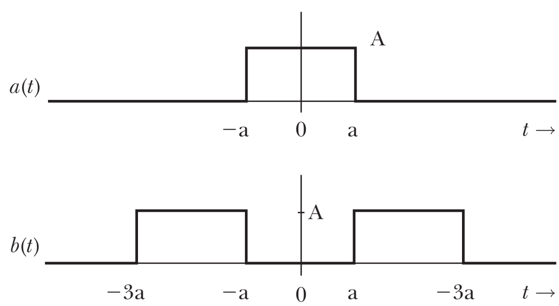



**FIGURE 1.32** Diagram for Question 1.7.

**1.8** Refer to Figure 1.33 and express this signal in terms of the unit rectangle.

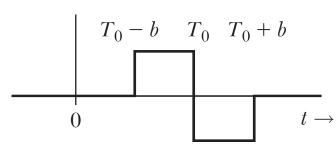



**FIGURE 1.33** Diagram for Question 1.8.

**1.9** Variations on the rectangle can be related to the unit rectangle rect$(t)$. In each of the sketches in Figure 1.34, express the signal $s(t)$ shown in terms of rect$(t)$.

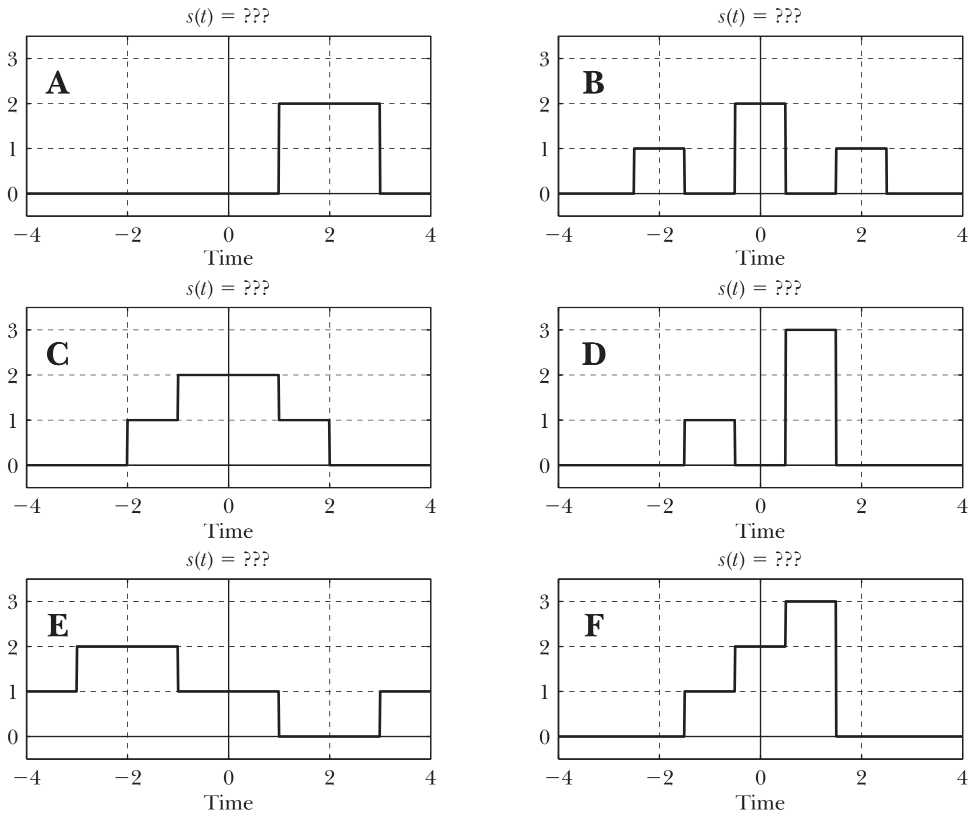



**FIGURE 1.34** Diagram for Question 1.9.

**1.10** Show how the unit rectangle rect$(t - 10)$ may be described in terms of the unit step $u(t)$.

**1.11** Consider the following signals $a(t) \cdots e(t)$:

**(a)** $a(t) = 2\sin(\pi t)$
**(b)** $b(t) = 3\cos(2t)$
**(c)** $c(t) = 10\sin(2\pi t + \pi/4)$
**(d)** $d(t) = \cos(\pi t + \pi/2)$
**(e)** $e(t) = 5\sin(\pi(t - 1))$

Write each of the signals $a(t) \cdots e(t)$ in terms of $A\cos(2\pi f_0 t + \Phi)$ by giving values for $(A, f_0, \Phi)$ in each case.

**1.12** Accurately sketch by hand each of the signals in Question 1.11 $a(t) \cdots e(t)$ and find the value at $t = 0$ in each case.

**1.13** Use MATLAB to sketch each of the signals in Question 1.11 $a(t) \cdots e(t)$ and find the value at $t = 0$ in each case.

**1.14** Evaluate the following integral by inspection.

$$\int_{-\infty}^{+\infty} \delta(t - \pi)\cos(2t - \pi)\,dt$$

**1.15** A voltage signal $v(t)$ travels through an amplifier and emerges delayed by 0.001 s and increased to ten times its original amplitude but otherwise undistorted. Write an expression for the new signal in terms of $v(t)$.

**1.16** A voltage signal $c(t)$ is a cosine with amplitude 5 V and frequency 2 kHz. The signal travels over a communications network and emerges delayed by 100 $\mu$s and reduced to one third its original amplitude but otherwise undistorted. Write an expression for the new signal in terms of cosine. How much phase shift (in radians) has occurred? Accurately sketch the output signal and show the phase change.

**1.17** Consider the sinusoid in Figure 1.35.

**(a)** Determine the terms $(A, f_0, \Phi)$ and write the signal as a general cosine.
**(b)** Show the above terms on a one-sided cosine graph and phase graph.

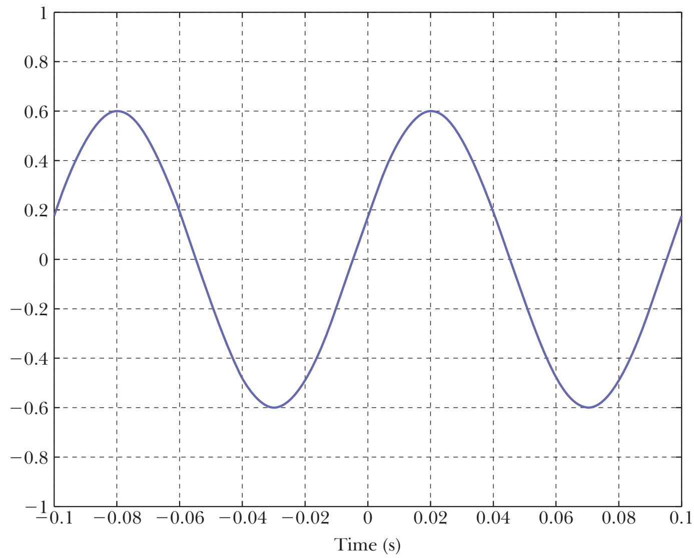



**FIGURE 1.35** Diagram for Question 1.17.

**1.18** What is the period of the dialtone(t) waveform from Figure 1.31? How does the period relate to the periods of its component signals?

**1.19** A signal $s(t)$ is required that consists of exactly 15 complete periods of a 10 Hz sine wave centered on the origin.

**(a)** Accurately sketch the appearance of $s(t)$.

**(b)** Give an exact expression for the signal $s(t)$ as the product of a unit rectangle and a sine function.

**1.20** Consider the four signals $(a, b, c, d)$ shown in Figure 1.36.

**(a)** Give a simplified expression for $c(t)$ in terms of $a(t)$ and $b(t)$.

**(b)** Give a simplified expression for $d(t)$ in terms of $a(t)$ and $b(t)$.

**(c)** Give a simplified expression for each of the signals $(a, b, c, d)$ in terms of $\text{rect}(t)$ and/or the generic cosine $A\cos(2\pi f_0 t + \Phi)$.

**1.21** Consider the signal $s(t) = a(t) - b(t) + c(t)$, where $\{a(t), b(t), c(t)\}$ are:

- $a(t) = \cos(2\pi t)$
- $b(t) = \frac{1}{3}\cos(6\pi t)$
- $c(t) = \frac{1}{5}\cos(10\pi t)$

**(a)** Identify the frequency of each of the component cosines $\{a(t), b(t), c(t)\}$.

**(b)** Sketch a system block diagram that shows how the three signals combine to make $s(t)$.

**(c)** Sketch the signal $s(t)$ on a one-sided cosine graph.

**(d)** Use MATLAB to plot the signal $s(t)$ on an interval from $t = -1$ to $t = +1$ s.

**1.22** Let $s(t) = a(t) - b(t) + c(t)$, where $\{a(t), b(t), c(t)\}$ are shown below and $\Phi_n = 0$. This is the same signal $s(t)$ from Question 1.21.

- $a(t) = \cos(2\pi t + \Phi_a)$
- $b(t) = \frac{1}{3}\cos(6\pi t + \Phi_b)$
- $c(t) = \frac{1}{5}\cos(10\pi t + \Phi_c)$

Consider the time-shifted signal $s(t - 0.1)$

**(a)** Write non-zero values for $\{\Phi_n\}$ such that $s(t - 0.1) = a(t) - b(t) + c(t)$. Each of these components must be shifted by 0.1 s.

**(b)** Sketch the signal $s(t - 0.1)$ on a one-sided cosine graph and phase graph.

**(c)** What is the slope of the line described by the three values $\{\Phi_a, \Phi_b, \Phi_c\}$ on the phase graph?

**(d)** Give an expression for the slope of the line for a general time delay $s(t - t_0)$ for $t_0 > 0$.

**(e)** Use MATLAB to plot the signal $s(t - 0.1)$ on an interval from $t = -1$ to $t = +1$ s.

**1.23** A system is designed for which an input signal emerges delayed by 1.0 seconds, followed by (an echo) another copy of the same signal 1.0 second after that and another copy of the same signal 1.0 second after that. Let the input signal be $a(t) = \text{rect}(t/4)$. Let the output signal be $b(t)$. Sketch the output signal and give an expression for $b(t)$ in terms of $a(t)$.

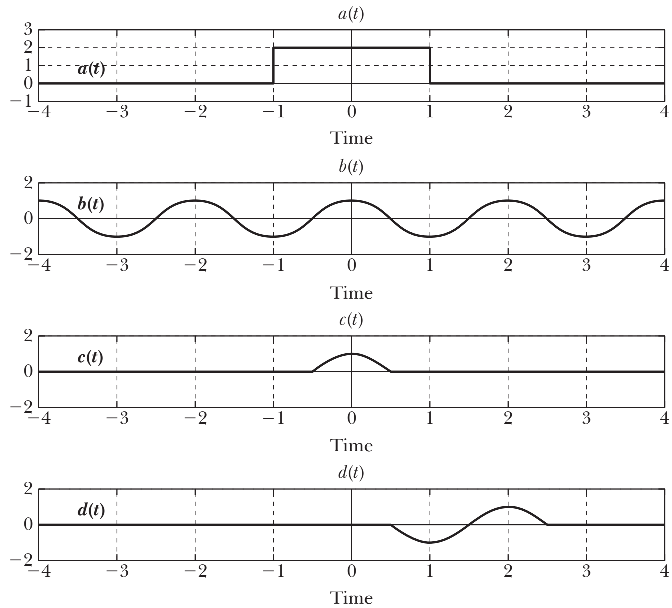



**FIGURE 1.36** Diagram for Question 1.20.

**1.24** A system is designed for which an input signal emerges delayed by 1.0 seconds, followed by (an echo) another copy of the same signal 1.0 second later. An input signal $a(t) = \text{rect}(t)$ arrives and 5.0 seconds later another input signal $b(t) = \text{rect}(t)$ arrives. Let the output signal be $c(t)$. Sketch the output signal and give an expression for $c(t)$ in terms of $a(t)$ and $b(t)$.

**1.25** A system is designed for which an input signal emerges delayed by 1.0 seconds, followed by (an echo) another copy of the same signal 1.0 second later. An input signal $a(t) = \text{rect}(t)$ arrives and 0.5 seconds later another input signal $b(t) = \text{rect}(t/2)$ arrives. Let the output signal be $c(t)$. Sketch the output signal and give an expression for $c(t)$ in terms of $a(t)$ and $b(t)$.

CHAPTER 2

# Classification of Signals

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- List the different signal types in the time domain
- Recognize odd and even signals and their properties
- Identify common periodic signals by their appearance
- Explain the difference between power and energy signals
- Calculate the energy and power and $V_{rms}$ of signals
- Describe new signals graphically in terms of common signals
- Use MATLAB to define and to plot common signals
- Compute the phase and magnitude of a complex signal
- Explain the properties defining a random signal
- Define a discrete time signal from a continuous time signal

A signal $s(t)$ may be classified by identifying certain characteristics that serve to predict its behavior or appearance, to verify computational results and in many cases, to greatly simplify the mathematics involved in signal analysis. This chapter defines several important categories of signals, each of which describes some significant characteristic, which, in turn, may dictate the mathematical approach one might take in signals analysis. Related computations and practical MATLAB examples are presented for each signal type.

## 2.1 Introduction

The list of categories shown below summarizes the kind of characteristics that could be identified for a signal $s(t)$. These classifications are not all mutually exclusive, and it is expected that every signal $s(t)$ can be classified according to each of these definitions. For example, the signal $s(t) = \cos(t)$ can be described as being periodic, even, power, real, continuous, and deterministic.

- Periodic or Non-periodic
- Odd or Even or Neither
- Power or Energy
- Real or Complex
- Continuous Time or Discrete Time
- Random or Deterministic

# 2.2 Periodic Signals

Sines and cosines are examples of periodic signals, repeating every $T$ seconds where the constant $T = 1/f_0$ is called the *period*. By definition, a signal $s(t)$ is periodic with period $T$ if the same value of $s(t)$ can always be found $T$ seconds away at $s(t + T)$, or:

**DEFINITION 2.1 Periodic Signal**

*Given a signal* s($t$) *if, for all* t *and for constant* T,

$$s(t + T) = s(t)$$

*then* s(t) *is a* periodic *signal, and* T *is its* period.

This property is especially useful in signals analysis when it is observed that computations performed over a single period of length $T$ seconds would simply be duplicated for all other periods of the signal.

A variety of useful and familiar periodic signals are described in this section. In each case, a short example illustrates how the periodic signal may be generated and displayed using MATLAB. While the most common periodic signals are available directly in MATLAB, any periodic waveform can be created by using the `pulstran()` function to duplicate a given signal shape at regular intervals along the time axis. In later chapters, it is shown that *all* periodic signals can be expressed as a linear combination of sine and cosine signal components.

## 2.2.1 Sinusoid

As discussed in Section 1.4, the sinusoid signal includes both sines and cosines.

MATLAB includes the functions `cos()` and `sin()` and both expect arguments in radians rather than degrees, as can be easily shown with an example.

```
cos(pi / 4)
ans =
  0.7071
```

The $s(t) = \cos(t)$ has a period of $2\ \pi$ s, as can be seen in Figure 2.1, and generated in MATLAB as:

```
t = -10 : 0.01 : 10;
a = cos(t);
plot(t, a)
axis([-10 10 -2 2 ])
grid on
```

## 2.2.2 Half-Wave Rectified Sinusoid

The *half-wave rectified sinusoid* is the result of passing a sinusoid through an ideal diode as shown in Figure 2.2 to give the waveform of Figure 2.3.

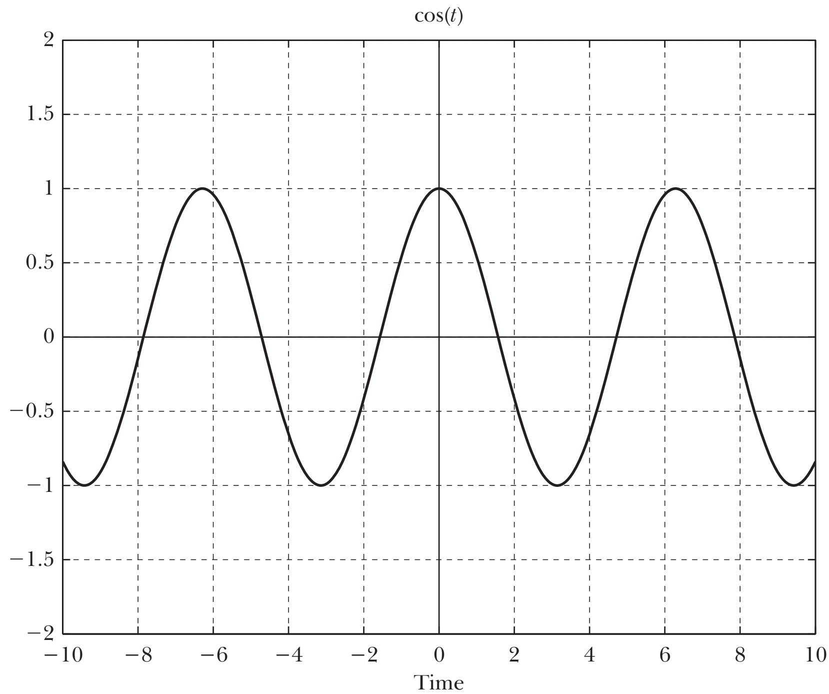



**FIGURE 2.1** Cosine—Period: $T = 2\pi$ s.

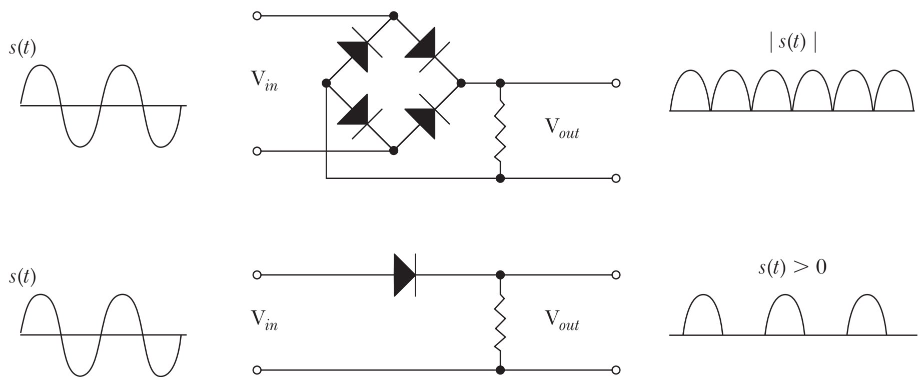



**FIGURE 2.2** Ideal diode circuits accomplish a *full-wave rectifier* (top) and *half-wave rectifier* (bottom). The effect of these circuits can be modelled mathematically as shown.

The electronic *half-wave rectifier* can be accomplished mathematically by passing only the positive part of a waveform. In MATLAB, when a conditional statement such as `(a > 0)` is used within an expression, it returns 1 or 0 for each corresponding element of the vector `a`, as in the code below:

```
t = -10 : 0.01 : 10;
a = cos(t);
b = a .* (a > 0); % half-wave rectifier
plot(t, b)
axis([-10 10 -2 2 ])
grid on
```

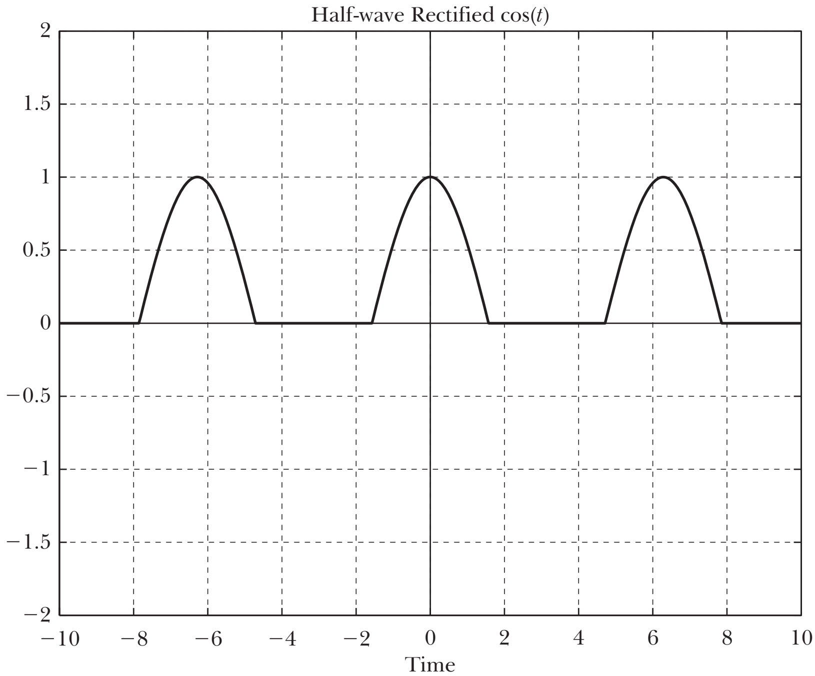



**FIGURE 2.3** Half-wave rectified cosine—Period: $T = 2\pi$ s.

## 2.2.3 Full-Wave Rectified Sinusoid

The *full-wave rectified sinusoid* is the result of passing a sinusoid through an ideal diode bridge as shown in Figure 2.2 to give the waveform of Figure 2.4.

The electronic *full-wave rectifier* is mathematically equivalent to the absolute value function or `abs(a)` in MATLAB. Note that the period of the resulting signal is half that of the original sinusoid, and this circuit configuration is sometimes employed as the basis for a *frequency doubler* application.

```
t = -10 : 0.01 : 10;
a = cos(t);
b = abs(a); % full-wave rectifier
plot(t, b)
axis([-10 10 -2 2 ])
grid on
```

## 2.2.4 Square Wave

The *square wave* is defined as a *periodic waveform that alternately for equal lengths of time assumes one of two fixed values* [2]. For example, a square wave may have alternating positive and negative values of equal magnitude and duration. The square wave is a special case of the rectangular pulse train shown in Figure 2.7 and can also be described as periodic rectangles in which each rectangle is one half period in length (50% Duty Cycle). The square wave is a built-in MATLAB function called `square(t)`. Like `cos(t)`, the default period of the square wave is $T = 2\pi$ s, and its range is $(-1,+1)$, as seen in Figure 2.5.

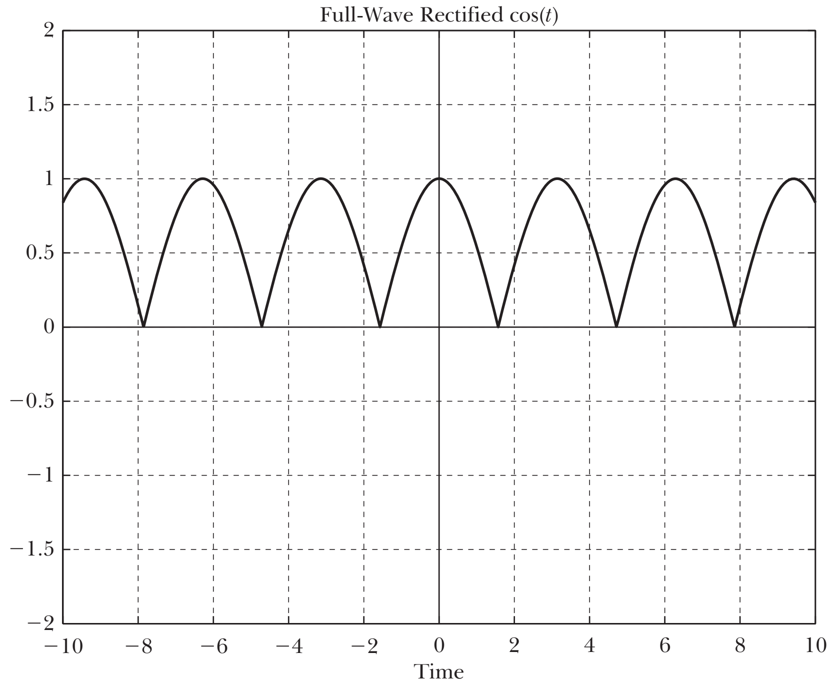



**FIGURE 2.4** Full-wave rectified cosine—Period: $T = \pi$ s.

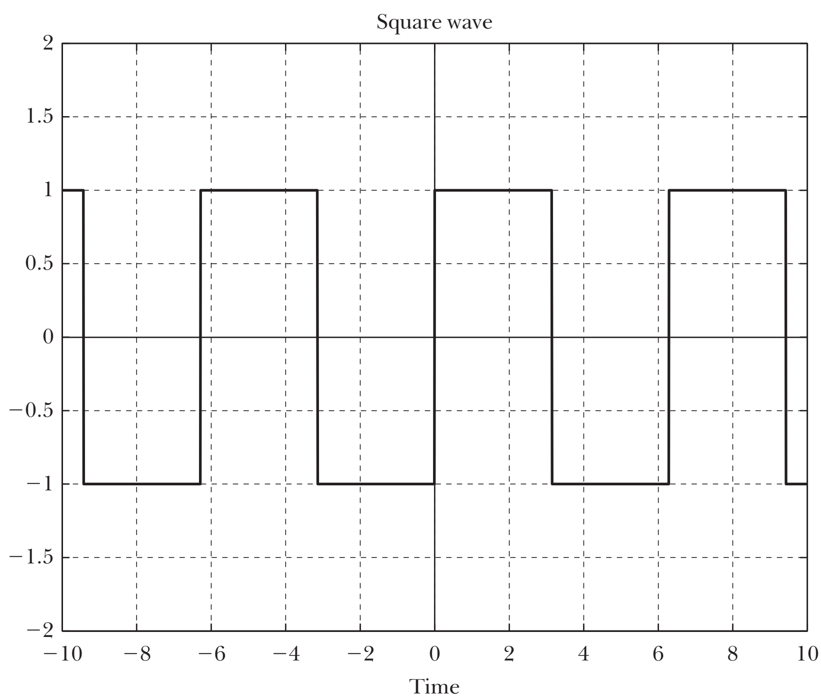



**FIGURE 2.5** Square wave—Period: $T = 2\pi$ s.

```
t = -10 : 0.01 : 10;
a = square(t);
plot(t, a)
axis([-10 10 -2 2 ])
grid on
```

## 2.2.5 Sawtooth Wave

The *sawtooth* wave is defined by a *ramp and return to initial value* [2], typically a linear rising slope followed by a discontinuous drop. This is also known as a *ramp* signal. The sawtooth wave is a built-in MATLAB function called `sawtooth(t)`. By default, its period is $T = 2\pi$ s, and its range is $(-1, +1)$, as seen in Figure 2.6.

```
t = -10 : 0.01 : 10;
a = sawtooth(t);
plot(t, a)
axis([-10 10 -2 2 ])
grid on
```

## 2.2.6 Pulse Train

The *pulse train* is defined generally as a *continuous repetitive series of pulse waveforms* [2]. The shape of the pulse within each period further characterizes the pulse train. Pulse waveforms of particular interest include pulses that are rectangular or triangular in shape.

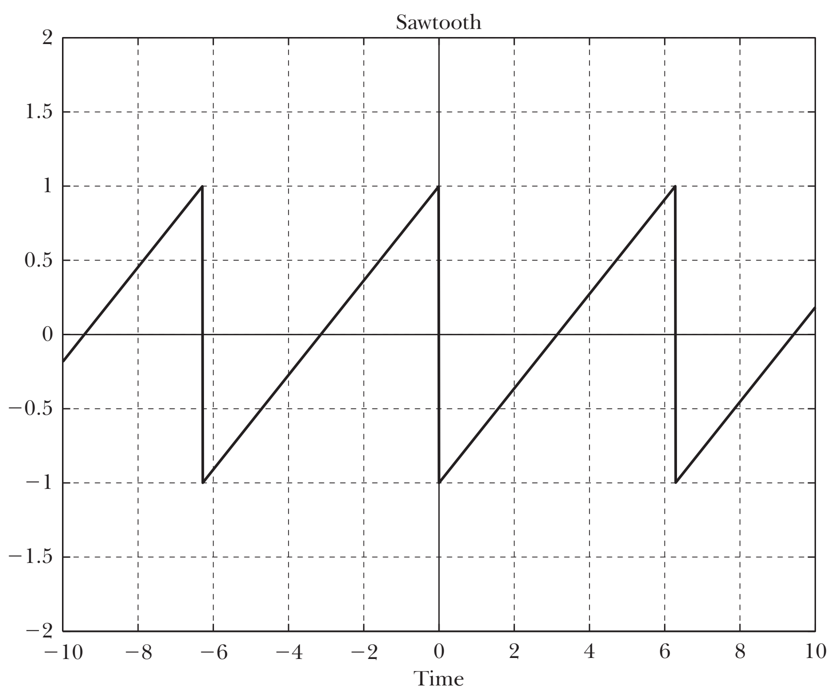



**FIGURE 2.6** Sawtooth Wave—Period: $T = 2\pi$ s.

### 2.2.7 Rectangular Wave

The *rectangular wave* or *rectangular pulse train* is characterized by its *duty cycle*, expressed in percent, indicating what fraction of the each period is occupied by the rectangular pulse. In particular, a rectangular pulse train with a 50 percent duty cycle has the square wave appearance of Figure 2.5.

The generalized pulse train is a built-in MATLAB function that includes a parameter to specify the pulse shape. A rectangular pulse train may be created by repeating a series of rectangular pulses defined by `rectpuls()`. In this case, the signal range is $(0, +1)$, as seen in Figure 2.7. A rectangular wave with a 25 percent duty cycle can also be generated directly using `square(t,25)`.

```
t = -10 : 0.01 : 10;
w = pi/2;         % define rectangle width
p = 2*pi;         % define rectangle spacing
d = -10 : p : 10; % define rectangle positions
s = pulstran(t, d, 'rectpuls', w);
plot(t,s)
axis([-10 10 -2 2 ])
grid on
```

In Figure 2.7, the pulse train has period $T = 2\pi$ s and a 25 percent duty cycle given by:

$$\text{duty cycle} = \frac{\text{pulse width}}{\text{period}} = \frac{\pi/2}{2\pi} = 0.25$$

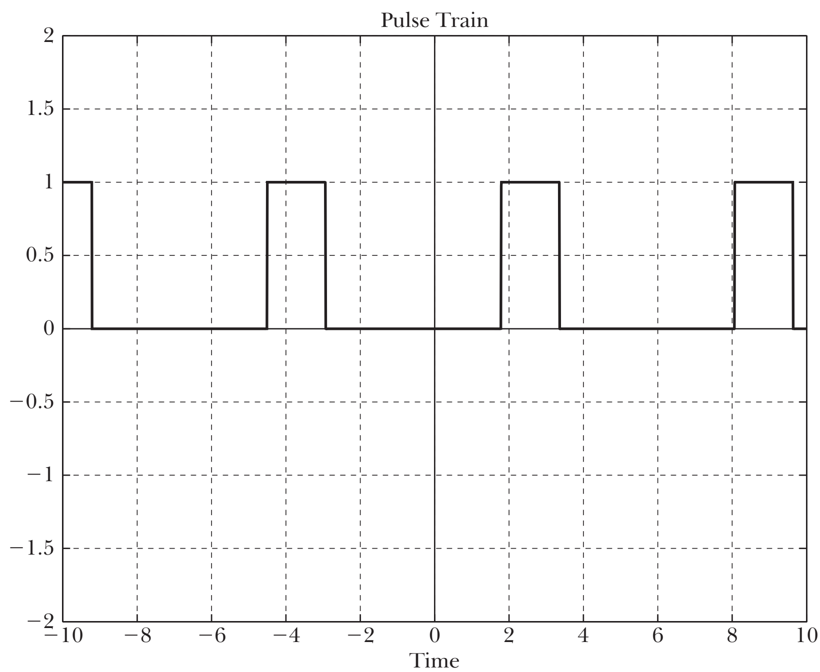



**FIGURE 2.7** Pulse Train—A regular series of rectangular pulses. Duty Cycle: 25%, Period: $T = 2\pi$ s.

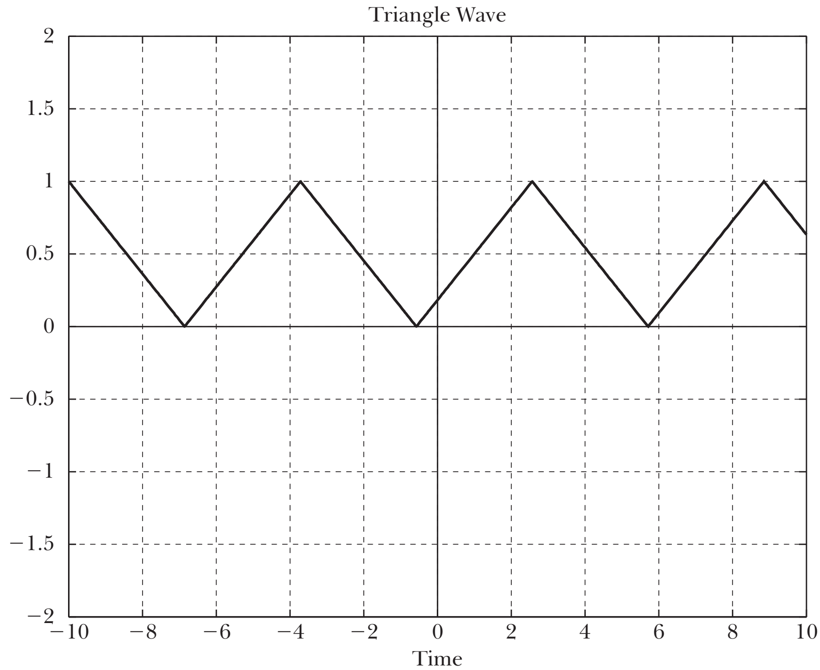



**FIGURE 2.8** Triangle Wave—A regular series of triangular pulses. Period: $T = 2\pi$ s.

## 2.2.8 Triangle Wave

The *triangle wave* or *triangular pulse train* describes a periodic series of triangles. The triangle wave may be created by repeating a series of triangular pulses using the `tripuls()` function with the pulse train `pulstran()` function. In this case, the signal range is $(0, +1)$, as seen in Figure 2.8. A triangle wave can also be generated directly using `sawtooth(t,0.5)`.

```
t = -10 : 0.01 : 10;
w = 2*pi;         % define triangle width
d = -10 : w : 10; % define triangle positions
s = pulstran(t, d, 'tripuls', w);
plot(t,s)
axis([-10 10 -2 2 ])
grid on
```

## 2.2.9 Impulse Train

The periodic signal consisting of an infinite series of unit impulses spaced at regular intervals $T$ is called variously an *infinite impulse train*, *comb function*, or *picket fence*, and may be defined as:[1]

$$\text{comb}(t, T) = \sum_{n=-\infty}^{+\infty} \delta(t - nT)$$

[1] This notation for $\text{comb}(t, T)$ is not a standard form.

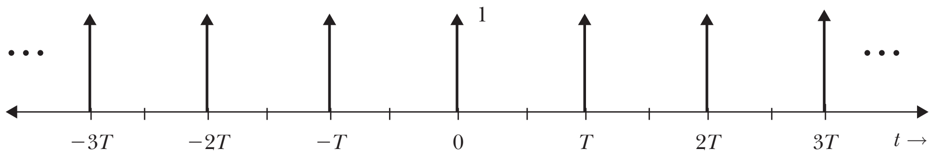



**FIGURE 2.9** This *impulse train* is an infinite series of impulse functions spaced at regular intervals $T$ along the time axis.

Each term $\delta(t - nT)$ is a unit impulse shifted to position $t = nT$, as shown in Figure 2.9.

The impulse train forms the basis for many periodic signals, including all the other periodic signals shown in this section. Beginning with the impulse train, if any shape such as a triangle is placed at every impulse, then arbitrary periodic functions can be assembled.

**DC Component in Periodic Signals** Periodic signals are distinguished by their shape and characterized by the appearance of a single period. Without changing either the period or the overall appearance, a periodic signal could be moved up and down with respect to the horizontal axis by the addition or subtraction of a constant value or *DC offset*. Similarly, shifting the waveform in time or changing its amplitude will have no effect on its general appearance when viewed on an oscilloscope.

## 2.3 Odd and Even Signals

The operation $s(t) \rightarrow s(-t)$ acts to reverse a signal in time, causing $s(t)$ to be reflected about the origin $t = 0$. The behavior or appearance of a signal under this reflection serve to categorize $s(t)$ as *odd*, or *even*. Signals with no such special reflection properties will be called *neither* odd nor even.

A signal such as $\cos(t)$ possesses the property of being unchanged when reflected about the origin $t = 0$; in other words, $\cos(t) = \cos(-t)$. The observation defines an *even* signal.

**DEFINITION 2.2 Even Signal**

*Given a signal* s(t) *if, for all* t,

$$s(-t) = s(t)$$

*then* s(t) *is called an* even *signal, or a* symmetrical *signal.*

Similarly, for every point in the signal $\sin(-t) = -\sin(t)$; the signal $\sin(t)$ is anti-symmetrical when reflected about the $t = 0$ axis. This observation defines an *odd* signal.

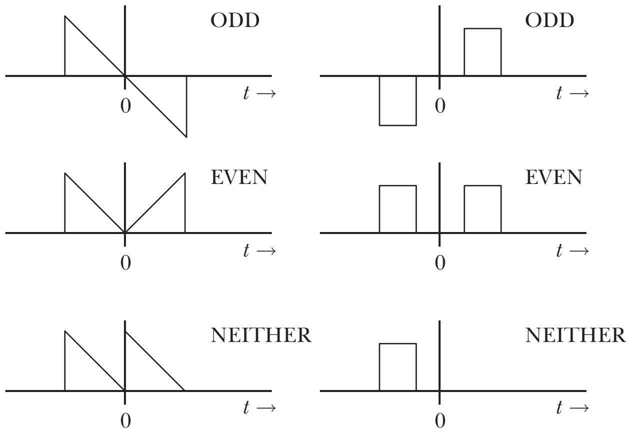



**FIGURE 2.10** Odd or even signals are defined by reflection about the origin.

**DEFINITION 2.3 Odd Signal**

*Given a signal* s(t) *if, for all* t,

$$s(-t) = -s(t)$$

*then* s(t) *is called an* odd *signal, or an* anti-symmetrical *signal.*

Many signals are neither *odd* nor *even*, and the definition of a *neither* signal is made here for completeness and in the interest of brevity.

**DEFINITION 2.4 Neither Signal**

*A signal* s(t) *that is neither* odd *nor* even *shall be called a* neither *signal.*

Examples of odd and even signals are shown graphically in Figure 2.10. Many signals such as rect$(t - 2)$ or the unit step function in Figure 1.12 are neither signals, but *odd-ness* or *even-ness* is a fundamental property that can be employed to simplify computations. In particular, much integration can be saved by observing the following theorems that may be proved by inspection:

**THEOREM 2.1** *(Integration of even signals)*

If $s(t)$ is even, then for constant *k:*

$$\int_{-k}^{+k} s(t)\, dt = 2\int_{0}^{k} s(t)\, dt$$

**THEOREM 2.2** *(Integration of odd signals)*

If $s(t)$ is odd, then for constant $k$:

$$\int_{-k}^{+k} s(t)\, dt = 0$$

## 2.3.1 Combining Odd and Even Signals

When two signals that are odd or even are multiplied or added together, the result is another function, which may also be odd or even.

Let $a_N(t)$ be an even function, and let $b_N(t)$ be an odd function. Then, by definition, $a_N(-t) = +a_N(t)$, and $b_N(-t) = -b_N(t)$. By observing the appropriate sign changes, Tables 2.1 and 2.2 show how odd and even signals combine under multiplication and addition.

Table 2.2 shows that the result of adding an odd function to an even function is a neither function. In particular, this is the case for adding a constant (even) to some odd function. This observation leads to two important conclusions.

1. No even function could ever be created by the addition (linear combination) of functions that includes any odd functions. The only linear combinations of functions that can result in an even function must consist exclusively of even functions.
2. No odd function could ever be created by the addition (linear combination) of functions that includes any even functions. The only linear combinations of functions that can result in an odd function must consist exclusively of odd functions.

**TABLE 2.1**

**Multiplication of Odd and Even Signals**

| | |
|---|---|
| $a_1(-t) \times a_2(-t) = +a_1(t) \times +a_2(t)$ | *even* $\times$ *even* = *even* |
| $a_1(-t) \times b_2(-t) = +a_1(t) \times -b_2(t)$ | *even* $\times$ *odd* = *odd* |
| $b_1(-t) \times b_2(-t) = -b_1(t) \times -b_2(t)$ | *odd* $\times$ *odd* = *even* |

**TABLE 2.2**

**Addition of Odd and Even Signals**

| | |
|---|---|
| $a_1(-t) + a_2(-t) = +a_1(t) + +a_2(t)$ | *even* + *even* = *even* |
| $a_1(-t) + b_2(-t) = +a_1(t) + -b_2(t)$ | *even* + *odd* = *neither* |
| $b_1(-t) + b_2(-t) = -b_1(t) + -b_2(t)$ | *odd* + *odd* = *odd* |

These properties of odd and even functions under addition illustrate the concept of *orthogonality*. Any odd function is *orthogonal* to any even function over a symmetric interval $[-t_0, +t_0]$ and vice versa, because no linear combination of odd functions will ever lead to an even function. The property of orthogonality is key to expressing signals in terms of a unique set of component signals, as will be discussed in Chapter 3.

### EXAMPLE 2.1 (Odd and Even Signals)

*Evaluate the integral*

$$c(t) = \int_{-\pi}^{+\pi} \cos(10\pi t)\, \sin(13\pi t)\; dt$$

**Solution:**
Since the product of even and odd functions is odd, then by inspection and from Theorem 2.2, $c(t) = 0$. This example illustrates the value of recognizing the properties of odd and even signals.

### EXAMPLE 2.2 (The Sum of Two Sinusoids)

*Consider the signal*

$$g(t) = \cos(2t) + \cos(3t)$$

*Determine if* g(t) *is odd or even and find its period.*

**Solution:**
The sum of two even signals will give an even result. This is confirmed in Figure 2.11.

| SIGNAL | Amplitude | Phase | Frequency | Period |
|---|---|---|---|---|
| $\cos(2t)$ | 1 | 0 | $1/\pi$ | $\pi$ |
| $\cos(3t)$ | 1 | 0 | $3/2\pi$ | $2\pi/3$ |

By inspection, the two cosines have periods $\pi$ and $2\pi/3$ s respectively. These periods would coincide with each other every $2\pi$ s, after two periods of the first and three periods of the second waveform. The overall period of the resulting signal $g(t)$ would be given directly as the Least Common Multiple (LCM) of the two non-zero cosine periods, as $LCM(\pi, 2\pi/3) = 2\pi$ s. In MATLAB, the function `lcm(a,b)` returns the LCM of the positive integers $a$ and $b$. While the signal $g(t)$ is periodic, its shape only vaguely resembles either of the component signals, as shown in Figure 2.11.

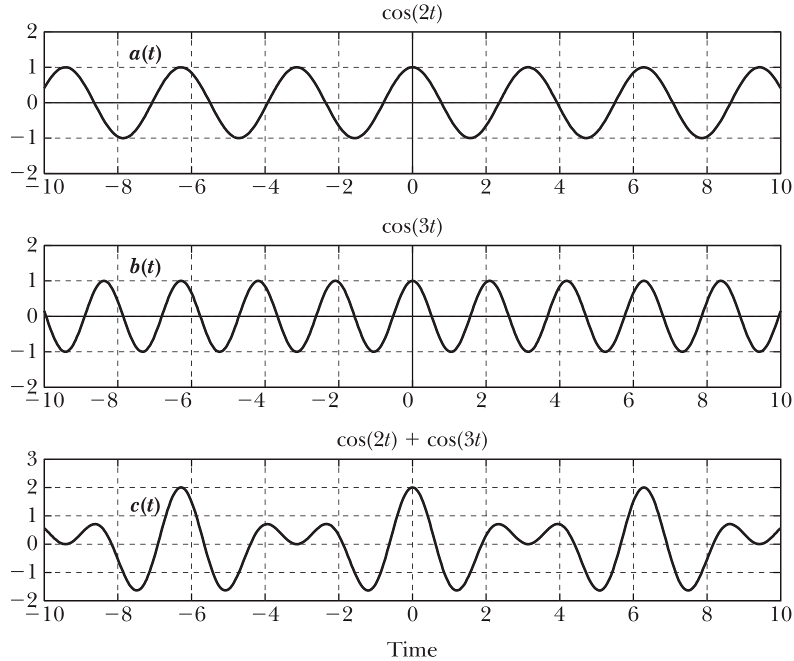



**FIGURE 2.11** The Sum of Two Cosines $c(t) = a(t) + b(t)$—The overall period is the LCM of the two periods—Period: $T = 2\pi$ s.

## 2.3.2 The Constant Value $s(t) = A$

The constant value $s(t) = A$ is an even signal. Adding a constant value (or *DC-offset*) to another signal does not affect the shape or appearance of the waveform; it simply moves the signal up or down.

An even signal will remain an even signal if a constant is added to it (even + even = even), while in general, an odd signal will become a neither signal when a constant is added to it. In many cases, the addition of a constant makes the only difference between a neither and an otherwise odd signal. In such cases, it may be useful to remove the DC-offset early in computation, then to take advantage of the properties of the remaining odd signal. For example:

$$s(t) = 0.5 + \cos(t) \quad \text{is even}$$
$$s(t) = 0.5 + \sin(t) \quad \text{is neither}$$

## 2.3.3 Trigonometric Identities

Trigonometric identities serve as an example of combining sine (odd) and cosine (even) functions using addition and multiplication. Conversely, the rules for combining odd and even functions may help in remembering these useful identities. Examine the three identities below in terms of odd and even functions. For example, $\sin(x) \times \sin(y)$ is the product of two odd functions and must equate to the sum of two cosines (even).

$$2 \sin x \cos y = \sin(x - y) + \sin(x + y) \tag{2.1}$$
$$2 \cos x \cos y = \cos(x - y) + \cos(x + y) \tag{2.2}$$
$$2 \sin x \sin y = \cos(x - y) - \cos(x + y) \tag{2.3}$$

As a further memory aid, the exact order and signs required in these identities can easily be confirmed with a pocket calculator.

**EXAMPLE 2.3**

With $x = 1$ rad and $y = 4$ rad, verify the rule in Equation 2.2.

**Solution:**

$$\begin{aligned} LHS &= 2\cos(1)\ \cos(4) \\ &= 2 \times 0.540 \times -0.694 = -0.706 \end{aligned}$$

$$\begin{aligned} RHS &= \cos(1-4) + \cos(1+4) \\ &= \cos(-3) + \cos(5) = -0.990 + 0.284 = -0.706 \end{aligned}$$

Since $RHS = LHS$, the formula is shown to work for this specific example.

## 2.3.4 The Modulation Property

An important consequence of the above sine and cosine identities is the observation that when sines and cosines of different frequencies are multiplied (mixed) together, the result is always two sinusoids at the *sum and difference* of their two frequencies, respectively. Modern communication systems could not exist without recourse to this *modulation property* of mixing sinusoids. The modulation property is used extensively in electronic circuits to alter the frequency of a signal by multiplying it by a cosine.

**A Television Tuner Box** Cable and satellite providers generally transmit multiple television channels on different frequencies, and an external tuner box is used to select one channel for viewing. To this end, the frequency of the selected channel must be brought down to a single common frequency that is used to feed an attached television. This frequency conversion is accomplished using the modulation property of sinusoids, and the tuner box essentially contains a cosine generator and a mixer.[2]

Let the final target frequency be $f_0$ Hz and the channel $N$ of interest be at $f_N$ Hz. If the desired signal at $f_N$ is multiplied with some *mixing frequency* $f_m$, then the output will include the sum $f_{sum} = f_m + f_N$ and the difference $f_{diff} = f_m - f_N$. To achieve the

[2] In the days of analog television, such a box would be connected to the cable system and to a television set that remained tuned typically to Channel 3 regardless of the channel being watched.

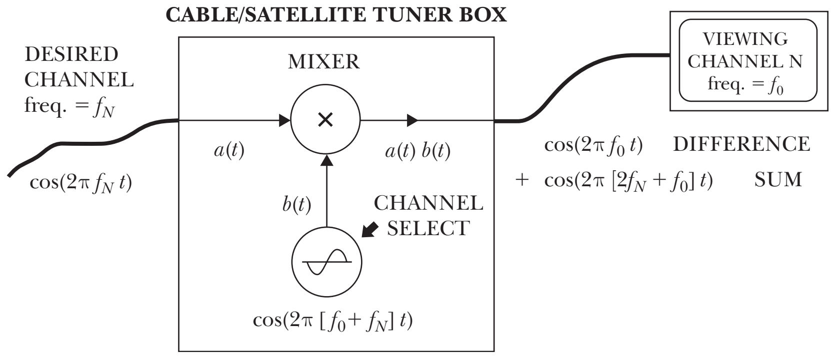



**FIGURE 2.12** A television tuner box uses the modulation property to select a channel by varying the frequency of an internal cosine generator.

desired frequency $f_0$, the mixing frequency can be chosen to be $f_m = f_N + f_0$ giving a difference $f_{diff} = f_0$ and a sum $f_{sum} = 2f_N + f_0$, which may be discarded. The signal of interest is found in the difference term at the frequency $f_0$.

In Figure 2.12, if the input from the cable (containing all available channels) is fed to this mixer, then the desired channel at $f_N$ can be selected by setting the cosine generator to the frequency $f_m = f_N + f_0$ after which a lowpass filter isolates only the signal components near $f_0$ Hz. Different channels at other frequencies can be viewed by changing the oscillator frequency in the box and without changing anything on the television receiver.

| | |
|---|---|
| Desired channel | $f_N$ |
| Mixing frequency | $f_m = f_N + f_0$ |
| Sum | $f_{sum} = 2 \times f_N + f_0$ |
| Difference | $f_{diff} = f_0$ |

## EXAMPLE 2.4 (The Product of Two Sinusoids)

*What is the period of the signal* $g(t) = \cos(2t) \times \cos(3t)$?

**Solution:**

From the modulation theorem, the product of two cosines should give one cosine at the sum of the two frequencies plus another at the difference of the two frequencies. By inspection, the components in the sum are $\cos((2+3)t)$ and $\cos((2-3)t)$. Since cosine is an even signal, the even terms are $\cos(5t)$ and $\cos(t)$. This observation may be confirmed by applying a trigonometric identity from Section 2.3.3.

$$g(t) = \cos(2t) \times \cos(3t)$$

$$= \frac{1}{2}[\cos(t) + \cos(5t)]$$

| SIGNAL | Amplitude | Phase | Frequency | Period |
|---|---|---|---|---|
| $\frac{1}{2}\cos(t)$ | $1/2$ | 0 | $1/2\pi$ | $2\pi$ |
| $\frac{1}{2}\cos(5t)$ | $1/2$ | 0 | $5/2\pi$ | $2\pi/5$ |

The new period is given by LCM$[2\pi, 2\pi/5] = 2\pi$ s, as seen in Figure 2.13.

**Check**: From the graph, the product of two even signals has led to an even result. The final waveform resembles a rapidly changing cosine slowing oscillating up and down; these are the sum (high) and difference (low) frequency components.

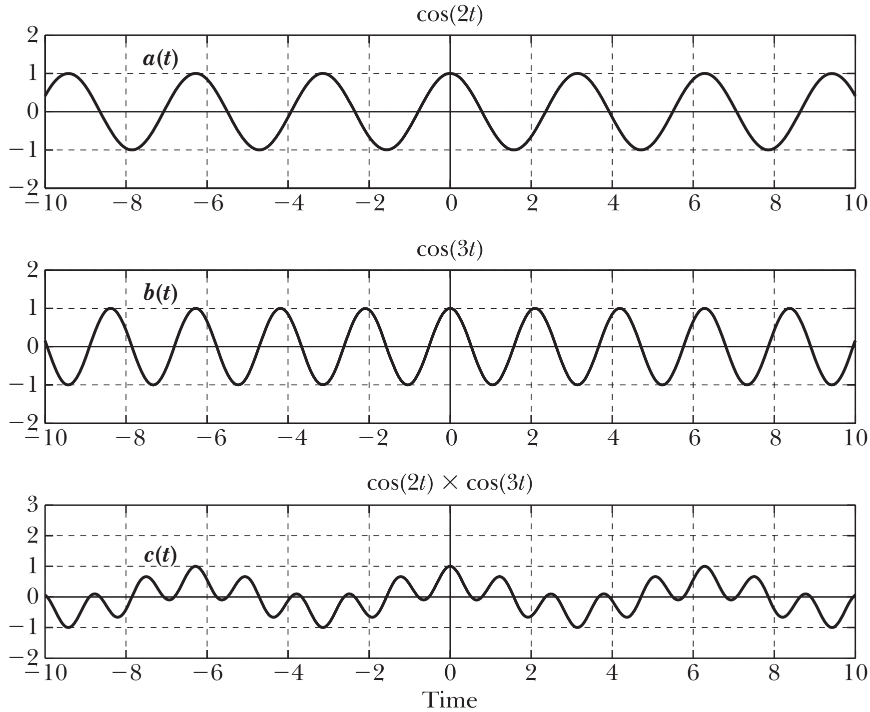



**FIGURE 2.13** The product of Two Cosines $c(t) = a(t) \times b(t)$—The overall period is the LCM of the two periods—Period: $T = 2\pi$ s.

**Squaring the Sinusoid** Squaring a sinusoid such as $s(t) = A\sin(2\pi f_0 t)$ gives a raised cosine with amplitude $A^2/2$ and a frequency that has doubled.[3] Multiplying any signal $s(t)$ by itself yields $s^2(t)$, which is strictly non-negative. Furthermore, if any signal $s(t)$ is odd or even, the squared signal $s^2(t)$ will always be even.

For the squared cosine, use the identity $2\cos x \cos y = \cos(x - y) + \cos(x + y)$ from Eqn. 2.2 and let $x = y$, giving:

$$2\cos y \cos y = \cos(y - y) + \cos(y + y)$$

$$cos^2 y = \frac{1}{2}[1 + \cos(2y)] \tag{2.4}$$

[3] It is a common error to imagine that the squared sinusoidal signal resembles a full-wave rectified cosine.

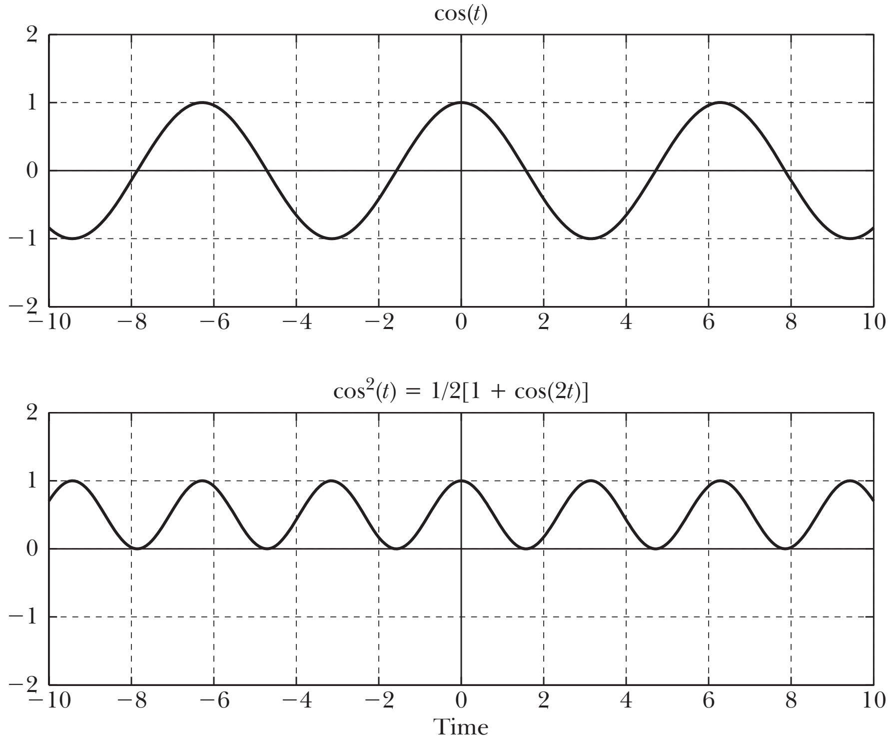



**FIGURE 2.14** Squaring a cosine doubles its frequency—New Period: $T = \pi$ s.

For the squared sine, use the identity $2\sin x \sin y = \cos(x-y) - \cos(x+y)$ from Eqn. 2.3 and let $x = y$, or:

$$sin^2 y = \frac{1}{2}[1 - \cos(2y)] \tag{2.5}$$

Consider the signal $s(t) = A\cos(2\pi f_0 t)$, and let $y = 2\pi f_0 t$, then from Eqn. 2.4, $s^2(t) = \frac{1}{2}A^2[1+\cos(4\pi f_0 t)]$ that is strictly positive, even, and has a value of $s^2(0) = A^2$ at the origin ($t = 0$). Moreover, the cosine term has double the frequency of the original cosine. The result is shown in Figure 2.14 for $A = 1$ and $f_0 = 1/2\pi$ Hz

The appearance of $\cos^2(2\pi f_0 t)$ is a raised cosine with double the frequency $f_0$.

## EXAMPLE 2.5 (Manipulating Odd and Even Signals)

*Let* s(t) = 3 *sin*(t) *What is the appearance of* $s^2$(t)? Sketch *$s^2(t)$* as a function of time.

**Solution:**

This signal $s^2(t)$ should be non-negative, even, and have a value of zero at the origin. Figure 2.15 confirms the observation that squared signals must be strictly non-negative. An odd signal becomes even when squared.

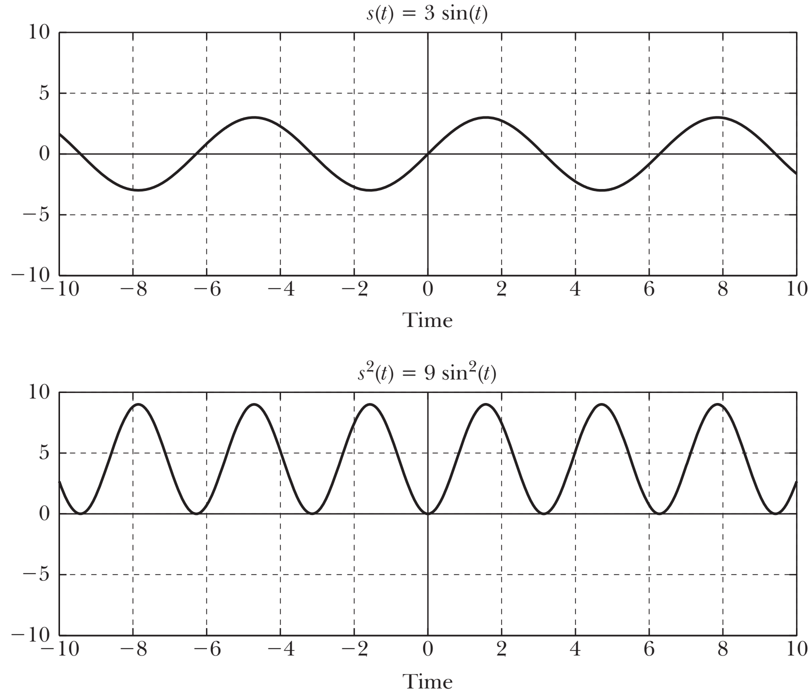



**FIGURE 2.15** When the odd signal $s(t)$ is squared, the result must be both even and non-negative.

## 2.4 Energy and Power Signals

The practical use of signals in electrical engineering often involves the calculation of the power or the energy contained in a signal. For example, when a signal passes through a system, the input and output energy or power may be compared to measure losses in the system. Because a signal is a voltage or current changing in time, then power and energy may also be useful as a means of classifying different signal types.

Using Ohm's law, the instantaneous power in a real signal $s(t)$ at a given time $t$, is equal to $power(t) = \frac{V^2(t)}{R}$ or $I^2(t)R$; however, in the study of signals, it is commonplace to simplify these definitions by assuming a $1\Omega$ resistive load. With $R = 1\Omega$, the instantaneous signal power is identical for both voltage and current signals, or:

$$power(t) = s^2(t)$$

The energy in a signal $s(t)$ is given by the integral of power over all time, or:

**DEFINITION 2.5 Signal Energy**

*For a signal* s(t), *let* E *be defined such that*

$$E = \int_{-\infty}^{+\infty} s^2(t)\, dt$$

*then* E *is called the* energy *in the signal.*

The power (or *average power*) in a signal is given by the average instantaneous power computed in the limit as:

**DEFINITION 2.6 Signal Power**

*For a signal* s(t), *let* P *be defined such that*

$$P = \lim_{x\to\infty} \frac{1}{x} \int_{-\frac{x}{2}}^{+\frac{x}{2}} s^2(t)\, dt$$

*then* P *is called the* power *in the signal.*

For some signals, such as $s(t) = \cos(t)$, the energy will be infinite, while the average power will be finite. In particular, periodic signals will have infinite energy, since the squared magnitude is positive and this energy integral can only grow to infinity as the signal is defined *for all time*. Signals can be classified accordingly as:

**DEFINITION 2.7 Energy Signal**

*Let the energy in a signal* s(t) *be* E, *where:*

$$E = \int_{-\infty}^{+\infty} s^2(t)\, dt$$

*if* $0 < \mathrm{E} < \infty$, *then* s(t) *will be called an* energy signal.

**DEFINITION 2.8 Power Signal**

*Let the average power in a signal* s(t) *be* P, *where:*

$$P = \lim_{x\to\infty} \frac{1}{x} \int_{-\frac{x}{2}}^{+\frac{x}{2}} s^2(t)\, dt$$

*if* $0 < \mathrm{P} < \infty$, *then the signal* s(t) *will be called a* power signal.

From the above discussion, all power signals will have infinite energy. On the other hand, all energy signals will have zero power (as the term $\frac{1}{x}$ will ensure that this integral tends to zero as $x$ grows). If a periodic signal is turned on and off, the resulting $s(t)$ has finite energy. An exponentially decaying signal is another signal with finite energy. These results are summarized below.

| | Energy Signal | Power Signal |
|---|---|---|
| Energy | finite | infinite |
| Power | zero | finite |

The average or mean voltage of a power signal is the square root of the effective signal power, or the square root of the power averaged over all time. The *square root of the mean squared voltage* or *root mean squared* ($V_{rms}$) level of a signal is given by:

$$\text{root mean squared voltage} = \sqrt{\text{average power}} = \sqrt{\lim_{x\to\infty} \frac{1}{x} \int_{-\frac{x}{2}}^{+\frac{x}{2}} s^2(t)\,dt}$$

## 2.4.1 Periodic Signals = Power Signals

For a periodic signal with period $T$, the average power in $s(t)$ is the same over one period as for the entire signal and reduces to:

$$\textit{average power} = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} s^2(t)\ dt$$

**EXAMPLE 2.6 (Power and $V_{rms}$)**

*Determine the power and* $V_{rms}$ *of the voltage signal* $s(t) = A\cos(2\pi f_0 t)$.

**Solution:**

Let the period be $T = 1/f_0$, then,

$$\textit{average power} = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} [A\cos(2\pi f_0 t)]^2\ dt = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} A^2 \times \frac{1}{2}[1+\cos(4\pi f_0 t)]\ dt$$

$$= \frac{A^2}{2} + 0$$

Since the integral of a cosine over any complete period is zero, only the constant term remains, leaving:

$$\textit{average power} = \frac{A^2}{2}$$

$$V_{rms} = \sqrt{\textit{average power}} = \frac{A}{\sqrt{2}}$$

The familiar factor of $0.707A = \frac{A}{\sqrt{2}}$ emerges from this computation. Note that since this $V_{rms}$ calculation is performed over a complete period, it would be the same for any frequency cosine. For a given periodic waveform, $V_{rms}$ is a function of amplitude only.

**$V_{rms}$ Does not Equal $A/\sqrt{2}$ for All Periodic Signals** The root mean squared voltage of an arbitrary periodic signal with amplitude $A$ is not always given by $V_{rms} = A/\sqrt{2} = 0.707A$. This relationship refers specifically to a sinusoidal signal with no DC offset, as computed above. Of course, in a study of alternating current circuits, waveforms are implicitly sinusoidal, and the factor $1/\sqrt{2}$ is often applied routinely during calculations in power systems or electronics, leading to the

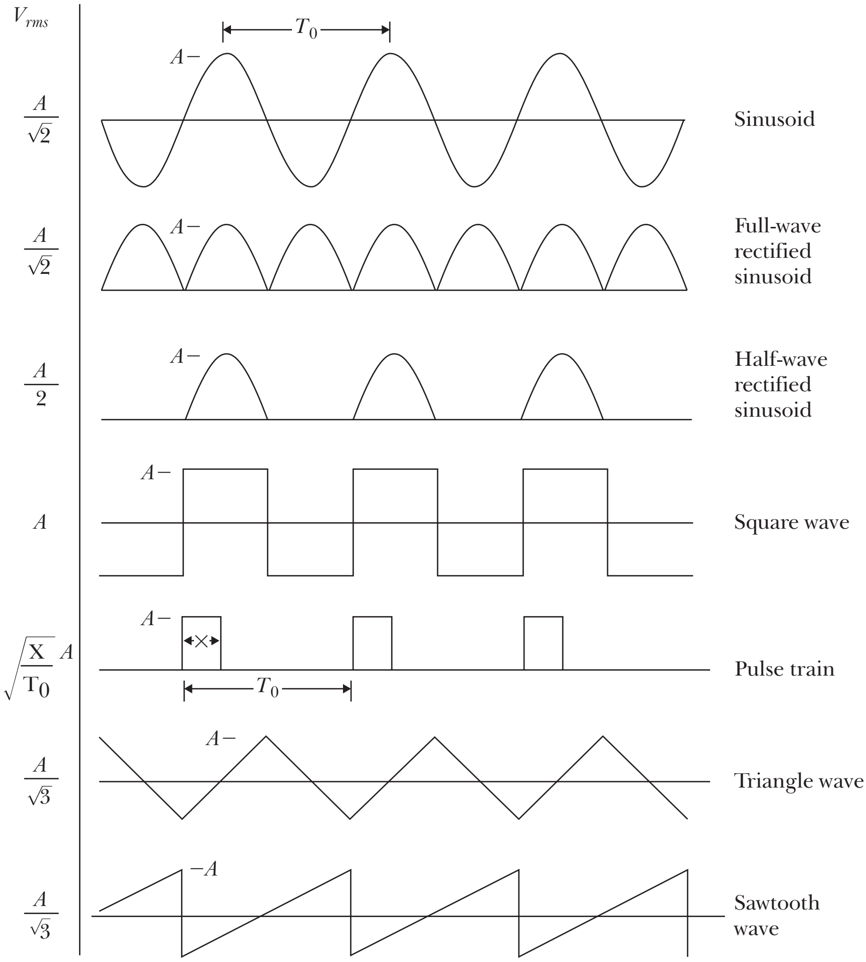



**FIGURE 2.16** $V_{rms}$ for some common periodic signals. Each signal has amplitude $A$.

common misconception that the result applies to all periodic signals. Worse, because of this, an inexpensive AC voltmeter may simply scale a reading by this same factor, assuming a sinusoidally varying input voltage. Only a *true rms* voltmeter can give accurate readings for a square wave and other periodic signals; such instruments are generally more expensive than ordinary AC voltmeters. Figure 2.16 shows the calculated $V_{rms}$ for various periodic signals, each of amplitude $A$.

**MATLAB Exercise 1: Computation of $V_{rms}$** In the section, the $V_{rms}$ value is found for a sawtooth waveform with period $T = 2\pi$ s and amplitude $A = 5$. Numerical computation of $V_{rms}$ requires integration of the squared waveform over a complete period. From Figure 2.16, $V_{rms} = A/\sqrt{3}$, independent of period. Therefore, the expected answer is $V_{rms} = 5/\sqrt{3} = 2.887$.

The same result can readily be confirmed using MATLAB. First, define a period $T = 2\pi$ s using a spacing increment 0.01, for 100 signal points per unit time.

```
t = 0 : 0.01 : 2*pi;
```

Next, use the built-in sawtooth function to define a sawtooth waveform with an amplitude of 5.

```
s = 5 * sawtooth(t); % default period is 2 pi
```

The waveform must now be squared. Multiply the signal point by point with itself. Call the squared signal `s2`.

```
s2 = s .* s;
```

Numerical integration in MATLAB is accomplished using the `trapz()` function.

```
integral = trapz(t, s2)
integral =
          52.2860
```

The mean value is found by dividing the integral by one period $T = 2\pi$, after which the square root gives the root mean value.

```
result = sqrt (integral / (2*pi))
result =
        2.8846
```

This confirms the expected result $V_{rms} = 2.89$ given by the formula $V_{rms} = A/\sqrt{3}$ for a sawtooth wave.

## 2.4.2 Comparing Signal Power: The Decibel (dB)

When the magnitudes or powers of two signals are compared, the ratio of the two quantities is a simple fraction and has no units. The *decibel* (dB) is a unit of measure describing the ratio of the power of two signals on a logarithmic scale. For example, in the ideal amplifier $a(t) \to ka(t)$ the gain in power is typically expressed in dB. The use of the decibel is commonplace in electrical engineering as it easily expresses very large or small ratios, and the overall effect of a signal passing through multiple systems can be found by simply adding the respective gains.

**DEFINITION 2.9 Decibel**

*If* $\mathrm{P}_1$ *and* $\mathrm{P}_2$ *are the power in signals* $\mathrm{s}_1(\mathrm{t})$ *and* $\mathrm{s}_2(\mathrm{t})$ *then:*

$$n = 10 \log\left[\frac{P_1}{P_2}\right]$$

*is the power ratio measured in* decibels *and written as* n *dB.*

The decibel is defined as a *power* ratio, so when comparing the $V_{rms}$ voltages $V_1$ and $V_2$ in two signals $s_1(t)$ and $s_2(t)$, it is assumed that power $P_1 = V_1^2$ and $P_2 = V_2^2$ to construct the corresponding ratio:

$$10 \log\left[\frac{P_1}{P_2}\right] = 10 \log\left[\frac{V_1}{V_2}\right]^2 = 20 \log\left[\frac{V_1}{V_2}\right] dB$$

TABLE 2.3

**Decibel Units (dB)**

| Ratio | Power | Voltage |
|---|---|---|
| 10 | +10 dB | +20 dB |
| 2 | +3 dB | +6 dB |
| 1 | 0 dB | 0 dB |
| $\frac{1}{2}$ | −3 dB | −6 dB |
| $\frac{1}{10}$ | −10 dB | −20 dB |
| $\frac{1}{100}$ | −20 dB | −40 dB |
| $\frac{1}{1000}$ | −30 dB | −60 dB |

Table 2.3 shows a change in power or magnitude (voltage or current) as a ratio and the corresponding change in power expressed in dB. For example, in the top row of the table, a 10 times difference in power is described by +10 dB, while a 10 times difference in voltage equates to a 100 times difference in power (+20 dB).

In Table 2.3 it may be observed that:

- An increase in magnitude (power) is a positive dB change, while a decrease in magnitude (power) is a negative dB change, and 0 dB represents no change in magnitude (power).
- The overall result of several magnitude (power) changes may be added in dB units.

**EXAMPLE 2.7** **(Systems in Series)**

*A voltage signal* $\mathrm{s(t)}$ *passes through two systems in series, described by* $\mathrm{a(t)} \rightarrow 0.5\,\mathrm{a(t)}$ *then* $\mathrm{b(t)} \rightarrow 10\,\mathrm{b(t)}$. *What is the overall change in power expressed in dB?*

**Solution:**

The voltage is divided by 2 in the first system (−6 dB), then multiplied by 10 in the second (+20 dB), so the overall system is like $c(t) \rightarrow 5c(t)$. The corresponding power changes in dB can be added to give the overall power gain as $-6 + 20 = 14$ dB.

**Check:** A voltage gain of 5 is found directly as: $20 \log\left[\frac{5}{1}\right] = 13.98$ dB.

## 2.5 Complex Signals

While signals in the laboratory are necessarily real, it is mathematically useful to allow for complex signals. A complex signal is defined as having *real* and *imaginary* parts, where imaginary components are the product of a real function with $j = \sqrt{-1}$. A complex signal $s(t)$ can be defined as;

$$s(t) = a(t) + j\,b(t) \tag{2.6}$$

where $a(t)$ and $b(t)$ are real functions and its *complex conjugate* as:

$$s^*(t) = a(t) - j\,b(t) \tag{2.7}$$

The real and imaginary parts of $s(t)$ are given by:

$$\textbf{Real part: } \mathrm{Re}[s(t)] = \mathrm{Re}[a(t) + j\, b(t)] = a(t) \tag{2.8}$$

$$\textbf{Imag part: } \mathrm{Im}[s(t)] = \mathrm{Im}[a(t) + j\, b(t)] = b(t) \tag{2.9}$$

Using the usual vector notation for the Argand (complex) plane, the *magnitude* (or *modulus*) and phase of a signal $s(t) = a(t) + j\, b(t)$ can be defined as:

$$\textbf{Magnitude of } \boldsymbol{s(t)}\textbf{:} |s(t)| = \sqrt{s(t)s^*(t)} = \sqrt{a^2(t) + b^2(t)} \tag{2.10}$$

$$\textbf{Phase of } \boldsymbol{s(t)}\textbf{:} \Phi(t) = \tan^{-1}\left(\frac{\mathrm{Im}[s(t)]}{\mathrm{Re}[s(t)]}\right) = \tan^{-1}\left(\frac{b(t)}{a(t)}\right) \tag{2.11}$$

**Check:** For a purely real signal, $b(t) = 0$, and the magnitude is simply $|s(t)| = |a(t)|$ while the phase term becomes $\Phi(t) = 0$.

The complex exponential $e^{j\theta}$ has a special role in signals analysis as it incorporates both real and imaginary sines or cosines. Real sinusoids can conveniently be expressed as complex functions by *Euler's identity*:

$$e^{j\theta} = \cos(\theta) + j\sin(\theta) \tag{2.12}$$

The identity can be rewritten with $\theta = 2\pi f_0 t$ to give the general form of an expression for all sinusoids, both real and complex.

$$e^{j2\pi f_0 t} = \cos(2\pi f_0 t) + j\sin(2\pi f_0 t) \tag{2.13}$$

This form of the general sinusoid has the further advantage that the derivative of an exponential is easily computed by hand, as are the corresponding integrals. The identities shown below for both sine and cosine will prove to be especially important.

$$\cos(2\pi f_0 t) = \frac{e^{+j\,2\pi f_0 t} + e^{-j\,2\pi f_0 t}}{2} \tag{2.14}$$

$$\sin(2\pi f_0 t) = \frac{e^{+j\,2\pi f_0 t} - e^{-j\,2\pi f_0 t}}{2j} \tag{2.15}$$

To fully represent a complex signal graphically requires a three-dimensional graph as (real, imaginary, time). In two dimensions, one of two conventions is generally employed to graph complex signals:

1. As real and imaginary components, often on the same graph in different colors (or using a dotted line for the imaginary part).
2. As magnitude and phase, drawn on separate graphs.

The utility of using a complex signal representation stems in part from the second graphical form above, since the terms *Magnitude graph* and *Phase graph* bring to mind the one-sided cosine *Amplitude* vs. *frequency* and *Phase* vs. *frequency* graphs, both of which are required to represent both sines and cosines. The implication of this comparison is that these two graphs can be combined into a unified (complex) mathematical form. This assertion is borne out by the fact that both sine and cosine signals can be directly expressed in terms of the complex exponential $e^{-j\,2\pi f_0 t}$. While there is no such thing as a complex signal in the laboratory, use of such a representation is commonplace in electrical engineering as a mathematical tool—one that ultimately simplifies signal manipulations.

**EXAMPLE 2.8** **(Plotting a Complex Signal)**

Sketch the complex exponential signal $s(t) = e^{-jt}$.

**Solution:**

A complex function can be graphed in two different ways, either as real and imaginary parts, plotted on the same graph, or as magnitude and phase, plotted on different graphs.

1. Recognizing that $e^{-j2\pi f_0 t} = \cos(2\pi f_0 t) - j\sin(2\pi f_0 t)$, and that this $s(t)$ is this same expression with $f_0 = 1/2\pi$, the real and imaginary parts of this signal can be sketched as the (real) cosine and (imaginary) sine as seen in Figure 2.17 (top).

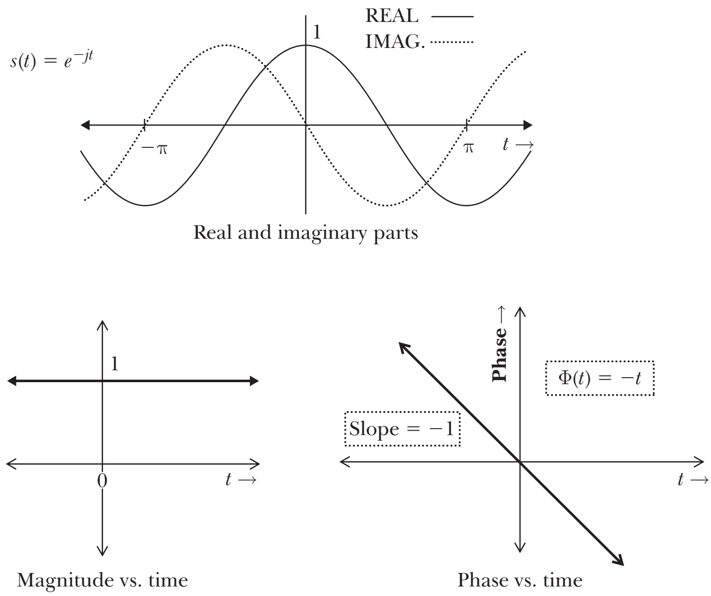



**FIGURE 2.17** Complex Signal $s(t) = e^{-j2\pi f_0 t}$ as real and imaginary parts (top) and as magnitude and phase (bottom).

2. From Equation 2.10, the magnitude $|s(t)|$ is constant 1. From Eqn. 2.11, the phase changes linearly with frequency, as shown in Figure 2.17 (lower part).

In Figure 2.17 the magnitude $|s(t)| = 1$. Consequently, if another signal $g(t)$ was multiplied by $e^{-j2\pi ft}$, the magnitude of $g(t)$ would not be affected. Furthermore, note that this linear phase plot is identical in form to the phase change associated with shifting a time domain signal as seen in Section 1.4.2. This observation suggests that shifting a signal in time is directly related to this frequency-related multiplication by $e^{-jt}$. This relationship will often be encountered.

**MATLAB Exercise 2: Complex Signals** Complex numbers may be readily manipulated in MATLAB where both the variables $j$ and $i$ are predefined to equal

$\sqrt{-1}$. Care must be taken to not overwrite $i$ and $j$ with different values.[4] When displaying complex values, MATLAB labels the imaginary part using the letter $i$ as is conventional in mathematics. For example, it may be confirmed with a calculator[5] that the value of $e^{-j}$ has two parts with real part $\cos(-1) = 0.5043$ and imaginary part $\sin(-1) = -0.8415$.

```
exp(-j)
ans =
   0.5403 - 0.8415i
```

The complex function $a(t) = e^{-jt}$ can be defined as shown below. The real and imaginary parts of a complex variable can be found using `real()` and `imag()`, respectively. This signal may be shown as real and imaginary parts on the same graph as in Figure 2.18 (top).

```
t = -10:0.01:10;          % define time axis
a = exp(-j*t);            % define complex signal
re = real(a);             % real part of a
im = imag(a);             % imag part of a
subplot(3,1,1);           % plot 1 of 3
plot(t,re,t,im,':');
grid on;
```

The phase and magnitude can now be calculated using `angle()` and `abs()` to produce the second form of the graph in Figure 2.18. As a practical matter, any

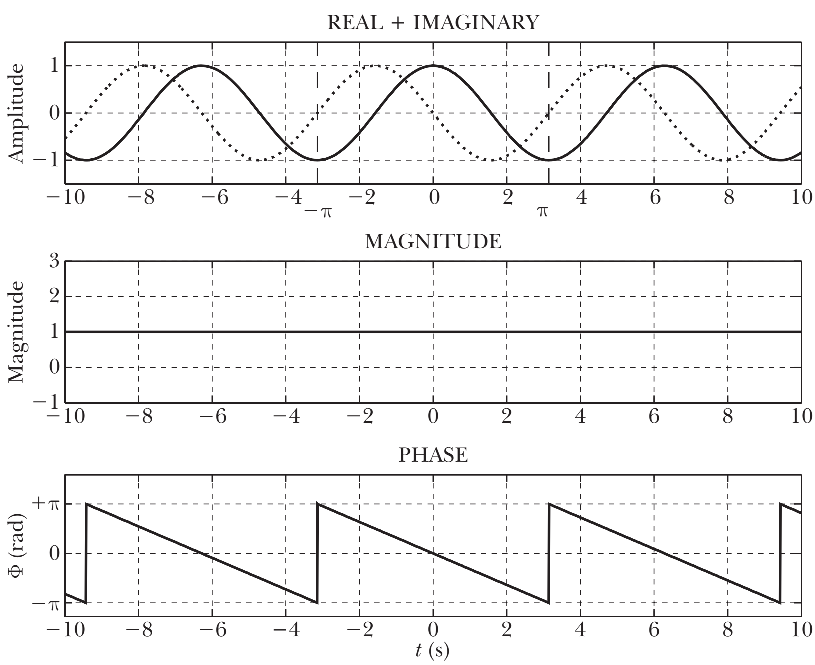



**FIGURE 2.18** Complex Signal $s(t) = e^{-jt}$ showing real and imaginary parts, magnitude and phase. Period = $2\pi$ s.

[4] The command `clear j` would restore $j$ to its default value.
[5] Be sure that the calculator mode is set to radians rather than degrees.

change in phase of $2\pi$ rad may be reduced to no phase change as seen by the apparent phase discontinuities; *the phase change is actually a continuous straight line*. This MATLAB code continues from above:

```
mg = abs(a);          % magnitude of a
ph = angle(a);        % phase of a
subplot(3,1,2);       % plot 2 of 3
plot(t,mg);           % plot the magnitude
grid on;
subplot(3,1,3);       % plot 3 of 3
plot(t,ph);           % plot the phase
grid on;
```

## 2.6 Discrete Time Signals

Consider the impulse train of Section 2.2.9. For each term $c(nT) = \delta(t - nT)$ an impulse is shifted to a different position in time. If a continuous signal $s(t)$ is multiplied by $c(nT)$, then each impulse takes on the area of the signal at each time $nT$:

$$s(t) \times \text{comb}(t,T) = \sum_{n=-\infty}^{+\infty} s(t)\,\delta(t - nT) \tag{2.16}$$

and the resulting sequence of impulses with different areas appears graphically as arrows having heights that follow the variations in $s(t)$, as shown in Figure 2.19. In this way, the impulse train serves to extract values of $s(t)$ (called *samples*) at regularly spaced intervals $nT$. The resulting array of values $s[n]$ is called a *discrete time signal*. Unlike a continuous time signal $s(t)$, the discrete time signal exists only at the specific times $nT$ and yet the values in $s[n]$ nonetheless reflect the variations that characterize the continuous function $s(t)$. The subject of discrete signals is fully covered in Chapter 8.

Any signal manipulated or stored using a digital computer is necessarily represented as a discrete signal where specific memory locations may hold each of the values $s[n]$. Using MATLAB, for example, the mathematical function $\cos(t)$ is continuous but a fixed number of points is necessarily chosen to represent the signal for display or calculation. In the example code below, 51 values of time are defined as `t = -10:0.4:10` to represent a set of sample times taken every 0.4 s, and the corresponding samples of the variable $c$ describe a discrete time signal. Sample points from a discrete time signal are best seen on a graph using the `stem()` function as in Figure 2.20. Another choice of plot maintains the level of each sample point until the next sample time[6] by using the `stairs()` plotting command that gives a continuous stepped appearance as in Figure 2.21. In either case, the discrete time signal $\cos(nT)$ is shown for $T = 0.4$ *sec*.

```
t = -10 : 0.4 : 10;     % one sample every T=0.4 sec
c = cos(t);             % sampled values as cos(nT)
stem (t, c, '.')
axis([-10 10 -2 2 ]);
grid on;
```

[6] Also known as a *sample and hold*.

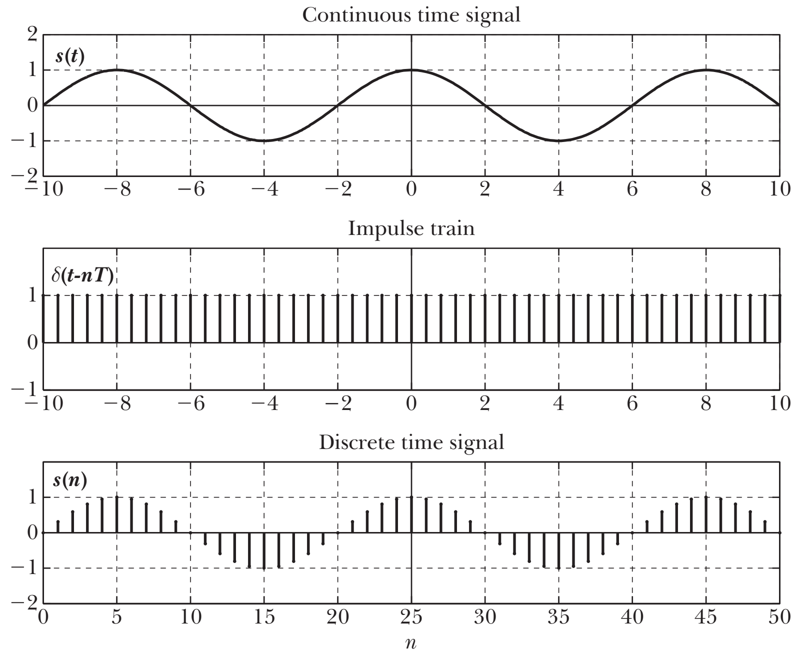



**FIGURE 2.19** A discrete time signal $s[n]$ can be created by multiplying the continuous time signal $s(t)$ by an impulse train $c(t)$. The elements of $s[n]$ hold the corresponding value $s(nT)$ as found in the area of each impulse.

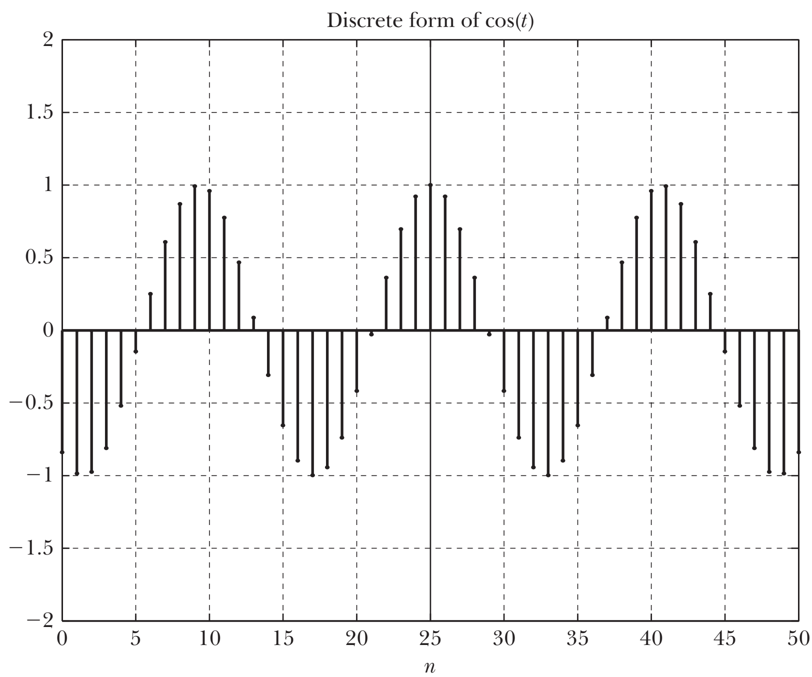



**FIGURE 2.20** A discrete time signal $s[n] = \cos(nT)$, as discrete samples.

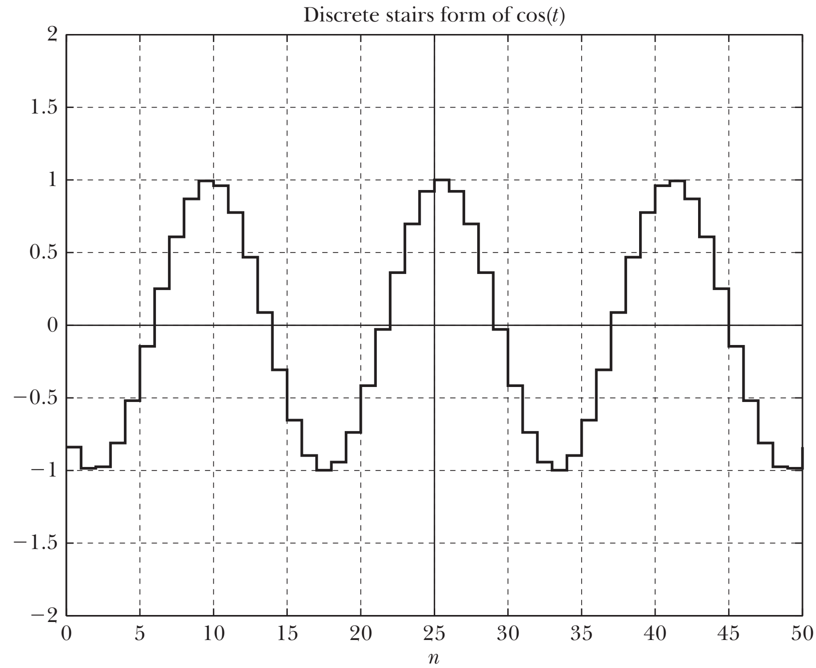



**FIGURE 2.21** A discrete time signal $s[n] = \cos(nT)$, as stepped values.

## 2.7 Digital Signals

Digital signals are fundamental to applications in modern *digital signal processing* (DSP) where a signal is to be processed in a computer. Digital signals may be thought of as practical discrete time signals, since getting an analog signal $s(t)$ into digital form requires the use of voltage sampling hardware (*analog-to-digital converter*), which necessarily constrains each value of $s[n]$ to a fixed set of discrete voltage values. The resulting signal is *both discrete time and discrete voltage*. In an analog-to-digital converter, sample values are acquired as $N$-bit integers, which correspond to one of $2^N$ discrete voltage levels. In Figure 2.22, the signal $cos(t)$ is shown as a digital signal with regular 4-bit samples taken every $T = 0.4$ s; this digital signal is constrained to one of 16 voltage levels.

The cosine signal in Figure 2.22 may appear somewhat distorted because of the discrete time and voltage restrictions. Because each sample value is rounded off to the nearest allowed voltage level, digitizing a signal in this way introduces roundoff error or noise into the resulting digital signal. This effect may be minimized by using more steps; for example, with 10 bits there would be $2^{10} = 1024$ voltage levels. For large $N$ the digital signal resembles the stepped discrete time signal of Figure 2.21.

## 2.8 ◎ Random Signals

If a signal $s(t)$ is *random*, the value of $s(t)$ at any given time cannot be known in advance. Random signals are instead characterized by their statistical properties over time. In comparison to the *deterministic* signals seen so far and that can be expressed as an explicit function of time, the random signal is *non-deterministic* and its value at any given time cannot be predicted.

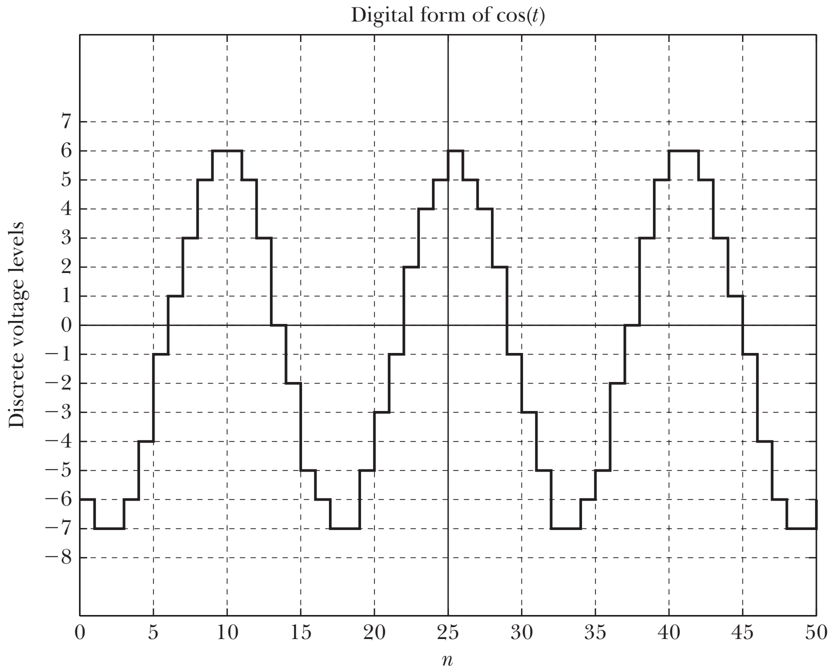



**FIGURE 2.22** A digital signal $s[n] = \cos(nT)$ using 4 bits = 16 discrete voltage levels.

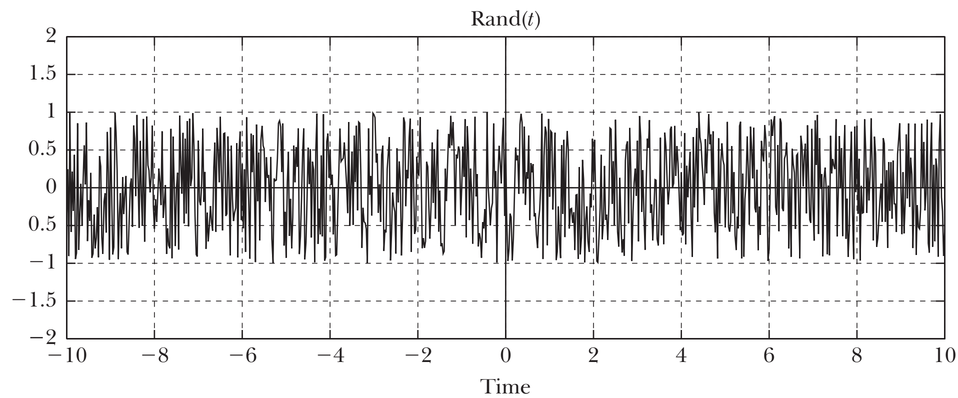



**FIGURE 2.23** A random signal is described by its statistical properties, such as the *mean* and *standard devation* of its value over time.

Random signals are encountered as the unavoidable (hopefully small) *noise* inevitably added to signals of interest. For example, the random signal $s(t)$ may represent the ever-present hiss in a stereo amplifier or heard when an AM radio is tuned between active stations.

Random signals are also useful as test input signals to determine the frequency response of analog and digital circuits. In this case, the ever-changing nature of the random signal implies that the characteristics of a circuit can be tested across a range of input frequencies with a single test signal.

Consider the random signal $s(t)$ shown in Figure 2.23 and existing for all time. The exact value of the waveform at any time $t$ is unpredictable, yet it can be seen that the values seem to fluctuate around 0. This observation defines a *zero mean* random signal. The statistics of the values taken on by the random $s(t)$ over time further distinguish one random signal type from another. By inspection, the above random signal takes on values that are distributed in the range $(-1, +1)$. The random signal is a power signal since the time integral of the squared signal goes to infinity. These properties can be used to model random signals and to study the effect of noise on real systems.

The `rand()` function in MATLAB generates *pseudo-random* values distributed uniformly in the interval (0.0, 1.0). The term pseudo-random refers to the fact that the values generated by a computer program may have the properties of a random signal but are totally predictable if the generating algorithm is known. A random signal plot may be created as:

```
t = -10:0.01:10;            % define time interval t
r = 2*rand(size(t))-1;      % random value range [-1,+1]
plot (t,r);
axis([-10 10 -2 2 ]);
grid on;
```

## 2.9 Useful Hints and Help with MATLAB

While MATLAB is a powerful tool for visualizing and analyzing signals, it must be used with care if results are to be accurate and meaningful. A few common pitfalls can be readily avoided once they are identified.

Care has been taken to limit example code to functions present in the Student Version of MATLAB. In particular, certain useful functions such as `square()` and `rectpuls()` are part of the add-on *Signals Toolbox* that is included with the Student Version. Any such functions that are provided with the Student Version may be optional in other configurations of MATLAB. If some known function cannot be found, this toolbox may be missing from the installation being used.

MATLAB, like all software on a digital computer, represents signals as discrete samples, meaning that the time axis is never continuous. Depending on how a user defines $t$ and the time interval, unexpected or inaccurate results may be obtained.

When defining the time axis as `t = -10:0.01:10;` the second parameter defines the spacing $T$ between time points; in this case, `0.01`. Most of the potential problems listed here are related to differences between ideal continuous signals and the corresponding stored discrete values in a computer.

1. Time interval—In theory, periodic signals like $\cos(t)$ and even the constant value $k$ are defined over all time. When represented in a computer, only a finite range of values for $t$ is available. Consequently, sinusoids must be cut-off somewhere in time, and results will be affected accordingly. For periodic signals, problems can often be avoided by taking care to define $t$ to span an integral number of periods.
2. Integration accuracy—The accuracy of the trapezoidal approximation used by `trapz()` to perform numerical integration is a function of the chosen time increment, where a smaller increment gives better results. To integrate the signal $s(t)$ where `t = -10:0.01:10;` the function `trapz(s)*0.01` is scaled by the time increment; alternatively, the form `trapz(t,s)` may be

used with no adjustment necessary. Care must be taken in checking numerical answers against theoretical results.

3. Vertical lines—Rectangles that do not appear exactly square, impulses that look like triangles, or jagged vertical lines will inevitably be seen in plots, depending on the time increment and the chosen display axes. The time increments suggested for the examples in this chapter have been chosen to minimize such effects.
4. Defining the impulse $\delta(t)$—A workable approximation to an impulse can be created in MATLAB as a vector that is zero everywhere except at the point of interest, where a single value 1 can be placed.
   For example, consider three impulses to be placed at $(0, \pi, 5)$ s:

```
t = -10 : 0.01 : 10;
impA = (t == 0);                  % impulse at t= 0
impB = (t == 5);                  % impulse at t= 5
impC = (t == pi);                 % impulse at t= pi
plot(t, impA + impB + impC);      % plot 3 impulses
axis([-10 10 -2 2 ]);
grid on;
```

   In practice, any of these definitions could fail. Given that $t$ is defined in terms of a series and a fixed-time increment, the limitations of finite word size binary arithmetic mean that $t$ may never exactly satisfy $(t == 0)$ or $(t == 5)$. It is certain that $(t - \pi)$ will never be satisfied on this interval; consequently the impulse at $t = \pi$ is absent on this graph.
5. Vector size—While using a smaller time increment can increase the accuracy of results, the speed of computation will be slowed, and the vector lengths can become unmanageably long. In general, either the time increment or the number of points can be specified when defining a time axis.

```
% Specify an increment, number of points is implicit
t = -10:0.1:10;    % defines 201 points
t = -10:0.01:10;   % defines 2001 points
t = -10:0.001:10;  % defines 20001 points

% Specify a number of points, increment is implicit
t = linspace(-10,10,200);    % defines 200 points
t = linspace(-10,10,2000);   % defines 2000 points
t = linspace(-10,10,20000);  % defines 20000 points
```

## 2.10 Conclusions

A signal $s(t)$ defined as a time-varying voltage or current often has specific properties that characterize its appearance or behavior. Several commonplace signal types have been introduced and their properties explored. Signals analysis is made easier by recognizing fundamental signal types and classifications, and by expressing signals in terms of those fundamental components. Many signals will be more readily appreciated if viewed in the frequency domain, where their properties and behavior can be visualized from a different perspective. The mathematics that allow signal decomposition and a unique frequency domain representation of any signal $s(t)$ will be explored in the next chapter.

# End-of-Chapter Exercises

**2.1** Consider a complex function of time $z(t) = x(t) + jy(t)$.

**(a)** What is the complex conjugate of $z(t)$?
**(b)** What is the magnitude of $z(t)$?

**2.2** Show that the even part of the product of two functions is equal to the product of the odd parts plus the product of the even parts.

**2.3** Identify which of the following signals are even or odd or neither.

$$\begin{aligned} a(t) &= 2\sin(\pi t) \\ b(t) &= 3\cos(2t) \\ c(t) &= 10\sin(2\pi t - \pi/4) \\ d(t) &= \cos(\pi t + \pi/2) \\ e(t) &= 5\sin(\pi(t-1)) \end{aligned}$$

**2.4** Express each of the signals in Figure 2.24 as a linear combination of signals of the form $s(t) = A\cos(2\pi f_0 t + \Phi)$:

| | |
|---|---|
| $w(t) = 30\sin(100\pi t - \pi/2)$ | $a(t) = x(t) + y(t)$ |
| $x(t) = 30\cos(100\pi t)$ | $b(t) = x(t) + y(t) + z(t)$ |
| $y(t) = -10\cos(300\pi t)$ | $c(t) = x(t) \times y(t)$ |
| $z(t) = 6\cos(500\pi t)$ | $d(t) = w(t) \times w(t)$ |

**FIGURE 2.24** Periodic Signals for Questions 2.4 to 2.10.

**2.5** Consider the signals in Figure 2.24.

**(a)** Identify each signal as being odd, or even, or neither.
**(b)** Compute the value of the signal at t = 0.

**2.6** Carefully sketch by hand the signals $(w, x, y, z)$ from Figure 2.24 as amplitude vs. time, and identify the amplitude and period of the signal on your graph. Check your sketches for consistency as odd and even signals.

**2.7** Use MATLAB to plot the signals $(a, b, c, d)$ from Figure 2.24 as amplitude vs. time, and identify the amplitude and period of the signals on your graph. Check each plot against the answers from Question 2.5.

**2.8** Carefully sketch by hand each of the eight signals in Figure 2.24 as amplitude vs. frequency, or the *one-sided cosine graph* and identify the amplitude $A$ and frequency $f_0$ of the (cosine) signal components on each graph.

**2.9** Evaluate the definite integral below making use of the rules for odd and even computations. The signals $a(t)$ and $b(t)$ are defined in Figure 2.24.

$$s(t) = \int_{-10}^{+10} a(t) \times b(t)\, dt$$

**2.10** Evaluate the definite integral below, making use of the rule for integrals involving $\delta(t)$. The signal $a(t)$ is defined in Figure 2.24.

$$s(t) = \int_{-10}^{+10} a(t) \times \delta(t-1)\, dt$$

**2.11** Let $s(t)$ be some well-behaved function of time, and define $a(t)$ and $b(t)$ as:

$$\begin{aligned} a(t) &= |s(2\pi t + 3)| \\ b(t) &= [s(4\pi t + 5)]^2 \end{aligned}$$

**(a)** What simple observation can be made about a graph of the signals $a(t)$ and $b(t)$?
**(b)** Use MATLAB to plot $a(t)$ and $b(t)$, and confirm the above answer for the case of $s(t) = sin(t)$.

**2.12** Show how a full wave rectified cosine with period $2\pi$ sec (see Figure 2.4) can be modelled as the product of a cosine and a square wave.

**2.13** Show how the half-wave rectified cosine with period $2\pi$ sec (see Figure 2.3) can be modelled as the product of a cosine and a square wave.

**2.14** The trigonometric identities relating sines and cosines can be manipulated to give additional rules. Show that the relationships below are consistent with the rules for combining odd and even functions.

**(a)** $\sin(x+y) = \sin(x)\cos(y) + \cos(y)\sin(x)$
**(b)** $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$
**(c)** $\sin(x) + \sin(y) = 2\sin([x+y]/2)\cos([x-y]/2)$
**(d)** $\cos(x) + \cos(y) = 2\cos([x+y]/2)\cos([x-y]/2)$

**2.15** Consider the signals $a(t) = \cos(\pi t)$ and $b(t) = \mathrm{rect}(t)$. Sketch each of the following signals by hand and without the use of a calculator.

**(a)** $a(t)$ and $b(t)$ on the same graph.
**(b)** $c(t) = a(t) \times b(t)$
**(c)** $d(t) = a(t) \times b(t-2)$
**(d)** $e(t) = [1 + a(t)]$
**(e)** $f(t) = e(t) \times b(t/2)$

**2.16** Consider a pulse train with a repetition rate of 1843 Hz and a percent duty cycle of $P$ percent, where $0 \le P \le 100$. The signal is at 1 V for $P$ percent of its period and is 0 V otherwise.

The following hints may be useful in the question:

- Let the period be $T$. (The exact period is not relevant to the $V_{rms}$ calculation. In any case, the specific value for $T$ could be inserted afterwards.)
- Sketch the waveform to help set up the integration.
- Using a pulse centered on the origin may save some math (make an odd or even function).

**(a)** Compute $V_{rms}$ for this waveform, as a function of $P$.
**(b)** Use the result from (a) to find $V_{rms}$ for $P = 50\%$ (square wave with DC-offset).
**(c)** Use the result from (a) to find $V_{rms}$ for $P = 25\%$ (pulse train).
**(d)** Use the result from (a) to find $V_{rms}$ for $P = 75\%$ (pulse train).
**(e)** Use the result from (a) to find $V_{rms}$ for $P = 100\%$ (DC value . . . checks your result.)
**(f)** Use the result from (a) to find $V_{rms}$ for $P = 0\%$ (Zero . . . checks your result.)

**2.17** Consider the signals:

$$a(t) = 34\cos(200\pi t)$$
$$b(t) = 99\sin(150\pi t)$$
$$c(t) = \mathrm{rect}(t)$$

**(a)** Complete the table:

| | Odd or even or neither | Value at $t = 0$ ? | Energy or Power? | Period (sec) |
|---|---|---|---|---|
| $a(t)$ | | | | |
| $b(t)$ | | | | |
| $c(t)$ | | | | |
| $a(t) + b(t)$ | | | | |
| $a(t) \times b(t)$ | | | | |

**(b)** Referring to the signals defined above, make a neat labeled sketch by hand of the signals below.

$$f(t) = 5c(t - 10) - 5c(t + 10)$$
$$g(t) = -2c(t/10)$$
$$a^2(t) = 1156\cos^2(200\pi t)$$

**2.18** Refer to the signal $s(t)$ in Figure 2.25 and answer the following questions.

**(a)** Identify $s(t)$ as odd, even, or neither.
**(b)** Find the period of $s(t)$.
**(c)** Find the value $s(t)$ at $t = 0$.
**(d)** Determine the two sinusoidal components that add to give this signal.
**(e)** Is the above answer unique?

To answer the final part (d), use your observations from a through c above, and make some reasonable assumptions about the component signals.

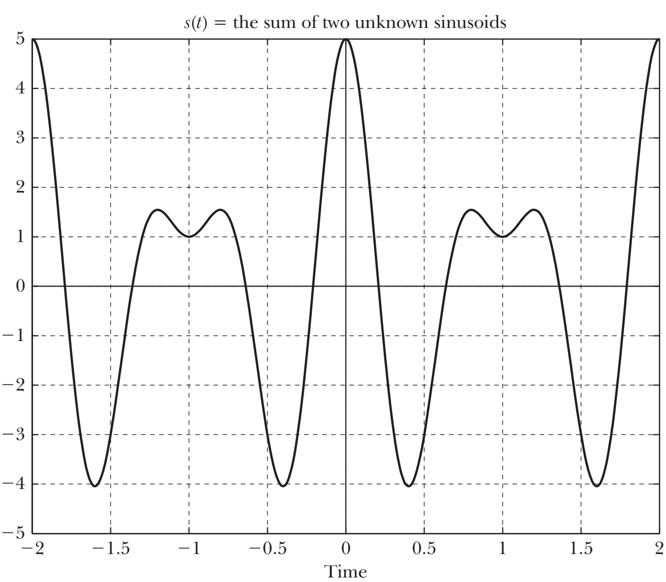



**FIGURE 2.25** Diagram for Question 2.18.

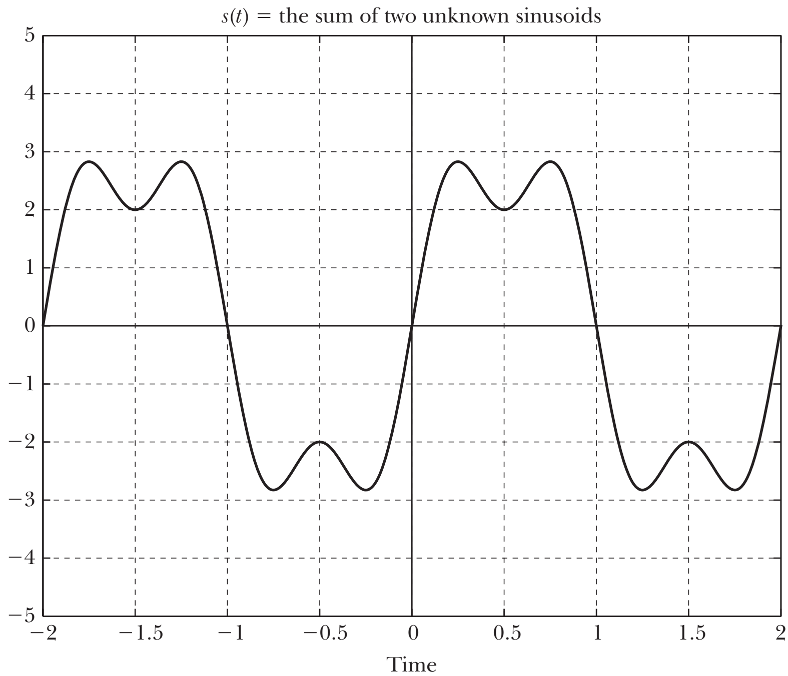



**FIGURE 2.26** Diagram for Question 2.19.

**2.19** Repeat Question 2.18 using the signal $s(t)$ in Figure 2.26.

**2.20** Use the MATLAB `square()` function to create a square wave with amplitude 10 V and period 5 s.

- **(a)** Show the steps necessary to create this signal.
- **(b)** Does the MATLAB square wave function create an odd or an even signal?
- **(c)** How could you modify your steps to create an even function?

**2.21** Use the MATLAB `sawtooth()` function to create a sawtooth wave with amplitude 2 V and period 7 s.

- **(a)** Show the steps necessary to create this signal.
- **(b)** Does the MATLAB sawtooth function create an odd or an even signal?
- **(c)** Can a sawtooth ever be an even function?

**2.22** Determine the energy and power in each of the signals below:

$$a(t) = 5\cos(2\pi t)$$

$$b(t) = 5\cos(\pi t)$$

$$c(t) = 5\sin(\pi t)$$

$$d(t) = 5\,\text{rect}(t/3)$$

$$e(t) = \delta(t)$$

- **(a)** Use MATLAB to compute $V_{rms}$ for a 1000 Hz triangle wave with amplitude 3 V.
- **(b)** What is the power in the above signal?
- **(c)** Compare the above answers to Figure 2.16.

**2.23** Sinusoidal inputs to a system are output with different magnitudes depending on frequency as shown in Figure 2.27, where the magnitude variations are shown in dB. If the input signal is $s(t) = A\cos(2\pi f_0 t)$, determine the magnitude and power of the output signal at the three frequencies $f_0 = \{f_A, f_B, f_C\}$ in the graph.

**2.24** Sinusoidal inputs to a system are output with different magnitudes depending on frequency as shown in Figure 2.28, where the magnitude variations are shown in dB. If the input signal is $s(t) = A\cos(2\pi f_0 t)$, determine the magnitude and power of the output signal at the three frequencies $f_0 = \{f_A, f_B, f_C\}$ in the graph.

**2.25** Consider the continuous signal $s(t) = \cos(22\pi t)$ and a train of impulses spaced every 0.1 seconds. A discrete signal is formed as the product of the impulse train with the cosine $s(t)$.

- **(a)** What is the period of $s(t)$?
- **(b)** Construct a table showing values of the discrete signal on the time interval $[0, 1]$ seconds.

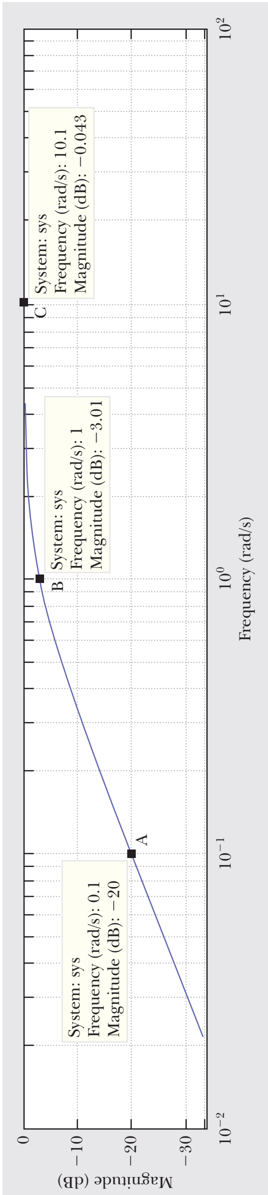



**FIGURE 2.27** Diagram for Question 2.24.

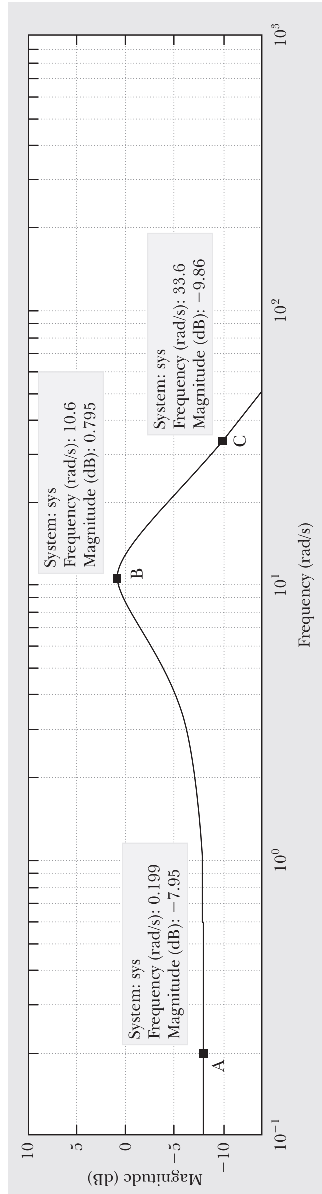



**FIGURE 2.28** Diagram for Question 2.25.

**(c)** From the table, make a sketch of the discrete signal.

**(d)** Comment on the period of the waveform in your sketch. Suggest how this problem may be overcome.

**2.26** Consider the continuous signal $s(t) = \cos(22\pi t)$ and a train of impulses spaced every 0.1 seconds. A discrete signal is formed as the product of the impulse train with the cosine $s(t)$.

**(a)** Use MATLAB to find values of the discrete signal on the time interval [0, 1] seconds.

**(b)** Plot the discrete signal and comment on its period.

**(c)** Repeat this question using impulses spaced every 0.01 seconds.

CHAPTER 3

# Linear Systems

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Describe the properties of a linear time invariant system
- Explain the difference between linear and nonlinear systems
- Describe the concept of response function
- Identify an ideal linear system
- Apply the convolution integral to find the output of a linear system
- Use MATLAB to define signals and to perform convolution operations
- Derive the output of a linear system given the response function
- Explain causality in terms of the impulse response of a linear system
- Apply the properties of a linear system to simplify system block diagrams

The mathematical modelling of systems allows the study of both signals and systems from a purely theoretical perspective. For an engineer, the understanding of the expected behavior of a system as it interacts with signals allows the efficient design and testing of systems both inside and outside the laboratory. In particular, *linear systems* represent an important class of systems based on the general properties of the system's interaction with signals. Most of the systems described in this text will fall into the category of linear systems. For some nonlinear systems (such as transistor amplifiers), the operating conditions can often be constrained to satisfy the conditions of linearity. In other cases, a much more involved and nonintuitive analysis is required, but the study of general noninterior systems is beyond the scope of this discussion.

## 3.1 Introduction

Certain systems exhibit general properties that characterize their behavior and consequently define the mathematical tools that will be useful in signals analysis. The most important class of such systems are referred to as *linear systems*. A linear system is defined in terms of how it interacts with input signals and specifically how it behaves when a linear combination of input signals is presented to the system. Systems that do not fit this general behavior are called *nonlinear systems*.

## 3.2 Definition of a Linear System

Every system responds to an input signal in a distinctive way, and this *response* characterizes the system. For example, if a signal $a(t)$ is input to some arbitrary system, the output may be observed as $a(t) \to a_2(t)$. This output $a_2(t)$ will reflect the response of the system to the input $a(t)$. Similarly, a different input signal $b(t)$ could pass through the same system to give $b(t) \to b_2(t)$. Mathematically, the input and output signals in both cases are related by the same *response function* that defines this system's behavior.[1] By introducing some modest constraints on the system's behavior, the category of a *linear system* can be defined.

### 3.2.1 Superposition

One simple requirement is sufficient to define a linear system. A system is *linear* if the following property holds true for input signals $a(t)$ and $b(t)$:

**DEFINITION 3.1 Superposition**

$$\textit{If } \mathrm{a(t)} \to \mathrm{a_2(t)}$$
$$\textit{and } \mathrm{b(t)} \to \mathrm{b_2(t)}$$
$$\textit{then } [\mathrm{a(t)} + \mathrm{b(t)}] \to [\mathrm{a_2(t)} + \mathrm{b_2(t)}]$$

In other words, when the signals $a(t)$ and $b(t)$ added together enter this system, the output will be the sum of the outputs produced by each signal individually. Just as the input signals are added together, the individual output signals are superimposed; this behavior is know as the property of *superposition.* All linear systems must obey superposition. It is a significant step towards generalizing the behavior of this system for all possible input signals.

It follows that for an input of $2a(t)$, the output is $2a(t) \to 2a_2(t)$, since:

$$a(t) + a(t) \to a_2(t) + a_2(t)$$

By extension, the definition of a linear system can be generalized to include any linear combination of input signals. If a linear combination of signals $a(t)$ and $b(t)$ is input to a linear system, the output will be equivalent to the linear combination of the corresponding output signals, considered separately. Consequently, with constants $k_a$ and $k_b$, the definition of superposition becomes:

$$[k_a a(t) + k_b b(t)] \to [k_a a_2(t) + k_b b_2(t)]$$

A system can be classified as being linear or nonlinear by checking if it obeys the superposition property. Several examples will reveal some important properties of linear systems that emerge from the superposition property.

[1] It is customary to label the response function as $h(t)$.

### 3.2.2 Linear System Exercise 1: Zero State Response

Consider a system containing a 1.5 V battery such that every input signal is output with 1.5 V added to it. The system *responds to* any input signal by adding 1.5 V. In this case, the operation *add 1.5 V* is the response function of the system, as shown in Figure 3.1. It will be shown that this is *not* a linear system.

To establish linearity, the property of superposition can be checked by observing the system response to test inputs $a(t)$ and $b(t)$ and then to both inputs added together.

Let two input signals be $a(t)$ and $b(t)$ such that:

$$a(t) \to a(t) + 1.5$$
$$b(t) \to b(t) + 1.5$$

Let *RHS* be the output for the system for the sum $s(t) = [a(t) + b(t)]$:

$$RHS : s(t) \to s(t) + 1.5 = [a(t) + b(t)] + 1.5$$

Let *LHS* be the expected output for a linear system, or the sum of the individual outputs observed for $a(t)$ and $b(t)$ separately.

$$LHS : a(t) + b(t) \to [a(t) + 1.5] + [b(t) + 1.5] = [a(t) + b(t)] + 3.0$$

Since $LHS \neq RHS$, this is *not* a linear system.

**Zero Input → Zero Output** Closer inspection reveals that this could not be a linear system if any non-zero function is added to an input signal. This implies that for the input signal $a(t) = 0$, the output signal is expected to be zero. Put another way, *if a linear system is at rest* then, until some input arrives, the output will be zero. The resulting response to a new input is called the *zero state response.*

**Initial Conditions** In practice, linear systems may have certain non-zero *initial conditions* at the moment an input signal arrives. For example, a capacitor in a circuit may not be fully discharged following some past inputs. Such a situation adds a *zero input response* to the *zero state response* to give the overall system output. Initial conditions will be assumed to be zero unless otherwise stated.

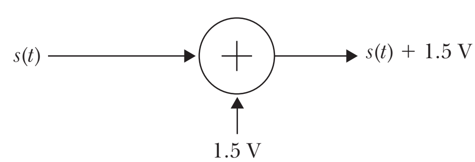



**FIGURE 3.1** Addition of a Constant—The output of this system is the sum of the input signal $s(t)$ and 1.5 V.

### 3.2.3 Linear System Exercise 2: Operating in a Linear Region

Consider a system containing an ideal diode (or *half-wave rectifier*) as shown in Figure 3.2. It will be shown that this is *sometimes* a linear system.

Mathematically, this system passes only positive values of an input signal, and any negative values will be set to zero. For an input signal $a(t)$:

$$a(t) \to a_2(t) = \begin{cases} a(t) & \text{if } a(t) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

To establish linearity, the property of superposition can be checked by observing the system response to test inputs $a(t)$ and $b(t)$ and then to both inputs added together. A specific example will demonstrate graphically that a system is linear *only under certain input conditions.*

Let two input signals be $a(t) = \cos(2\pi f_0 t)$ and $b(t) = 1$, as shown in Figure 3.3. The output signals $a_2(t)$ and $b_2(t)$ are shown, along with their sum, after passing through the ideal diode circuit. Another input signal $s(t)$ is created from the sum of the two input signals: $s(t) = [a(t) + b(t)]$. The output signal $s_2(t)$ does not equal $a_2(t) + b_2(t)$, and therefore, this is *not* a linear system. Any negative values of $a(t)$ are blocked by the sytem.

By inspection, this system is linear in the region $a(t) > 0$ because the system is $a(t) \to a(t)$ under this condition.

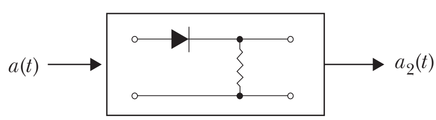



**FIGURE 3.2** This nonlinear system $a(t) \to a_2(t)$ includes an ideal diode.

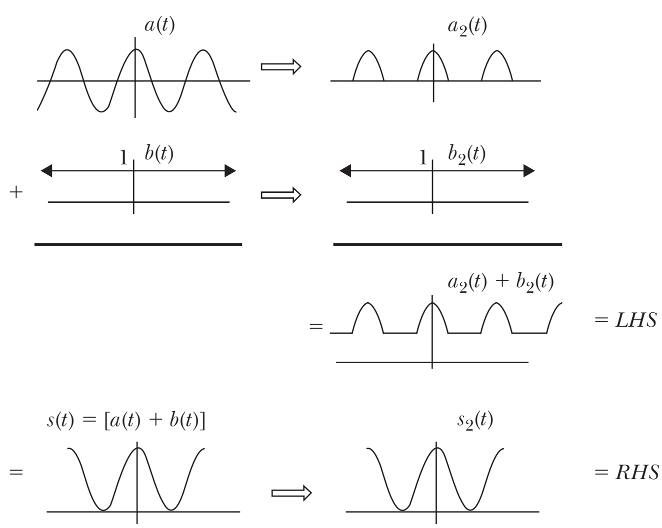



**FIGURE 3.3** The system $a(t) \to a_2(t)$ implements a half-wave rectifier. This sketch demonstrates violation of the superposition property of linear systems.

**Nonlinear Components** In electronic circuits, diodes and other semiconductor components are known to be *nonlinear* devices. On the other hand, many electronic circuits such as the one described previously may nonetheless be linear within the confines of a certain operating region. Transistor circuits such as amplifiers are typically designed to function at an operating point where (locally) they can be treated as linear circuits and undesirable operation (e.g., output distortion) can result whenever this condition is violated.

### 3.2.4 Linear System Exercise 3: Mixer

Consider a system containing a mixer, such that all input signals are multiplied by another signal $c(t)$, as shown in Figure 3.4. It will be shown that this *is* a linear system.

To establish linearity, the property of superposition can be checked by observing the system response to test inputs $a(t)$ and $b(t)$, and then to both inputs added together.

Let two input signals be $a(t)$ and $b(t)$ such that:

$$a(t) \to a(t) \times c(t)$$

$$b(t) \to b(t) \times c(t)$$

Let *RHS* be the output for the system for the sum $s(t) = [a(t) + b(t)]$:

$$RHS : s(t) \to s(t) \times c(t) = [a(t) + b(t)] \times c(t)$$

Let *LHS* be the expected output for a linear system, or the sum of the individual outputs observed for $a(t)$ and $b(t)$ separately.

$$LHS : a(t) + b(t) \to [a(t) \times c(t)] + [b(t) \times c(t)] = [a(t) + b(t)] \times c(t)$$

Since *LHS* = *RHS,* this is a linear system.

**A System Is Defined by Its Response Function** In the above examples, the behavior or response of the system was described by a mathematical function. The response function of a system describes its behavior for all input signals. In the first example, the response function was simply *add 1.5 V to the input signal.* Even if the inner workings of the box (the battery circuit) were unknown, its behavior could nonetheless be described and predicted from knowledge of the response function.

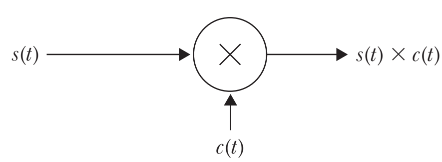



**FIGURE 3.4** Multiplier (Mixer)—The output of this system is the product of the input signal $s(t)$ with the signal $c(t)$.

Indeed, if another box is built with a complicated microprocessor-based acquisition and processing system that simply implements the response function *add 1.5 V to the input signal*, it might be impossible to tell the difference between this deluxe box and the simple battery circuit. When studying signals and systems, the response function is all that is needed to analyze the effect of linear systems on signals.

### 3.2.5 Linear Time-Invariant (LTI) Systems

Ideally, once the behavior or response function of a system is known, it should not behave differently at a later time in response to the same input signal. If the response function of a system changes over time, then it is futile to base predictions of future behavior on an observation of the response function. Mathematically, the concept of *time invariance* says that the system response does not depend on when the input occurs. In other words, the system should behave in an identical way if ever this same signal is introduced at another time. If a linear system is also time invariant, it is called a *linear time-invariant* (LTI) system, otherwise, the system response is said to be *time varying* or *linear time variant* (LTV).

**DEFINITION 3.2 Time Invariance**

*A system defined by the response:*

$$a(t) \rightarrow a_2(t)$$

*is* time invariant *if, for all* $t_0$

$$a(t - t_0) \rightarrow a_2(t - t_0)$$

*where* $t_0$ *is a constant.*

For example, if a fixed-gain linear amplifier simply multiplies an input signal by a constant, the system is linear time invariant as it always multiplies by the same constant at all times. In designing any electronic circuit, it would be reasonable to expect that the values of the circuit components do not change over time (time invariant). In contrast, if a system multiplies an input signal by a time-dependent function such as $\cos(t)$, or $s(t) \rightarrow s(t) \times \cos(t)$, the system is *not* time invariant as can be shown by applying the definition of time invariance, giving the result:

Let $a_2(t)$ be the output for an input signal $a(t)$:

$$a(t) \rightarrow a_2(t) = a(t)\cos(t)$$

Let $a_3(t)$ be the output for a delayed input signal $a(t - t_0)$:

$$a(t - t_0) \rightarrow a_3(t) = a(t - t_0)\cos(t)$$

For time invariance, expect that $a_3(t) = a_2(t - t_0)$:

$$a_2(t - t_0) = a(t - t_0)\cos(t - t_0) \neq a_3(t)$$

Since the $a_3(t) \neq a_2(t - t_0)$ in general, the system is not time invariant.

**DEFINITION 3.3 Linear Time-Invariant System**

*A system that is both linear and time invariant is called a* linear time-invariant *(LTI) system.*

### 3.2.6 Bounded Input, Bounded Output

An important requirement of practical systems is that the output should be *stable*. In other words, the response of a system to some ordinary input is not expected to produce an output signal that grows towards infinity or that unceasingly oscillates with a steadily increasing amplitude.

When the microphone of a public address system is held too close to a loudspeaker or when the output volume is turned too high and can be picked up by the microphone, the penetrating squeal that results is an example of an unstable system. In theory, the output volume would grow to infinity under these conditions. While the actual output sound may get very loud, it is ultimately limited by the audio amplifier circuit (which by that point is outside of its range of linear operation).

An input signal will be called *bounded* if its magnitude is always smaller than some finite value, and the output must be similarly limited if a system is to be stable.

**DEFINITION 3.4 Bounded Signal**

*If* s(t) *is a continuous time signal and*

$$|s(t)| < M \text{ for some } M < \infty$$

*then* s(t) *is a* bounded *signal.*

**DEFINITION 3.5 Bounded Input, Bounded Output (BIBO) Stability**

*If a system is such that every bounded input signal produces a bounded output signal, then the system has bounded-input, bounded-output (BIBO) stability.*

### 3.2.7 System Behavior as a Black Box

A significant observation can now be made. Suppose that the exact response function of the above mixer was unknown and the system was presented as only a black box with only its inputs and outputs available for study. Now, by placing a signal $a(t)$ on the input and by observing the output signal, the unknown response function would reveal itself through the output signal, where $a(t) \rightarrow a(t)\, c(t)$. A second experiment performed using $b(t)$ would reveal that $b(t) \rightarrow b(t)\, c(t)$. Further experimentation will eventually reveal that the secret of this black box, hidden in its response function, is to *multiply the input signal by* $c(t)$. Now, once the response function is known, it describes how the system would behave for any and all input signals.

In practice, it will be shown that a single observation for a known input signal (an impulse) to a linear system can be used to determine the response function uniquely, and thereby to fully describe the behavior of the system for all possible input signals. This is an important observation, and a convincing example of the role that signals theory plays in electrical engineering. Furthermore, LTI systems will form an important class of black box devices, precisely because such predictions are possible. To further study the behavior of a linear system, it is necessary to analyze in more detail the interaction between an input signal and a linear system.

## 3.3 Linear System Response Function $h(t)$

Any specific linear system is characterized by its *response function* usually written as $h(t)$. The exact output from a given linear system at any specific time is a function of the current input signal, the response function, and the state of the system as determined by the history of previous input signals. In particular, the history of previous input signals is especially important in fully describing the output from a linear system. In general, any non-zero input signal has a lingering effect on the output, which must be accounted for as it combines with the same or other input signals. The overall interaction between a signal and the response function is described by the mathematical operation called *convolution*, and a signal is said to be convolved with the response function of a linear system to produce the output signal. To completely describe the interaction between signals and systems requires an understanding of the convolution relationship between a system response function and an input signal. Once this association is established, study and analysis of the system can take place in either the time domain or the frequency domain.

## 3.4 Convolution

The introduction of the convolution integral follows naturally from the superposition property of linear systems. When a signal is fed into a system, the effect of that input may persist for some time, even after the input disappears. For example, striking a bell with a hammer produces a ringing that only gradually fades away. If the bell is struck again before all the echoes have ceased, the new sound will be affected by the fact that the system was not at rest before the second input occurred. The overall response of a system to some input must therefore take into consideration throughout the entire time frame before, during, and after the arrival of a specific input signal. Yet, because of the superposition property, the instantaneous output values are formed from the sum of the individual input responses at that time. This is the underlying principle leading to the convolution integral.

In Figure 3.5, an impulse results in an output signal that goes to 1 V for exactly 2s. When an impulse input arrives at $t = 1$s, the output is high from $t = 1$ to $t = 3$s. When an impulse input arrives at $t = 2$s, the output is high from $t = 2$ to $t = 4$s. When both inputs arrive, the superposition property requires that the output be the sum of the two outputs taken separately as shown. Consequently, the overall effect of the second impulse depends on the lingering effect of the earlier input. In general, multiple or continuous input signals at every point of time must be combined in this way with every other input value at every other point in the past. The final result is the convolution of an input signal with the response function $h(t)$. In the continuous case, the summation operation becomes an integral known as the *convolution integral*.

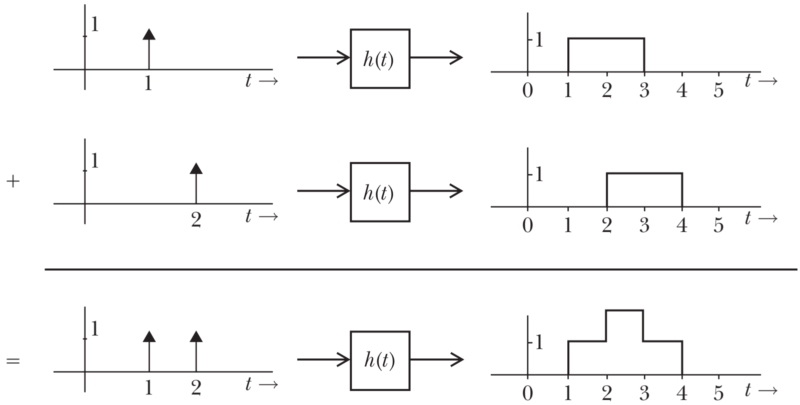



**FIGURE 3.5** Response Function Superposition—The output of a system in response to two impulses is the sum of the outputs for each impulse taken separately. This effect is a consequence of the superposition property that defines linear systems.

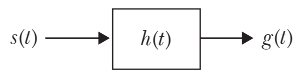



**FIGURE 3.6** Response Function—The input signal $s(t)$ leads to the output signal $g(t)$, after interacting with a linear system characterized by the *response function* $h(t)$.

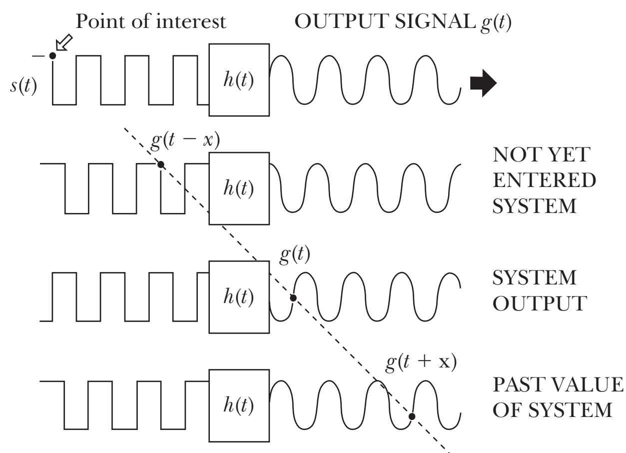



**FIGURE 3.7** Convolution: $g(t) = s(t) * h(t)$.

## 3.4.1 The Convolution Integral

Consider a linear system distinguished by some response function $h(t)$. Let an arbitrary input signal $s(t)$ pass through the linear system, and call the resulting output signal $g(t)$, as shown in Figure 3.6.

Imagine that a specific point on an input signal $s(t)$ can be *tagged* and carefully followed as it proceeds through the linear system. In Figure 3.7, the variable $x$ is used to plot the progress of the tagged point at three different points in its journey through the system. At the moment the signal emerges from the linear system, it may be labelled $g(t)$. Relative to this moment in time, the point $g(t-x)$ is some time before the point of interest entered the system. Sometime later, the same point labelled $g(t+x)$ no longer contributes.

To describe the output signal $g(t)$ for every point in the signal $s(t)$, every point in $s(t)$ may be tagged and followed on its path through the system. This requires that the output be computed as many times as there are points in $s(t)$. Furthermore, it can be observed that it should make no difference, mathematically, whether the signal passes through the system's response function or if the reference point is changed such that system's response function is shifted back through the input signal as in Figure 3.8. With an appropriate change of variables, the previous diagram has been rearranged to reflect the desired relationship between the input and output signals. In Figure 3.9, the output signal $g(t)$ is the convolution of the input $s(t)$ with the linear system's response function $h(t)$.

The above operation gives the system output $g(t)$ at a specific time $t$. By varying $t$, the response function $h(t-x)$ is mathematically shifted back across the input signal, and the product $s(x)\,h(t-x)$ describes the contribution of each tagged point $s(x)$ in the input signal to the eventual value of $g(t)$. To incorporate the effect of all previous inputs on the current output signal, the summation of all these contributions,

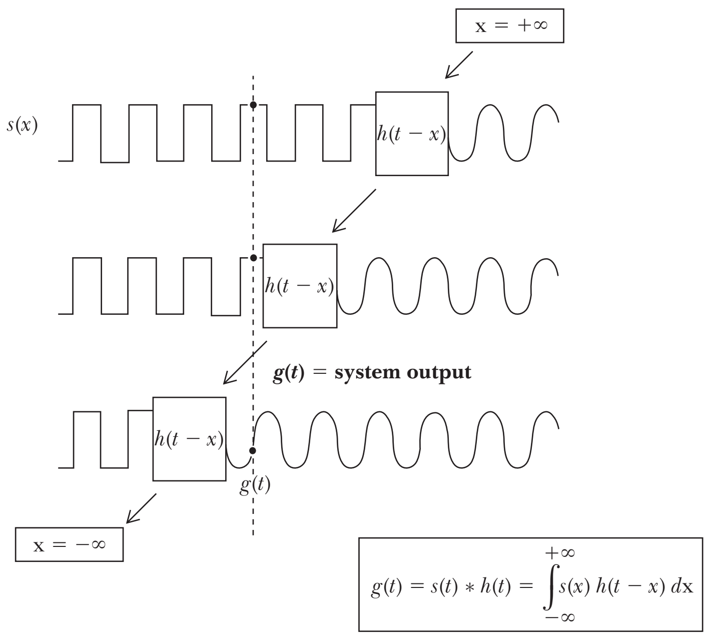



**FIGURE 3.8** Convolution: $g(t) = s(t) * h(t)$.

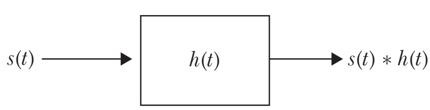



**FIGURE 3.9** Linear System Modelling—The output signal is the *convolution* of the input signal $s(t)$ and the system's response function $h(t)$.

in the limit, can be described by an integral over the variable $x$. This operation defines the signal $g(t)$ as the mathematical convolution of $s(t)$ and $h(t)$. The shorthand notation $g(t) = s(t) * h(t)$ is used to describe the resulting convolution integral.[2]

**DEFINITION 3.6** **Convolution**

*Let* s(t) *and* g(t) *be two continuous signals, then*

$$s(t) * h(t) = \int_{-\infty}^{+\infty} s(x)\, h(t-x)\, dx$$

*is called the* convolution *of* s(t) *and* g(t).

## 3.4.2 Convolution Is Commutative

By inspection, it should not matter which of the two signals is shifted across the other in the computation of the convolution integral. It can be readily shown that the order of the signals $s(t)$ and $h(t)$ does not affect the convolution result.

**THEOREM 3.1** *(Commutative)*

*If* g(t) *is the convolution of two continuous signals* s(t) *and* h(t),

$$g(t) = s(t) * h(t)$$

*then*

$$g(t) = h(t) * s(t)$$

**Proof:**
Consider the convolution $g(t) = s(t) * h(t)$ from Definition 3.6. Let $y = t - x$, then $x = t - y$, $dx = -dy$, and the limits also change sign, leaving the expected result.

$$\begin{aligned} s(t) * h(t) &= \int_{-\infty}^{+\infty} s(x)\, h(t-x)\, dx \\ &= -\int_{+\infty}^{-\infty} s(t-y)\, h(y)\, dy \\ &= +\int_{-\infty}^{+\infty} s(t-y)\, h(y)\, dy \\ &= h(t) * s(t) \end{aligned}$$

Therefore, the convolution operation is commutative.

[2] The convolution integral term $h(t-x)$ specifically includes an important minus sign, which distinguishes it from a similar operation called *correlation* in which the term $h(t+x)$ appears.

### 3.4.3 Convolution Is Associative

Provided that the convolution integrals exist, convolution is *associative*, or:

**THEOREM 3.2** *(Associative)*

*If* a(t) *and* b(t) *and* c(t) *are continuous signals, then*

$$[a(t) * b(t)] * c(t) = a(t) * [b(t) * c(t)]$$

**Proof:**
Consider:

$$a(t) * b(t) = \int_{-\infty}^{+\infty} a(x)\, b(t-x)\, dx$$

therefore

$$[a(t) * b(t)] * c(t) = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} a(x)\, b(y-x)\, dx\right] c(t-y)\, dy$$

which simplifies to:

$$LHS = [a(t) * b(t)] * c(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a(x)\, b(y-x)\, c(t-y)\, dx\, dy$$

Also:

$$b(t) * c(t) = \int_{-\infty}^{+\infty} b(y)\, c(t-y)\, dy$$

therefore

$$a(t) * [b(t) * c(t)] = \int_{-\infty}^{+\infty} a(x)\left[\int_{-\infty}^{+\infty} b(y)\, c(t-y-x)\, dy\right] dx$$

Now, let $z = x + y$ in the square brackets, then $y = z - x$ and $dy = dz$ and the limits do not change.

$$a(t) * [b(t) * c(t)] = \int_{-\infty}^{+\infty} a(x)\left[\int_{-\infty}^{+\infty} b(z-x)\, c(t-z)\, dz\right] dx$$

which simplifies to:

$$RHS = a(t) * [b(t) * c(t)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} a(x)\, b(z-x)\, c(t-z)\, dz\, dx$$

Therefore (with a difference in labelling), *RHS* = *LHS*, and the theorem is proved.

### 3.4.4 Convolution Is Distributive over Addition

Consistent with the definition of a linear system, the convolution operation obeys superposition; that is, the convolution of a signal with the sum of two signals is given by the sum of the convolutions of the individual signals. This is a statement of the distributive property wherein the signals $a(t) + b(t)$ are convolved with $h(t)$.

**THEOREM 3.3** *(Distributive)*

*If* a(t) *and* b(t) and h(t) *are continuous signals, then*

$$[a(t) + b(t)] * h(t) = [a(t) * h(t)] + [b(t) * h(t)]$$

**Proof:**
Consider the definition of $a(t) * h(t)$ and $b(t) * h(t)$ and examine their sum $[a(t) + b(t)]$:

$$\begin{aligned}
[a(t) + b(t)] * h(t)] &= \int_{-\infty}^{+\infty} [a(x) + b(x)]\ h(t-x)\ dx \\
&= \int_{-\infty}^{+\infty} a(x)\ h(t-x)\ dx + \int_{-\infty}^{+\infty} b(x)\ h(t-x)\ dx \\
&= [a(t) * h(t)] + [b(t) * h(t)]
\end{aligned}$$

Therefore, convolution is distributive over addition.

### 3.4.5 Evaluation of the Convolution Integral

In general, the convolution integral is not easily evaluated because for a given value $t$, computing the integral over all $x$ yields only a single value of $g(t)$. For example, consider the value of $g(t) = s(t) * r(t)$, at the specific times $t = \{1, 2, 3\}$.

$$\begin{aligned}
g(1) &= \int_{-\infty}^{+\infty} s(x) r(1-x) dx \\
g(2) &= \int_{-\infty}^{+\infty} s(x) r(2-x) dx \\
g(3) &= \int_{-\infty}^{+\infty} s(x) r(3-x) dx \\
&\cdots \\
g(t) &= \int_{-\infty}^{+\infty} s(x) r(t-x) dx
\end{aligned}$$

This entire integral must be recomputed at every time $t$ to obtain a complete description of the output signal $g(t)$. Each integral gives the area under the product $s(x)r(t-x)$ only for one $t$. The function $r(t-x)$ effectively shifts to a new starting position for each value of $t$. Consequently, the convolution operation is notoriously time-consuming even when performed by computer. In practice, the equivalent operation will be more easily accomplished in the frequency domain, and a much more efficient approach to convolution using the Fourier transform will be found in Chapter 6. The following exercises serve to illustrate the meaning of the convolution integral, as well as to demonstrate how its computation may be accomplished. The same exercises will then be accomplished using MATLAB.

**Graphical Exercise 1: Convolution of a Rectangle with Itself** Consider the rectangle $s(t)$ shown in Figure 3.10. This rectangle has unit height and width = 2. The convolution $g(t) = s(t) * s(t)$ will be computed by hand.

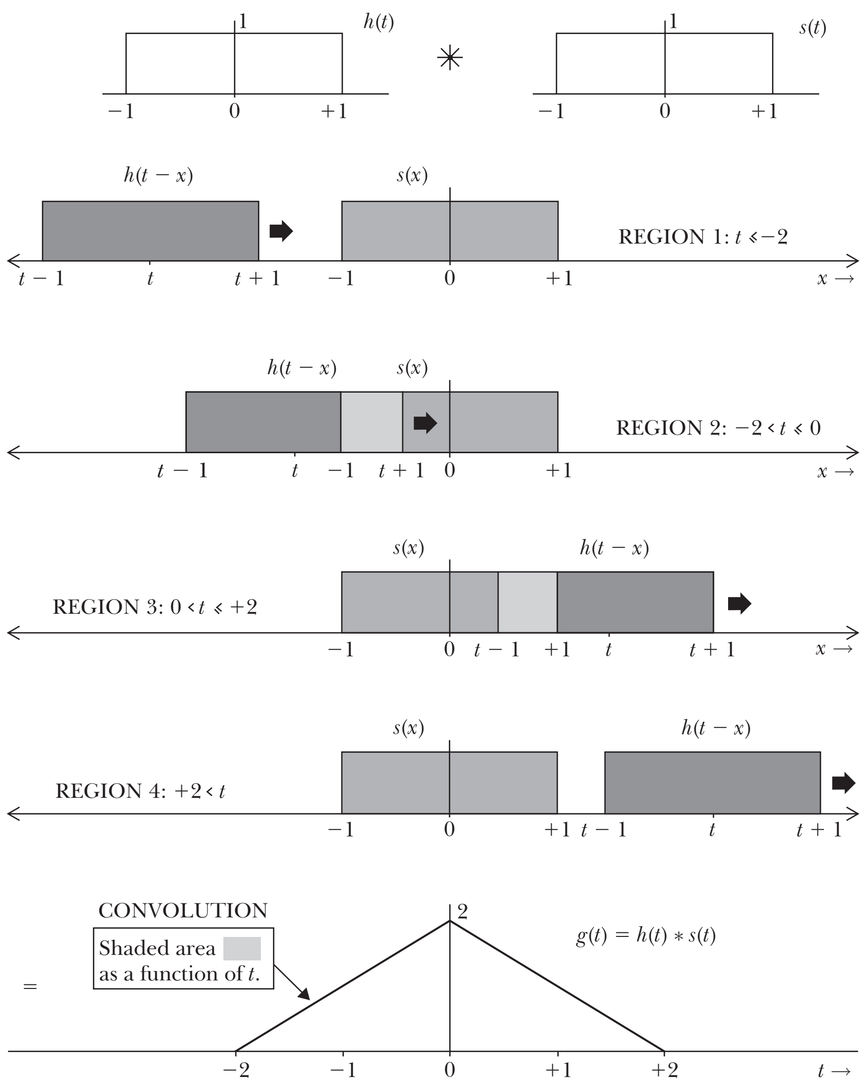



**FIGURE 3.10** The convolution of a rectangle with itself is accomplished piecewise and leads to a result with a distinctive triangular shape having twice the rectangle width and the square of the rectangle area.

**TABLE 3.1**

**The closed form solution $g(t)$ of the convolution integral within each region is obtained by inspection from Figure 3.10.**

| Region | Interval $[t_0, t_1]$ | $g(t) = \int_{-\infty}^{+\infty} s(x)\, h(t-x)\, dx$ |
|---|---|---|
| 1 | $[-\infty, -2]$ | 0 |
| 2 | $[-2, 0]$ | $2+t$ |
| 3 | $[0, +2]$ | $2-t$ |
| 4 | $[+2, +\infty]$ | 0 |

In a problem such as this, it is very useful to begin by sketching the situation. For clarity, let $h(t) = s(t)$ and compute $g(t) = h(t) * s(t)$. The two rectangles are shown in Figure 3.10. For each value of $g(t)$, the integral of the product $s(x)\ h(t-x)$ must be computed. For each value of $t$, the rectangle $h(t-x)$ effectively shifts to a new position. By inspection, for many of these positions, the above product term [and therefore $g(t)$] will equal zero.

By examining the two rectangles and the product $s(x)\ h(t-x)$, four distinct regions of interest may be identified as regions 1 through 4 on Figure 3.10. These regions are established by considering the appearance of the two signals for increasing $t$ as one rectangle moves to the right across the other:

1. The rectangles do not overlap and $s(x)\ h(t-x)$ is zero.
2. The rectangles overlap and the form of $s(x)\ h(t-x)$ is the same until they completely overlap.
3. The rectangles overlap and the form of $s(x)\ h(t-x)$ is the same until they no longer overlap.
4. The rectangles do not overlap and $s(x)\ h(t-x)$ is zero.

The corresponding closed form integral solutions are summarized in Table 3.1. Without identifying separate regions, an infinity of integrations must be performed. By identifying regions sharing a common product term $s(x)h(t-x)$, the problem becomes manageable as only one integral within each region must be solved over an interval $[t_0, t_1]$. Within each region, the convolution $g(t)$ is the integral of the overlapping signals (shown shaded) as a function of time.

These results describe a piecewise $g(t)$ over the four regions that gives the form of a triangle as shown in Figure 3.10. In conclusion, the convolution of a rectangle with itself is a triangle with twice the width of the rectangle and area equal to the area of the rectangle squared.

## 3.4.6 Convolution Properties

Two observations can be generalized for the convolution $g(t) = s(t) * h(t)$:

1. The area of $g(t)$ equals the product of the areas of $s(t)$ and $h(t)$.
2. The width of $g(t)$ equals the sum of the widths of $s(t)$ and $h(t)$.

**Graphical Exercise 2: Convolution of Two Rectangles** In this section, the previous example is repeated with one rectangle equal to half the width of the other.

Let $h(t)$ be a unit rectangle (Area = 1, Width = 1), then $s(t)$ has Area = 2, Width = 2.

Figure 3.11 shows that there are five regions of interest as shown in Table 3.2, since the smaller rectangle fits within the wider one as it shifts across. Throughout this interval of complete overlap, the shaded area is constant. In contrast to the previous result, this answer should be a trapezoid rather than a triangle. Applying the rules for convolution, it can be predicted that the area of the result will equal $1 \times 2 = 2$, and the total width of the result will be $1 + 2 = 3$. The result is shown with reference to the five regions in Table 3.2.

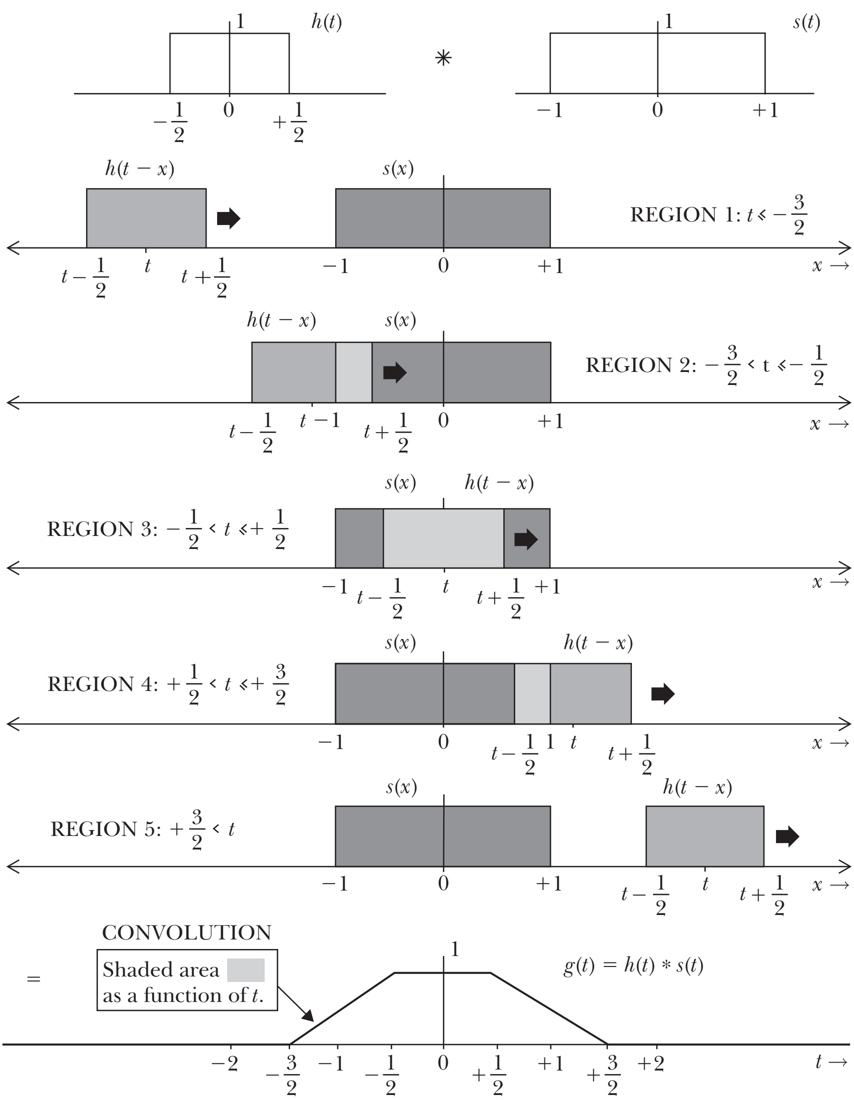



**FIGURE 3.11** The convolution of two different rectangles is accomplished piecewise and leads to a result with a distinctive trapezoidal shape having the sum of the rectangle widths and the product of the rectangle areas.

**TABLE 3.2**

**The closed form solution $g(t)$ of the convolution integral within each region is obtained by inspection from Figure 3.11.**

| Region | Interval $[t_0, t_1]$ | $g(t) = \int_{-\infty}^{+\infty} s(x)\,h(t-x)\,dx$ |
|---|---|---|
| 1 | $[-\infty, -3/2]$ | 0 |
| 2 | $[-3/2, -1/2]$ | $3/2 + t$ |
| 3 | $[-1/2, +1/2]$ | 1 |
| 4 | $[+1/2, +3/2]$ | $3/2 - t$ |
| 5 | $[+3/2, +\infty]$ | 0 |

This result is sketched on Figure 3.11. It can be seen that the width of $g(t)$ equals 3, and its area equals 2, as predicted. On the other hand, this example shows that this systematic graphical approach to solving the convolution equation will really only be effective for relatively simple signals. The next example shows a practical application and an important point to remember.

**Graphical Exercise 3: Convolution of a Rectangle and an Exponential Decay.** Consider a linear system having the response function $h(t) = e^{-at}\,u(t)$, for $a > 1$. This signal is zero for $t < 0$. This linear system might be a simple circuit consisting of an ideal resistor and capacitor as shown, and $h(t)$ can be sketched as shown in Figure 3.12. Assume that the capacitor is initially uncharged.

By definition, the function $h(t)$ is the zero-state response of the above system. From the resting state $(t = 0)$, it can be imagined that an impulse input would suddenly charge the capacitor, after which the exponential curve describes the characteristic voltage decay as the capacitor discharges at a rate determined by the RC time constant of this circuit. The response function $h(t)$ can now be used to describe the behavior of the RC circuit for various inputs.

**A Pulse Input Signal** The output signal $g(t)$ will be found for a rectangle $s(t)$ input to this circuit.

The output signal $g(t)$ is the convolution of the input signal $s(t)$ with the response function $h(t)$, as shown in Figure 3.13. Before proceeding, it is important to recall that the signal $h(-x)$ is the mirror image of the signal $h(x)$. This implies that the sketch of $h(x)$ must be *flipped around* before the convolution regions are defined graphically. This operation is of no consequence for even signals such as the rectangle used in the previous examples. Three regions are defined as shown in Table 3.3.

The resulting output signal follows what would be expected of the *charging-discharging* cycle of an RC circuit.

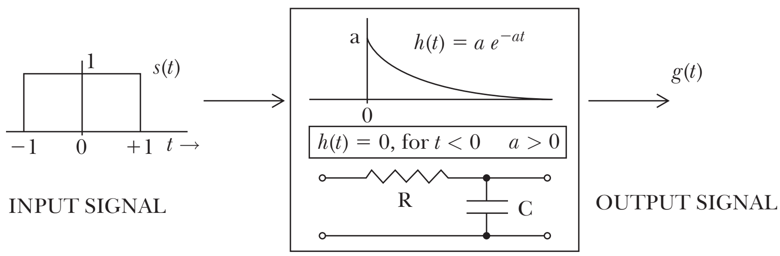



**FIGURE 3.12** The impulse response of an RC circuit intuitively follows the expected circuit behavior. The exact form is a decaying exponential function.

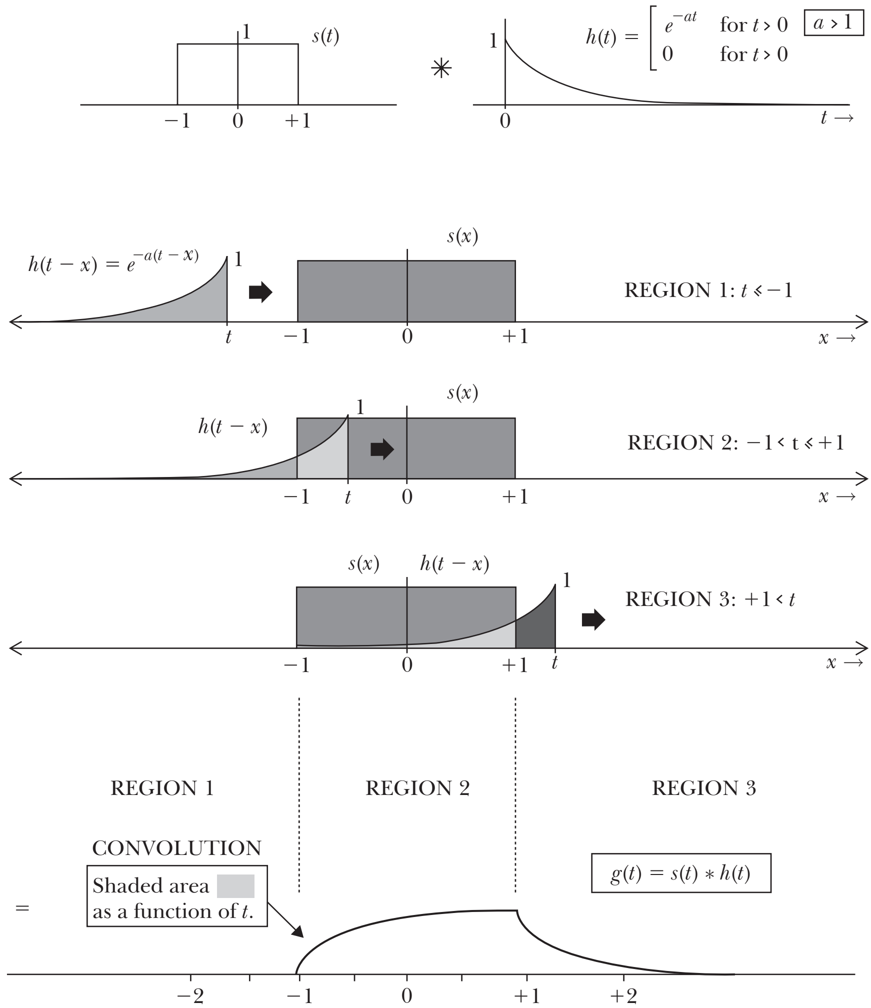



**FIGURE 3.13** The RC circuit behavior is contained in the impulse response function $h(t)$. The convolution of an input signal with $h(t)$ gives the output signal. The time reversed $h(-t)$ must be used when computing the convolution.

**TABLE 3.3**

**The closed form solution $g(t)$ of the convolution integral within each region is obtained by inspection from Figure 3.13.**

| Region | Interval $[t_0, t_1]$ | $g(t) = \int_{-\infty}^{+\infty} s(x)\, h(t-x)\, dx$ |
|---|---|---|
| 1 | $[-\infty, -1]$ | 0 |
| 2 | $[-1, +1]$ | $\frac{1}{a}[1 - e^{-a(t+1)}]$ |
| 3 | $[+1, +\infty]$ | $\frac{1}{a}[e^{-a(t-1)} - e^{-a(t+1)}]$ |

## 3.4.7 Convolution with MATLAB

The previous examples show that graphical convolution has its limitations and can be time-consuming. Fortunately, numerical methods can also be used for this important operation. The following MATLAB exercises confirm the previous observations and results calculated manually.

- Convolution in MATLAB may be accomplished through the operator `conv(a,b)`, where $a$ and $b$ are the vectors to be convolved. The result must be scaled by the defined time increment. The output array has a number of elements determined by the sum of the number of elements in $a$ and $b$.
- Numerical integration in MATLAB may be accomplished through the operator `trapz(a)`, where $a$ is the vector to be integrated over a defined time interval. The result must be scaled by the defined time increment. If $t$ defines a known time interval, then `trapz(t,a)` may be used with no scaling required.

**MATLAB Exercise 1: Convolution of a Rectangle with Itself** Let the signal $a(t)$ be a unit rectangle defined on a time axis 4 s wide using a step size of 0.01. The convolution of $a(t)$ with itself will be performed as shown in Figure 3.14.
**Check:** The result of a convolution has a width equal to sum of the widths of the input signals.

```
t = -2:0.01:2;          % step size = 0.01
a = rectpuls(t);        % define a unit rectangle
plot(t,a);
axis([-2 2 -1 2 ]);     % display interval = 4s
c = conv(a,a)*0.01;     % result scaled by step size
```

Because $a(t)$ is defined on an interval 4 s wide, then `c=conv(a,a);` spans an interval 8 s wide with the same time increment.

```
t = -4:0.01:4;           % define new time interval
plot(t,c);
axis([-2 2 -1 2 ]);      % display interval = 4s
```

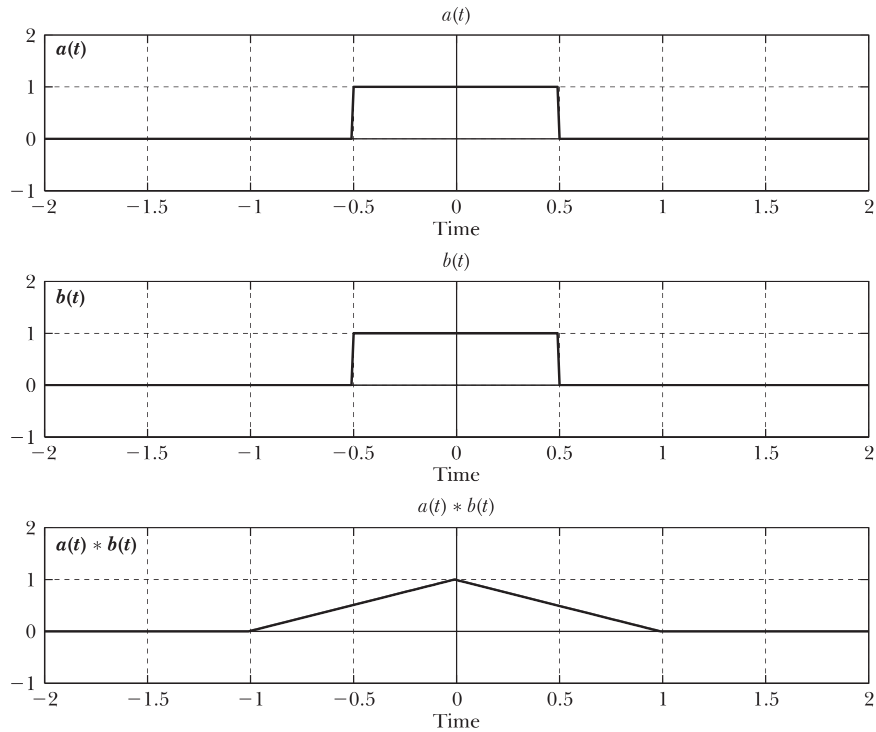



**FIGURE 3.14** The convolution of a unit rectangle with itself gives a triangle shape with Area = 1 and Width = 2. Compare to Figure 3.10.

**MATLAB Exercise 2: Convolution of Two Rectangles** Let the signal $a(t)$ be a unit-height rectangle with width 4, and the signal $b(t)$ be a unit-height rectangle with width 7. The convolution of the signals $a(t)$ and $b(t)$ is to be performed. The width and area of the resulting signal $c(t)$ will be calculated.

**Check:** It is expected that the $c(t)$ has $area = 4 \times 7 = 28$, and $width = 4 + 7 = 11$.

A rectangle $N$ units wide may be defined by scaling the unit rectangle as `rectpuls(t/N)` to stretch the unit rectangle in time.

```
t = −10:0.01:10;          % define a time interval
a = rectpuls(t/4);        % unit height rectangle, width=4
b = rectpuls(t/7);        % unit height rectangle, width=7
c = conv(a,b)*0.01;       % convolution

area = trapz(t,c)         % integration
area =
      28
```

This confirms the expected area (28) under the convolution. The expected width (14) is confirmed in Figure 3.15.

```
t = −20:0.01:20;           % define new time interval
plot(t,c);
axis([−10 10 −1 8 ]);      % display interval = 20s
```

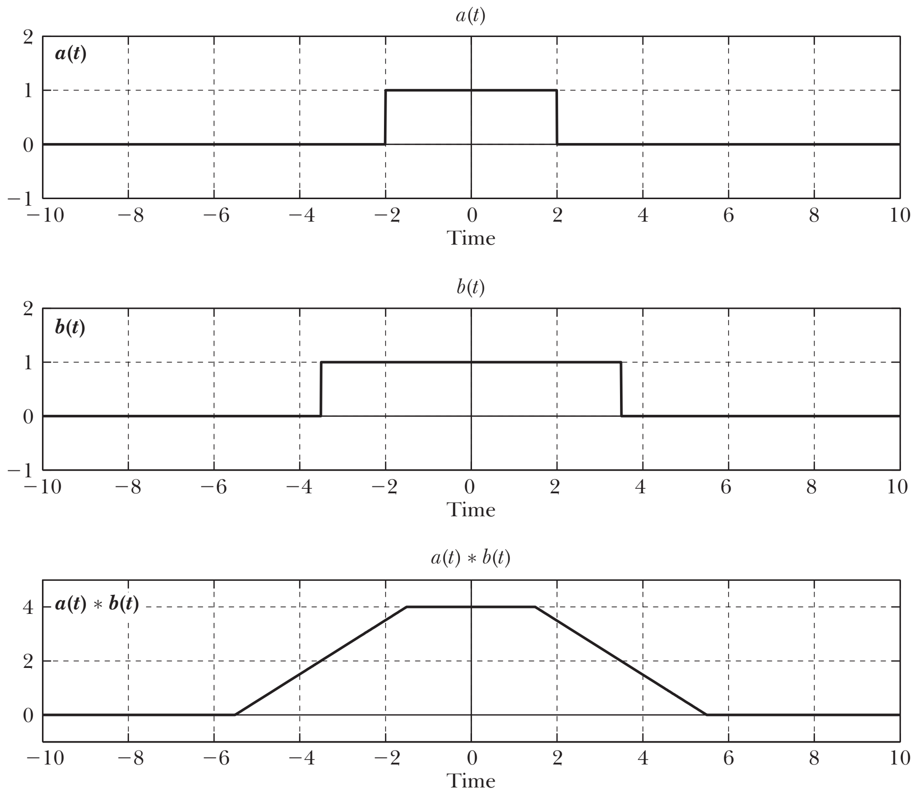



**FIGURE 3.15** The convolution of these two different width rectangles gives a trapezoidal shape with area = 28 and width = 11. Compare to Figure 3.11.

**MATLAB Exercise 3: Convolution of a Rectangle with an Exponential Decay** Let the signal $a(t)$ be a unit rectangle defined on a time axis 4 s wide and $b(t)$ be an exponential decay $b(t) = e^{-2t}$, defined for $t > 0$. The convolution of the signals $a(t)$ and $b(t)$ is to be performed, and the area of the resulting signal $c(t)$ will be calculated.

**Check:** It is expected that the $c(t)$ should have a total area equal to the area under $b(t)$ since the rectangle $a(t)$ has unit area.

```
t = −1:0.01:3;                      % define time interval
a = rectpuls(t);                    % define unit rectangle
b = exp(−2*t) .* (t>0);             % exponential for t>0
c = conv(a,b)*0.01;                 % convolution

t = −2:0.01:6;                      % define new time interval
plot(t,c);                          % plot the result
axis([−1 3 −0.5 1 ]);               % set display interval

LHS = trapz(t,b)                    % check the area under b
LHS=
     0.4938

RHS = trapz(t,c)                    % check the area under c
RHS=
     0.4938
```

Since *RHS* = *LHS,* the expected area result is confirmed. The convolution result $c(t)$ is shown in Figure 3.16.

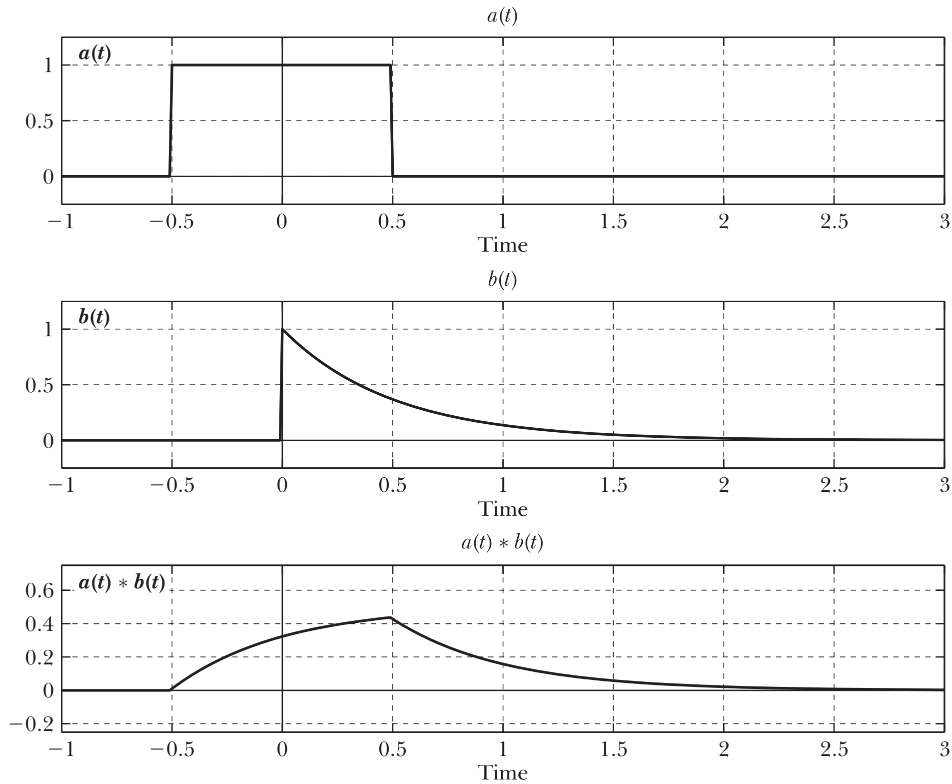



**FIGURE 3.16** The convolution of a unit rectangle with this exponential decay gives a distinctive shape modelling the effect of passing a pulse through a first-order lowpass filter (RC circuit). Compare to Figure 3.13.

## EXAMPLE 3.1 (Integration by Convolution)

*Show that convolution by a step function* u(t) *is equivalent to time integration, or*

$$a(t) * u(t) = \int_{-\infty}^{t} a(x)\, dx$$

**Solution:**

The convolution of a signal $a(t)$ by a step function $u(t)$ is given by:

$$a(t) * u(t) = \int_{-\infty}^{+\infty} a(x)\, u(t-x)\, dx$$

or, since $u(t-x) = 0$ for all $x > t$, and $u(t-x) = 1$ otherwise:

$$a(t) * u(t) = \int_{-\infty}^{t} a(x)\, dx$$

which is the *integral operator* describing the time integral of the function $a(t)$. This result is shown in Figure 3.17 for $a(t) = rect(t)$.

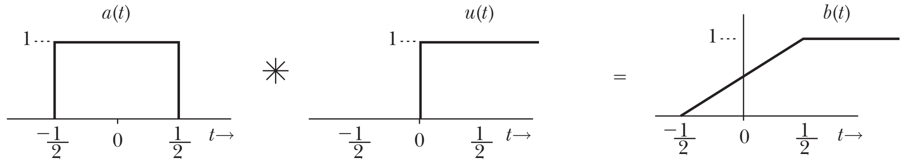



**FIGURE 3.17** Convolution by a unit step function $u(t)$ is equivalent to time integration. The signal $b(t)$ is both the time integral of $a(t)$ and the convolution of $a(t)$ with $u(t)$.

## 3.5 Determining $h(t)$ in an Unknown System

The convolution integral dictates the behavior of a linear system when some input signal $s(t)$ is applied. A specific linear system is characterized by a distinctive response function $h(t)$. As was stated previously, an unknown (black box) linear system would be fully understood if its hidden response function $h(t)$ could be determined. Short of breaking open the box to see the hidden circuit, one means of determining $h(t)$ would be to input some test signal $s(t)$ and then to observe the output $g(t)$. With enough test signals, it would be possible to work backwards through the mathematics of convolution to determine $h(t)$. Fortunately, there is one special test signal that yields $h(t)$ directly.

### 3.5.1 The Unit Impulse $\delta(t)$ Test Signal

Let the input signal to a linear system be a unit impulse, $s(t) = \delta(t)$. The output signal is $g(t) = h(t) * \delta(t)$, where:

$$g(t) = \int_{-\infty}^{+\infty} h(x)\,\delta(t-x)\,dx$$

Recall that an integral including a product with an impulse function $\delta(x)$ can be solved directly and consequently $g(t) = h(t) * \delta(t) = h(t)$.

In other words, when $\delta(t)$ is input to an LTI system, the response function $h(t)$ emerges. No other tests need to be performed to determine $h(t)$. For this reason, the response function $h(t)$ is often referred to as the *impulse response* of a linear system, since $h(t)$ is effectively the output or response of a linear system to an impulse input signal. In general, this result is:

**THEOREM 3.4** *[Convolution with $\delta(t)$]*

*If* g(t) *is the convolution of* s(t) *with* δ(t) *then*

$$g(t) = s(t) * \delta(t) = s(t)$$

In summary, the convolution of a signal $s(t)$ with an impulse $\delta(t)$ equals the same signal $s(t)$. It follows that, by shifting the impulse to $\delta(t - t_0)$, a copy of $s(t)$ can be placed anywhere in time, as shown in Figure 3.18, or:

$$s(t) * \delta(t - t_0) = s(t - t_0) \tag{3.1}$$

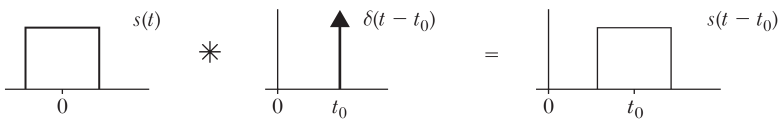



**FIGURE 3.18** The convolution of a signal $s(t)$ with an impulse located at $t = t_0$ gives a copy of $s(t)$ located at $t = t_0$.

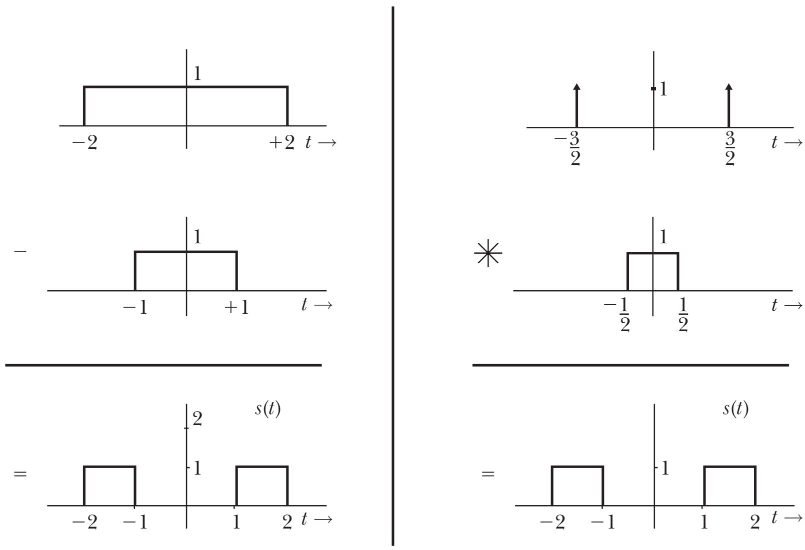



**FIGURE 3.19** This even pair of pulses $s(t)$ may be viewed as the difference between two rectangles or, alternatively, as the convolution of a unit rectangle with a pair of impulses.

## 3.5.2 Convolution and Signal Decomposition

The property that $s(t) * \delta(t - t_0) = s(t - t_0)$ is also very useful in constructing signals from basic signal types. In Chapter 1, new and interesting signals were built by scaling, shifting, adding, and multiplying other signals. Convolution can now be added to the available choice of signal manipulations.

Consider the signal $s(t)$ expressed as two shifted unit rectangles as in Figure 3.19.

$$s(t) = \text{rect}(t - 3/2) + \text{rect}(t + 3/2)$$

This signal was shown to be equivalent to the difference between a rectangle of width 4 and another of width 2, or:

$$s(t) = \text{rect}(t/4) - \text{rect}(t/2)$$

The same signal can now be modelled as a unit rectangle convolved with unit impulses located at $t = +3/2$ and $t = -3/2$, as shown on the righthand side of Figure 3.19.

$$\begin{aligned} s(t) &= [\delta(t + 3/2) + \delta(t - 3/2)] * \text{rect}(t) \\ &= \delta(t + 3/2) * \text{rect}(t) + \delta(t - 3/2) * \text{rect}(t) \\ &= \text{rect}(t - 3/2) + \text{rect}(t + 3/2) \end{aligned}$$

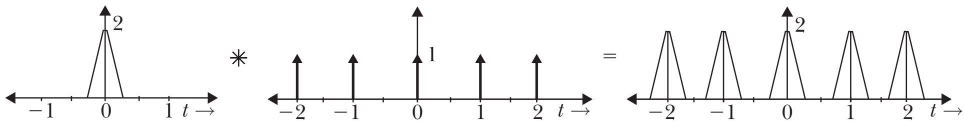



**FIGURE 3.20** A periodic signal can be modelled as the convolution of the single occurrence of a given shape with an impulse train.

The convolution leads to two copies of the unit rectangle, one centered at $t = -3/2$ and the other at $t = +3/2$. Graphically, convolution with an impulse is trivial to sketch, since only copies of the original signal need to be drawn. The convolution approach will serve in many situations where a signal must be expressed in terms of some component signals. Moreover, this example illustrates that there are often many different ways to describe a given signal. Other ways of expressing $s(t)$ were shown in Chapter 1. Each is correct, yet some may prove to be more useful or easier to work with in certain situations.

**Convolution and Periodic Signals** By extension, any periodic signal could be modelled as the single occurrence of a given shape convolved with a series of unit impulses spaced at equal intervals along the time axis (impulse train) as seen in Figure 3.20.

## 3.5.3 An Ideal Distortionless System

If a system is *distortionless*, then an impulse entering the system will emerge as an impulse, possibly changed in size, and possibly delayed in time. The impulse response $h(t)$ characterizing a distortionless system is defined as:

**DEFINITION 3.7 Distortionless System**

*If* h(t) *is the response function of an LTI system and*

$$h(t) = k\,\delta(t - t_0)$$

*for constant* k *and* $t_0 \geq 0$, *then* h(t) *defines a* distortionless system.

The output from such a system for any input $s(t)$ would be a copy of $s(t)$ delayed by time $t_0$ and amplified or attenuated by the constant $k$. This system response is illustrated in Figure 3.21; for the impulse response $h(t) = \delta(t - t_0)$, the input signal $s(t)$ emerges undistorted, although with a new amplitude and delayed by $t_0$ s.

**Deconvolution** In some applications such as image processing, the *reverse convolution* or *deconvolution* is attempted. This procedure involves going backwards through a linear system, starting with the output signal to find what the input signal used to be. For example, it should be possible to correct out-of-focus images by reversing the effect of light passing through a distorting lens. Deconvolution is a challenging task that is often not fully successful. The problem is complicated by the fact that random noise added to the output signal confounds the reverse operation. Unfortunately, its implications are beyond the scope of this discussion.

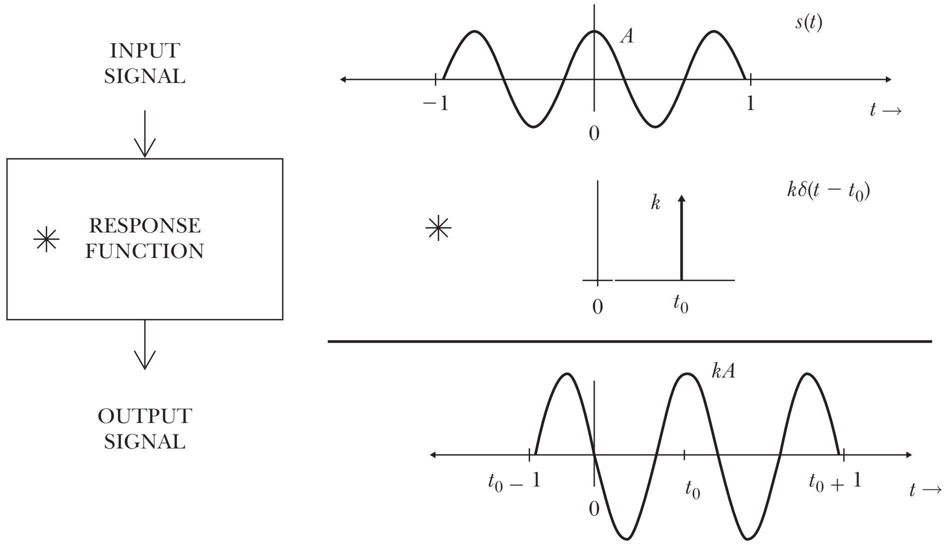



**FIGURE 3.21** An *ideal distortionless linear system* has the response function $h(t) = k\delta(t - t_0)$. The output is a copy of the input signal delayed by $t_0$ s and with a new amplitude determined by the constant $k$.

# 3.6 Causality

The examples here illustrate mathematically the operation of the convolution integral. A given system is defined by its impulse response function $h(t)$, which is convolved with any input signal. Mathematically, systems can be described that defy construction. While it is convenient to speak of *ideal systems*, $h(t)$ must nonetheless obey certain general conditions if a system is to exist beyond the pages of a book. The condition that $h(t)$ must describe a *causal* system is presented here.

The impulse response of a system $h(t)$ describes how the system responds to an impulse input occurring at time $t = 0$. The principle of causality is a statement that *the response function responds.* Any output must be caused by the input, or the output happens only after the input arrives. It is illogical to expect a system to respond to an impulse before the input arrives. In other words, the impulse response must be zero for all for $t < 0$. The response function $h(t)$ of a causal system is defined by the fact that $h(t)$ is zero for all $t < 0$, as shown in Figure 3.22.

**DEFINITION 3.8 Causal System**

*If* $\mathrm{h(t)}$ *is the impulse response of an LTI system, and* $\mathrm{h(t)} = 0$ *for all* $\mathrm{t} < 0$*, then* $\mathrm{h(t)}$ *is a* causal impulse response *describing a* causal system.

Any system that violates this rule must be capable of anticipating the arrival of an impulse at some future time, which is clearly not possible in ordinary circumstances.

The first-order RC lowpass filter has the exponential decay curve impulse function shown in Figure 3.23. This response function satisfies the condition for causality, and this is obviously a filter that can be constructed. It is a causal response function.

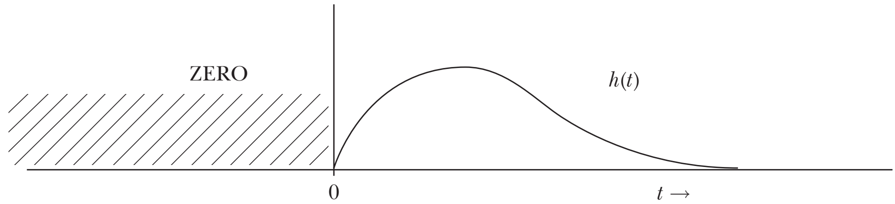



**FIGURE 3.22** A *causal* impulse response $h(t)$ is zero for $t < 0$. This implies that the input to a linear system at time $t = 0$ causes the output.

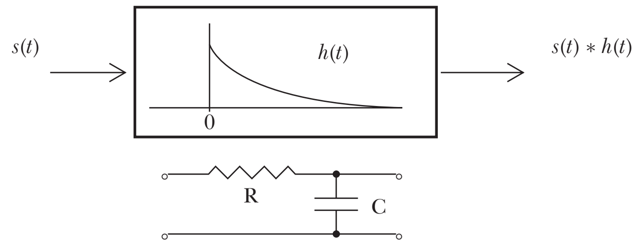



**FIGURE 3.23** A real-world causal impulse response (RC circuit). The system cannot respond until an input arrives: $h(t) = 0$ for $t < 0$.

### 3.6.1 Causality and Zero Input Response

The presence of initial conditions in a system is closely related to the concept of causality. Just as the output of a causal system depends only on past and present events, it is possible that the initial input to a system at time $t=0$ may be affected by the fact that certain circuit components are not at rest. Unless the initial conditions are zero, this zero input component must be considered in the response function. For example, a mechanical system may have a stretched spring, or an electrical system may have a charged capacitor that will influence the effect of an arriving signal. It can be argued that such occurrences are simply the result of past events, but, even if the events are forgotten, any lingering effects must be incorporated as *initial conditions* in any new calculation.

If the capacitor in the RC circuit of Figure 3.23 holds some non-zero charge prior to the arrival of an input signal at time $t = 0$, then the outcome will include both the expected output for the input alone (zero state response) plus the contribution from the charged component (zero input response). This result is consistent with the superposition property that defines linear circuits. It is as if the past contributions from all previous inputs are condensed into the initial conditions found at time $t = 0$. In practice, the behavior of inductors and capacitors is described by differential equations that themselves incorporate initial conditions into their solutions.

## 3.7 Combined Systems

It would be rare for a signal to pass through only one linear system, and it may be convenient to express a system in terms of component systems. For example, if $a(t)$ and $b(t)$ are added to enter a system described by $h(t)$, the superposition property

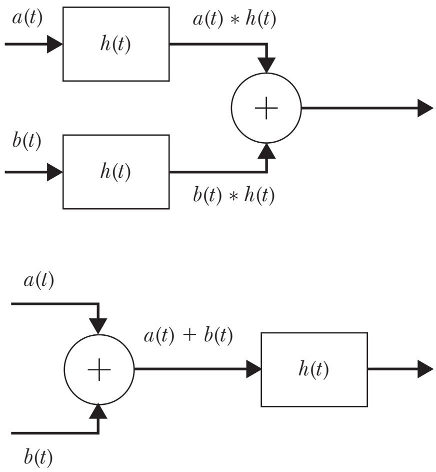



**FIGURE 3.24** Two equivalent linear systems: The input signals $a(t)$ and $b(t)$ emerge as $a(t) * h(t) + b(t) * h(t)$ in either case.

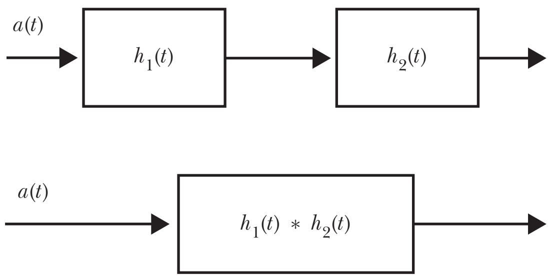



**FIGURE 3.25** Two equivalent linear systems: The input signal $a(t)$ emerges as $a(t) * h_1(t) * h_2(t)$ in either case.

immediately suggests that the system could be redrawn as if they each passed through separately, as shown in Figure 3.24.

Similarly, if a signal $a(t)$ passes through two linear systems in series, the overall effect may be modelled as a single system with response function $h_1(t) * h_2(t)$ where the order of evaluating the convolution integral does not matter, as shown in Figure 3.25.

**MATLAB Exercise 4: Systems in Series** Two linear systems have response functions $h_1(t)$ and $h_2(t)$, respectively, where:

$$h_1(t) = \begin{cases} e^{-2t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h_2(t) = \begin{cases} e^{-3t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

A signal $s(t)$ passes through the first system, and the output passes through the second system. The overall response $h_1(t) * h_2(t)$ will be performed using MATLAB.

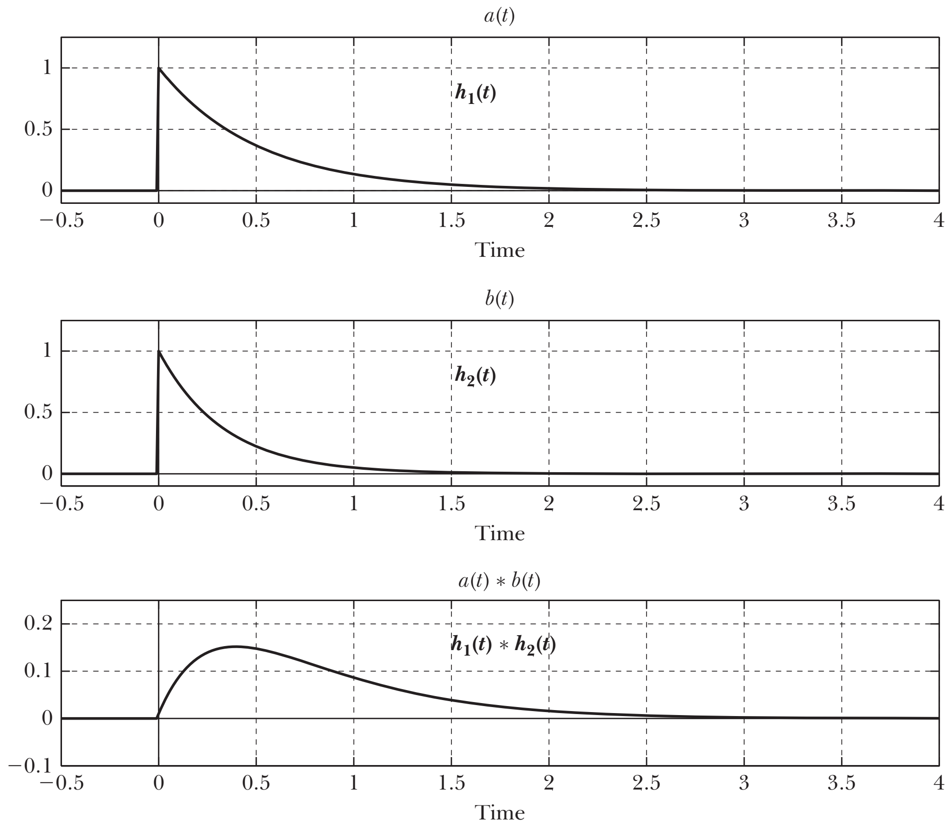



**FIGURE 3.26** The overall response function of two linear systems connected in series is given by the convolution of $h_1(t)$ and $h_2(t)$.

```
t = 0:0.01:10;              % time interval = 10s
h1 = exp(-2*t);             % define h1(t)
h2 = exp(-3*t);             % define h2(t)
h = conv(h1,h2)*0.01;       % convolve h1(t)* h2(t)
t = 0:0.01:20;              % new time interval = 20s
plot(t,h);
grid on;
```

The convolution result is plotted in Figure 3.26. The combination of two causal systems in series is also a causal system.

## 3.8 ◎ Convolution and Random Numbers

The convolution integral appears in a number of diverse applications, in particular in statistical analysis and the study of random noise in systems. The following example is based on the statistics of throwing dice. This analysis is simplified using MATLAB.

If a fair die is thrown, the possible outcomes (one of the digits 1 through 6) are equally probable with the likelihood of a given digit appearing being 1/6. This situation can be modelled by defining a variable with six equal values as shown below. This result is sketched as the *probability density function* in Figure 3.27A.

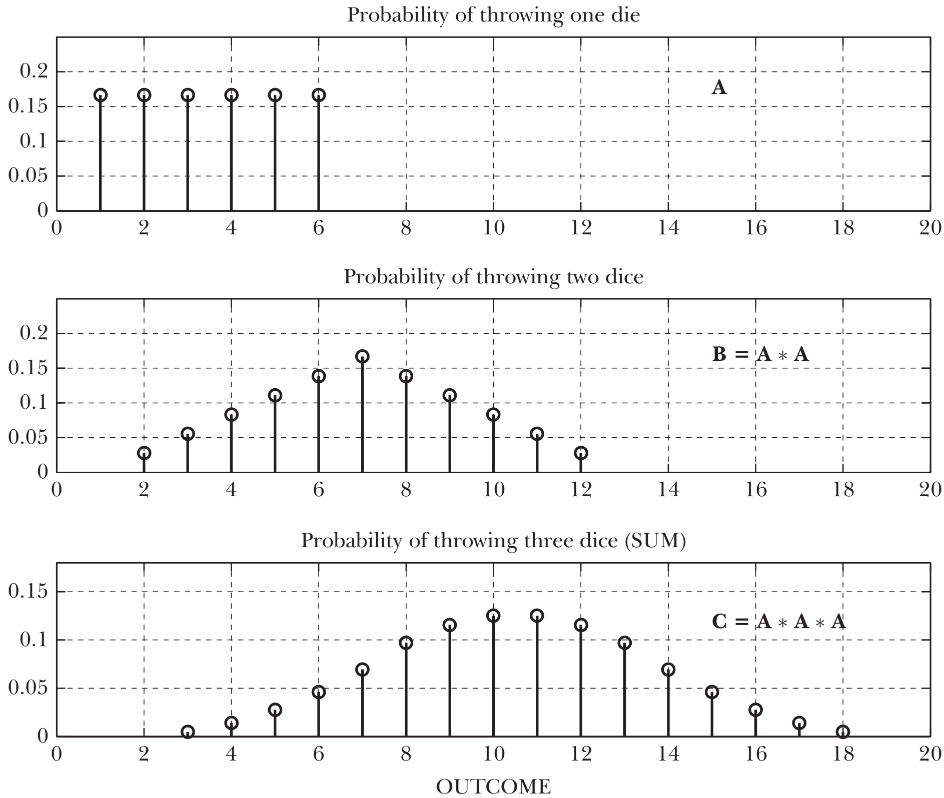



**FIGURE 3.27** The probability of any outcome from the toss of a fair die is 1/6 (A). When two dice are thrown (B), their sum varies from 2 to 12 with 7 the most likely outcome, as given by the convolution $B = A * A$. For three or more dice (C), the convolution operator can be repeatedly applied.

```
a = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6 ]
a = 0.1667  0.1667  0.1667  0.1667  0.1667  0.1667
```

When a pair of dice is thrown, the outcome is the sum of two independent events, each with the same probability. It is expected that of the $6 \times 6 = 36$ possible outcomes, some will be more likely than others, with 7 being the most likely sum. The overall probability density function for two dice can be obtained from the convolution of the single die case convolved with itself, or $B = A * A$, as shown in Figure 3.27B. Note that this operation is equivalent to convolving two rectangles to give a triangular shape.

The MATLAB convolution operator gives the (scaled) probability density function for two dice, where 7 is likely to be the sum 6 out of 36 times:

```
b = 36 * conv(a, a)
b = 1   2   3   4   5   6   5   4   3   2   1
```

This result may be confirmed by inspection. It follows that if three dice are thrown, the probability density function can be computed as the convolution $C = A * A * A$. The sum as shown in Figure 3.27C is between 3 and 18, giving a wider curve (consistent with the properties of the convolution operation). Observe that the curve shape is somewhat more rounded. By extension, if more dice were thrown, this curve would become wider and more rounded. In the limit, the curve approaches the classical Gaussian density function. This fundamental outcome generally emerges from the sum of many random numbers from a given density function. The implications of the *Central Limit Theorem* are directly relevant to signals analysis in the study of random noise in systems.

## 3.9 Useful Hints and Help with MATLAB

1. Convolution of Long Signals—The convolution function `conv(a,b)` convolves two signals $a$ and $b$, requiring one integration for every point. Consequently, the `conv(a,b)` operation will be relatively slow when $a$ and $b$ have many points.
2. Convolution of Periodic Signals—A sinusoidal signal exists for all time, but only a finite time interval of this signal can be stored for computation. A signal such as $\cos(t)$ may be defined in MATLAB over a short interval of time, but never for all time. When this signal is convolved with another, the convolution is effectively taking place with only the few cycles of cosine defined for the vector in which it is stored. This problem is most obvious at the extremes where the cosine representation begins and ends.

   Consider the convolution of $a(t) = \cos(2\pi t)$ with a unit rectangle centered at $t = 2$, both defined on a time interval [0,4], as seen in Figure 3.28.

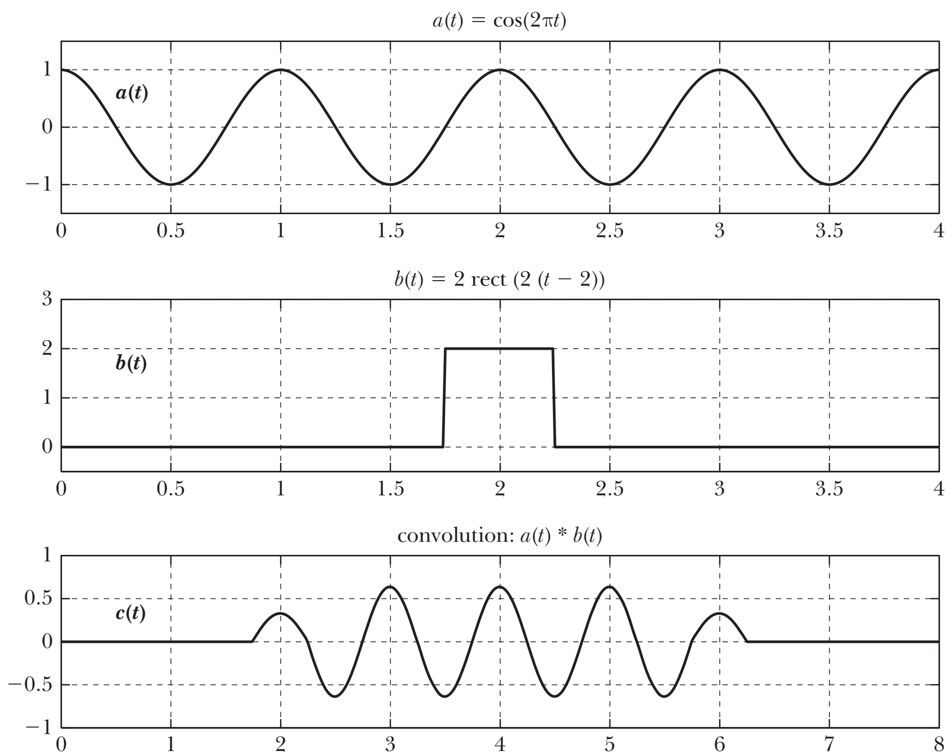



**FIGURE 3.28** The convolution of the cosine $a(t)$ with a unit area rectangle $b(t)$ is expected to give a result that exists for all time. In this example, the result $c(t)$ is constrained to a limited time frame just as the cosine signal $a(t)$ has been defined over only four seconds. The result is valid at the center.

```
t = 0:0.01:4;                   % define interval
a = cos(2*pi*t);                % define a(t)
b = 2*rectpuls(2*(t-2));        % define b(t)
c = conv(a,b)*0.01;             % convolution
```

The cosine signal $a(t)$ exists only in the interval [0,4] s and when convolved with the unit area rectangle $b(t)$ in Figure 3.28, the result only becomes non-zero across the width of the defined cosine signal and not over all time as expected of a cosine. The convolution result $c(t)$ exists only within a five-second interval and, moreover, is not accurate near the limits of this region. It is not possible to represent a signal defined for all time.

## 3.10 Chapter Summary

Many systems of practical interest can be classified as linear systems, and even some nonlinear systems can be treated as linear systems if constrained to operate in a *linear region.* Linear systems obey the property of superposition, whereby the sum of two input signals leads to an output equal to the sum of the corresponding outputs considered separately. Linear systems are characterized by their impulse response function, called $h(t)$.

If a signal $s(t)$ is input to a linear system, the output signal $g(t)$ is the convolution of $s(t)$ and $h(t)$, written $g(t) = s(t) * h(t)$. The convolution of a unit impulse $\delta(t)$ with any signal equals the same signal. Consequently, if a unit impulse $\delta(t)$ is input to a system, the output $g(t)$ will be $g(t) = \delta(t) * h(t)$, which simply equals the system's (unknown) $h(t)$, which is also known as impulse response for this reason. While the convolution operation is somewhat involved mathematically, it will later be seen that the same operation can be greatly simplified when approached from the frequency domain. Finally, a real system must be causal, and a real distortionless system introduces nothing more serious than a delay and amplification.

## 3.11 Conclusions

Linear time-invariant (LTI) systems form an important class of systems that will be used throughout this treatment of signals and systems. If an input signal can be simplified and described by a linear combination of component signals, then each of those components can be treated separately as they pass through a linear system, and then recombined to recover the complete output signal. These characteristics are therefore fundamental to a thorough study of signals and systems. The next chapter deals with the decomposition of various signals into unique components to simplify their analysis.

## End-of-Chapter-Exercises

**3.1** For each of the following systems shown as $s(t) \rightarrow s_2(t)$, describe the system as being linear and/or time invariant.

**(a)** $s_2(t) = 2s(t)$

**(b)** $s_2(t) = s^2(t)$

**(c)** $s_2(t) = \sin(t)s(t)$

**(d)** $s_2(t) = \log(s(t))$

**(e)** $s_2(t) = \frac{1}{s(t)}$

**(f)** $s_2(t) = \frac{d}{dt}s(t)$

**(g)** $s_2(t) = \int_{-\infty}^{t} s(x)\, dx$

**3.2** Consider the convolution $c(t) = a(t) * b(t)$, where $a(t)$ is an even triangle with width 2 and height 2, and $b(t)$ is a unit rectangle.

- **(a)** Sketch $a(t)$ and note its width and area.
- **(b)** Write $a(t)$ as a piecewise function of time in four intervals.
- **(c)** What are the expected width and area of the convolution $c(t)$?
- **(d)** Identify five intervals for computing the convolution $c(t)$.
- **(e)** Give an expression for the convolution $c(t)$ in each of the five intervals.
- **(f)** Sketch $c(t)$ and confirm the expected width and area from above.

**3.3** Repeat the convolution of Question 3.2 using MATLAB.

- **(a)** Confirm the appearance of $c(t)$ in a graph.
- **(b)** Confirm the area under $c(t)$.

**3.4** A cosine signal $a(t) = \cos(2\pi t)$ is input to an LTI system having the response function $h(t) = 2\ \mathrm{rect}(2t)$.

- **(a)** Is this a causal system?
- **(b)** Calculate the output $c(t)$.
- **(c)** What is the amplitude of $c(t)$?
- **(d)** This convolution was computed using MATLAB in Figure 3.28. Compare the above result to that in the figure.

**3.5** Consider the signals $a(t) = \cos(2\pi t)$ and $b(t) = \mathrm{rect}(t)$, and use MATLAB to determine the convolution $c(t)$. Explain the form of $c(t)$ with the aid of a sketch.

**3.6** A voltage signal $s(t)$ travels over a communications network and emerges delayed by 10 msec and reduced to one third its original amplitude but otherwise undistorted. Sketch the response function $h(t)$ for this linear system and give an expression for $h(t)$.

**3.7** A linear system has the response function $h(t) = b(t-5)$, where $b(t) = \mathrm{rect}(t/4)$.

- **(a)** Sketch the output of this system for an impulse input $\delta(t)$.
- **(b)** Sketch the input signal $w(t) = \delta(t) + \delta(t-2)$.
- **(c)** Sketch the output of this system for the input $w(t) = \delta(t) + \delta(t-2)$.

**3.8** Two linear systems have response functions $h_1(t)$ and $h_2(t)$, respectively, where:

$$h_1(t) = \begin{cases} e^{-2t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$h_2(t) = \begin{cases} e^{-3t} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

A signal $s(t)$ passes through the first system, and the output passes through the second system.

- **(a)** Find the overall response function of the two systems connected in series.
- **(b)** Sketch the overall response function.
- **(c)** Compare the above answer to the MATLAB solution of Section 3.7.

**3.9** Two linear systems have response functions $h_1(t)$ and $h_2(t)$, respectively, where:

$$h_1(t) = \delta(t-2)$$

$$h_2(t) = \mathrm{rect}(t)$$

A system $h(t)$ is constructed using these two systems in series.

- **(a)** Find the overall response function $h(t)$.
- **(b)** Sketch the overall response function.
- **(c)** Are $h_1(t)$ and $h_2(t)$ causal systems?
- **(d)** Is $h(t)$ a causal system?
- **(e)** Does the order of the two systems in series matter in the above answers?

**3.10** A linear system has the response function $h(t) = 10\delta(t-10)$.

- **(a)** Sketch the output of this system for an impulse input $\delta(t)$.
- **(b)** Sketch the input signal $w(t) = 2\mathrm{rect}(t/2)$.
- **(c)** Sketch the output of this system for the input $w(t)$.

**3.11** A linear system has the response function $h(t) = 3\delta(t-6) + 2\delta(t-7)$.

- **(a)** Sketch the output of this system for an impulse input $\delta(t)$.
- **(b)** Sketch the input signal $w(t) = 3\mathrm{rect}(t/3)$.
- **(c)** Sketch the output of this system for the input $w(t)$.

**3.12** A linear system has the response function $h(t) = \delta(t) - \delta(t-2)$.

- **(a)** Sketch the output of this system for an impulse input $\delta(t)$.
- **(b)** Sketch the input signal $w(t) = 10\cos(2\pi t)$.
- **(c)** Sketch the output of this system for the input $w(t)$.
  *Hint: This answer may be best solved graphically. Draw the two copies of the input signal on the same graph.*

**3.13** A linear system has the response function $h(t) = e^{-2t}$ defined for $t > 0$.

- **(a)** Sketch the output of this system for an impulse input $\delta(t)$.
- **(b)** Sketch the input signal $w(t) = \delta(t) + \delta(t-1)$.
- **(c)** Sketch the output of this system for the input $w(t) = \delta(t) + \delta(t-1)$.

**3.14** Refer to the signal $s(t)$ in Figure 3.29.

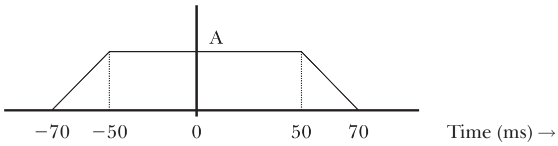



**FIGURE 3.29** Diagram for Question 3.14.

**(a)** Identify $s(t)$ as odd, even, or neither.
**(b)** Determine the two signals that convolved to give this signal.

**3.15** A signal $x(t)$ is convolved with itself. The result $s(t)$ is an even triangle centered on the origin, spanning $-5 < t < +5$ s, with a height of 5.

**(a)** Sketch $s(t)$.
**(b)** Explain how to find $x(t)$ from the information supplied.
**(c)** Make a neat labelled sketch of $x(t)$.

**3.16** A periodic signal $s(t)$ is a 1 V pulse train with 10 percent duty cycle and period 10 seconds. Show with the aid of a sketch how this signal can be formed as the convolution of an impulse train and a unit rectangle.

**3.17** A periodic signal $s(t)$ is a 2 V pulse train with 50 percent duty cycle and period 10 seconds. Show with the aid of a sketch how this signal can be formed as the convolution of an impulse train and a unit rectangle.

**3.18** A voltage signal $v(t)$ travels through an amplifier and emerges delayed by 0.001 s and increased to ten times its original amplitude but otherwise undistorted. Write an expression for the response function $h(t)$ for this amplifier system. Make a sketch of $h(t)$.

**3.19** An automobile with poor shock absorbers is observed bouncing along after striking a speed bump. The height of the front bumper gives the impulse response $h(t)$ as shown in Figure 3.30. This graph shows that when an impulse arrives at a system initially at rest, the output begins to oscillate with decreasing amplitude. Eventually the vehicle drives smoothly again, until another bump is encountered.

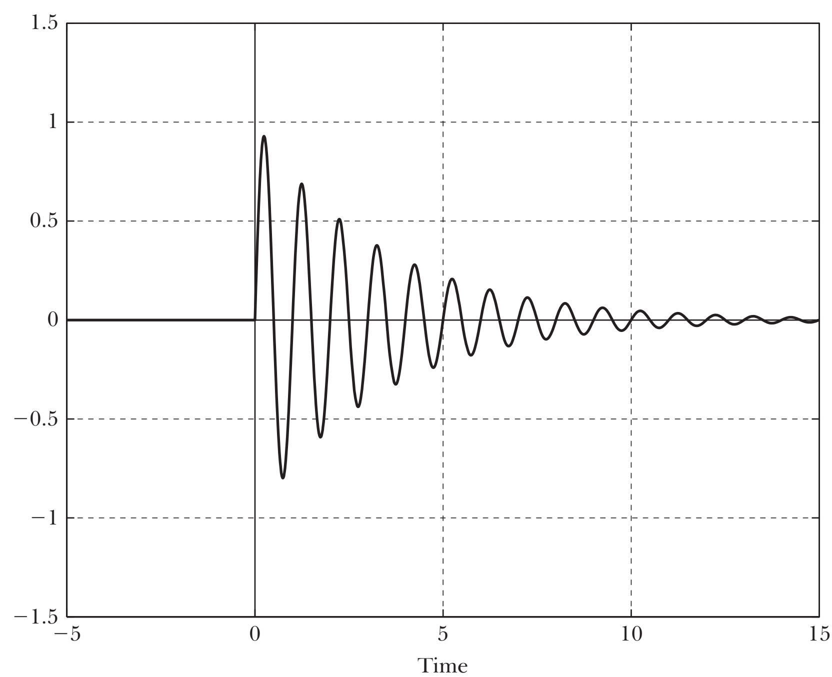



**FIGURE 3.30** Diagram for Question 3.20.

(a) Is this a causal system?
(b) From the graph, write an expression for $h(t)$ in terms of a sine wave, an exponential and a step function.
(c) Give an expression for the system output $h_2(t)$ in terms of $h(t)$ if a second identical bump is encountered at time $t = 5$ s.
(d) Sketch $h_2(t)$.

**3.20** A system is defined by the impulse response $h(t)$ from Figure 3.30, is to be modelled in MATLAB. Give a mathematical expression for this exponentially decaying sine wave. Use MATLAB to find the system response to a 1 Hz cosine input.

**3.21** A system is designed for which an input signal emerges delayed by 1.0 second, followed by two echoes consisting of another copy of the same signal 1.0 second reduced to half amplitude and another copy of the same signal reduced to one quarter the original amplitude 1.0 second after that. Write an expression for the response function $h(t)$ for this system. Make a sketch of $h(t)$.

**3.22** The linear system in Figure 3.31 has the (unknown) response function $h(t)$. A unit rectangle $rect(t)$ input to the system leads to the output signal is shown. Find $h(t)$.

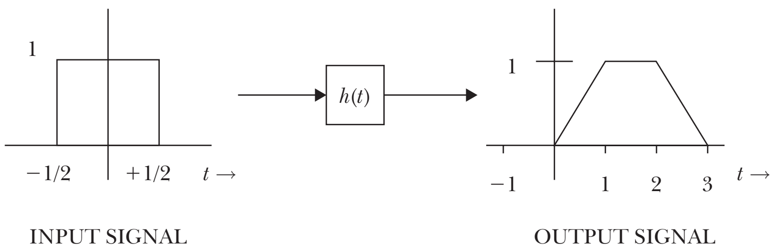



**FIGURE 3.31** Diagram for Question 3.22.

(a) Is $h(t)$ a linear system?
(b) Sketch the impulse response $h(t)$.
(c) Is this a causal system?

**3.23** The system in Figure 3.32 is an *integrator* with response function $h(t)$ given by:

$$h(t) = \int_{-\infty}^{t} s(x)\,dx$$

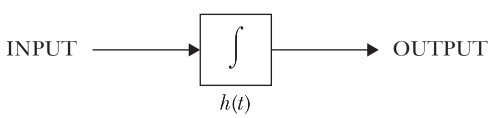



**FIGURE 3.32** Diagram for Questions 3.23 to 3.25.

**3.24** For the system of Figure 3.32, determine the output for the input signal $s(t) = 3\,\text{rect}(t/2)$.

**3.25** For the system of Figure 3.32, determine the output for the input signal $s(t) = \delta(t-2) + \delta(t-4)$.

CHAPTER 4

# The Fourier Series

In this chapter, arbitrary signals are expressed as linear combinations of component signals. Suitable sets of component signals are identified, and the *one-sided cosine-frequency* approach is extended to represent arbitrary periodic signals as the sum of orthogonal components. This discussion finally leads to the complex Fourier series as a generalized frequency-domain representation of periodic functions.

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Demonstrate how to express graphically one signal in terms of another
- Explain the concept of orthogonal functions
- List different possible sets of orthogonal functions
- Identify the orthogonal components of an arbitrary signal
- Use MATLAB to combine component signals into a target signal
- Define the orthogonal basis for a Fourier series representation
- Derive the Fourier series components of a periodic signal
- Explain how a Fourier series is affected by time-domain variations
- Apply Fourier series properties to identify a signal from its components
- Use MATLAB to identify the frequency components of a target signal

## Chapter Overview

This chapter is divided into three parts. Part One introduces the notion of orthogonal signals and explores the process of identifying the orthogonal components of a signal. There are many possible sets of orthogonal component signals, but in Part Two the use of components based on sinusoids is established as a particularly useful choice, leading to the *Fourier series* approximation to a periodic signal. In Part Three, the Fourier series is generalized to the complex components that are conventionally used to represent periodic signals and that form the basis for the *Fourier transform* in Chapter 5.

## 4.1 Introduction

Signals can take on a variety of shapes and forms, requiring a number of specialized techniques for their analysis. The study of arbitrary signals $s(t)$ can be greatly simplified if such signals can be expressed in terms of well-understood components. It has already been seen that simple mathematical manipulations of sines and cosines and rectangles and impulses can be used to create different signals. To ensure the systematic decomposition of arbitrary signals into unique and useful components, a careful study of signal synthesis is required. Once a suitable set of component signals is identified, further analysis of any signal $s(t)$ can be based on the clear understanding of a handful of basic signal types.

# 4.2 Expressing Signals by Components

When signals are added or multiplied together, the resulting signal may not resemble any of the original components. If the signals possess some special properties (odd, even, real, imaginary, periodic), then some properties of their sum or product can be predicted. In any case, even if a waveform is known to be composed of different (unknown) signals, finding those exact signals may prove to be difficult.

Consider the periodic signal $s(t) = 2\cos(t) + 3\cos(2t)$, defined over all time. If this signal is expressed as $s(t) = a(t) + b(t)$, where $a(t)$ and $b(t)$ are sinusoidal, can the signals $a(t)$ and $b(t)$ be uniquely determined from an examination of $s(t)$?

A graphical approach may be helpful. Before pursuing the usual path of plotting amplitude *vs.* time (which will require some effort and possibly a calculator), it is worthwhile to observe that the one-sided cosine amplitude *vs.* frequency graph can be drawn directly without the need for calculations, as shown in Figure 4.1. Working back from $s(t)$ begins with this study of the signal $s(t)$ as seen from the frequency domain.

On the other hand, if this same signal $s(t)$ was shown graphically (or viewed on an oscilloscope) as the time-domain waveform shown in Figure 4.2, the component signals are not so readily found.

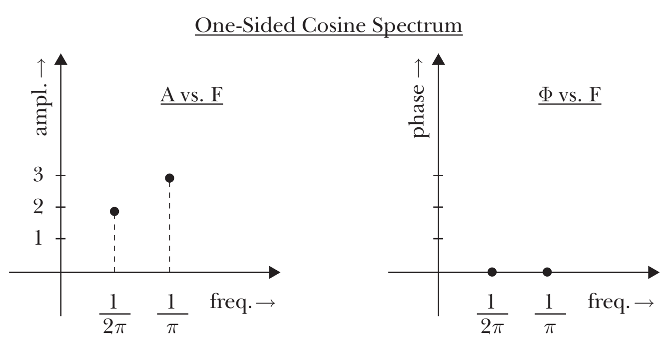



**FIGURE 4.1 One-Sided Cosine Graph** From this graph, the (*Amplitude*, *Frequency*, *Phase*) values associated with the two cosine signals contributing to $s(t)$ are $(2, 1/2\pi, 0)$ and $(3, 1/\pi, 0)$, respectively.

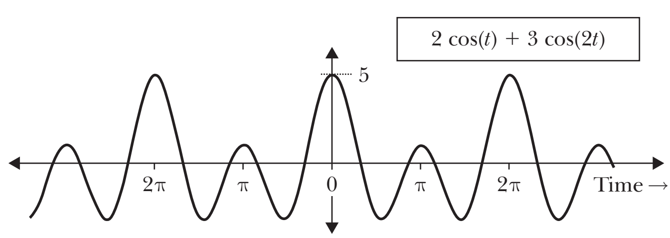



**FIGURE 4.2 Time-Domain Graph** It may be difficult to identify the two cosine components of $s(t) = 2\cos(t) + 3\cos(2t)$ working from the time-domain graph alone.

In this example, the use of an amplitude vs. frequency graph has clear advantages over the amplitude vs. time version. Not only can the graph be drawn directly from the equation but, significantly, the *cosine frequency components* can be identified directly from a sketch of the function.

It may not seem too surprising that two cosine components can be clearly identified on the one-sided cosine graph, which, after all, was introduced specifically for the signal $s(t) = 2\cos(t) + 3\cos(2t)$. To be useful, this amplitude vs. frequency representation should be applicable to general (periodic) signals that are not defined explicitly in terms of cosines. It should be demonstrated whether or not some other linear combination of cosines might also lead to $s(t)$. Similarly, a general method for finding the frequency components in an arbitrary $s(t)$ such as in Figure 4.3 must be identified. This chapter explores the answers to these and many other related questions.

**The Spectrum Analyzer** Just as the time-domain representation of a signal can be seen directly on an oscilloscope, the same signal can be viewed by frequency using a *spectrum analyzer* as seen in Figure 4.4. This display essentially presents a one-sided sinusoid plot with no phase information. Traditional analog spectrum analyzers

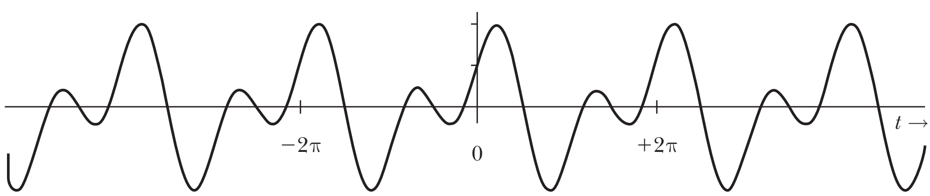



**FIGURE 4.3** How can the frequency components of an arbitrary waveform be identified?

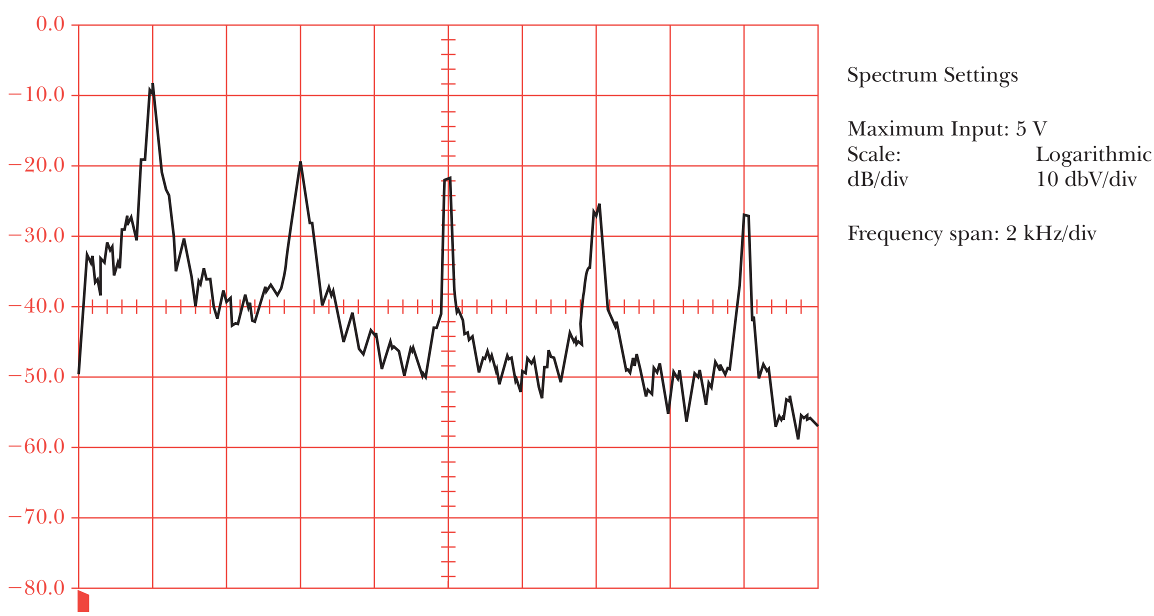



**FIGURE 4.4** This *spectrum analyzer* display shows signals directly as amplitude vs. frequency. In contrast, an oscilloscope displays signals as amplitude vs. time. In this case, the input signal is a time-domain square wave.

employ tuned-frequency circuits much like a radio receiver and are significantly more expensive than oscilloscopes, depending on the accuracy and range of frequencies to be measured. On the other hand, modern digital oscilloscopes often have built-in or add-on spectrum analyzer capabilities. These devices use microprocessors and effectively perform the computations described in this chapter to extract orthogonal frequency components.

### 4.2.1 Approximating a Signal $s(t)$ by Another: The Signal Inner Product

In Chapter 1, various signals were added and subtracted, and, in doing so, different signals were created. One aim of this chapter is to show that complicated signals can be expressed generally as a linear combination of two or more component signals, which, individually, will prove easier to analyze. On the other hand, given a signal $s(t)$ that is formed from some unknown $a(t) + b(t)$, it may not be evident which $a(t)$ and $b(t)$ led to this signal. To be useful, the signals $a(t)$ and $b(t)$ should form a unique decomposition of $s(t)$, and they should be relatively easy to identify, given only $s(t)$.

In general, signals often share common features, and one signal can often be approximated by using a certain fraction (component) of another, although the approximation with some arbitrary component signal may not be very good. Logically, an odd signal would be expected to be a linear combination of odd components. For example, an odd square wave $s(t)$ might be roughly approximated using a sine wave $a(t)$ as shown in Figure 4.5. The problem becomes how to determine the sine wave amplitude $A\,a(t)$ that provides the best approximation to the square wave.

Consider a square wave $s(t)$ with a period of 100 ms, and let the sine wave be $a(t) = \sin(20\pi t)$ to have the same period as the square wave. At any time $t$, the difference between the square wave $s(t)$ and the approximation $A\,a(t)$ is given by error $(t) = [s(t) - A\,a(t)]$. This difference reflects the goodness of the approximation as shown shaded in the figure and varies with amplitude $A$. The relative instantaneous error may be computed as the square of the difference at every point $t$. The *squared error* is given by $error^2(t) = [s(t) - A\,a(t)]^2$. The overall squared error[1] may then be determined by summing (integrating) this squared error over one complete period of the signal. The best $A$ is the value that minimizes the squared error. To find this *least squared error* graphically, the amplitude $A$ may be adjusted until the shaded area is minimized; in Figure 4.5, this error is shown very close to the minimum.

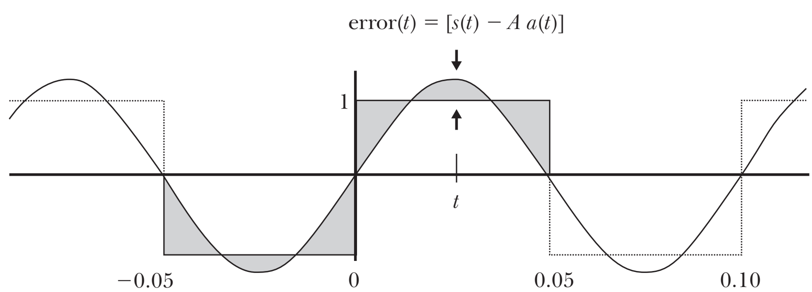



**FIGURE 4.5 Odd Square Wave** $s(t)$ approximated by $a(t) = A\sin(20\pi t)$. (Best approximated when $A = 4/\pi$.)

[1] Squaring ensures, at least, that positive and negative errors do not cancel each other.

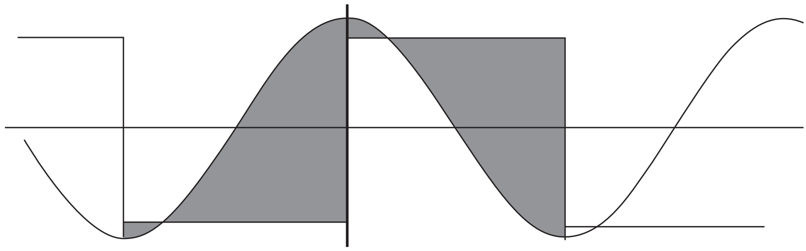



**FIGURE 4.6 Odd Square Wave** $s(t)$ approximated by $A\,a(t) = A\cos(20\pi t)$. (Best approximated when $A = 0$.) Compare to Figure 4.5.

Mathematically, the derivative with respect to $A$ of the squared error may be computed and set to zero, to find the amplitude $A$ giving the least squared error. This *best-fit* value in this particular example will be shown to be $A = 4/\pi$ or *slightly higher than* the square wave amplitude.

To further explore this approach, consider Figure 4.6 in which the same odd square wave is approximated with a cosine waveform $b(t) = \cos(20\pi t)$. Now, it is expected that an even cosine $B\,b(t)$ cannot be a good approximation to an odd square wave. Indeed, the error shown shaded in this figure can only grow larger as the cosine amplitude is increased, and, by inspection, this error is minimum for $B = 0$. In other words, the best approximation is no cosine at all. This result reflects the fact that no component of a cosine can usefully contribute to approximating an odd square wave. In general, no component of any even signal can usefully contribute to approximating any other odd signal, and vice versa.

## 4.2.2 Estimating One Signal by Another

The problem of determining the *best approximation* of an arbitrary signal $s(t)$ by some other signal can be approached generally. Let the component signal be $a(t)$, and find the constant $A$ for which the product $A\,a(t)$ best approximates the signal $s(t)$. The *best fit* will be defined as the minimum cumulative squared difference between the signal $s(t)$ and its approximation $A\,a(t)$ over the time interval $[t_0, t_1]$. In other words, find a value $A$ for which the integral of $\text{error}^2(t) = [s(t) - A\,a(t)]^2$ is minimum over the defining region; this is the least squared approximation to $s(t)$ using $a(t)$:

$$\text{total squared error}(t) = \int_{t_0}^{t_1} [s(t) - A\,a(t)]^2\,dt \tag{4.1}$$

To minimize the total squared error, compute the derivative with respect to $A$, and set to zero.

$$\begin{aligned}\frac{d}{dA}[\text{total squared error}(t)] &= \frac{d}{dA}\int_{t_0}^{t_1} [s(t) - A\,a(t)]^2\,dt \\ &= \frac{d}{dA}\int_{t_0}^{t_1} [s^2(t) - 2A\,a(t)s(t) + A^2 a(t)a(t)]\,dt \\ &= -2\int_{t_0}^{t_1} a(t)s(t)\,dt + 2A\int_{t_0}^{t_1} a(t)a(t)\,dt\end{aligned}$$

This derivative is zero for:

$$A = \frac{\int_{t_0}^{t_1} a(t)s(t)\, dt}{\int_{t_0}^{t_1} a(t)a(t)\, dt} \tag{4.2}$$

This is the value for $A$ to minimize the squared error between $s(t)$ and the approximation $A\,a(t)$. This result is shown to be a minimum by examining the sign of the second derivative. Since that integral must be positive, the value of $A$ corresponds to a minimum, as required.

Note that in the special case of a symmetric interval $[-b,+b]$ the integral in the numerator will be zero if the product $s(t)a(t)$ is odd. As in the previous example, whenever $s(t)$ is odd and $a(t)$ is even, or vice versa, there are no components of an odd signal in an even signal.

To be completely general, the above result should be extended for a complex signal $s(t)$. In this case, the magnitude of the squared difference includes the complex conjugate of the (now complex) $a(t)$ used in the derivation, as shown below. Of course, the complex conjugate has no effect on real signals, so it may simply be ignored unless complex signals are involved.

$$A = \frac{\int_{t_0}^{t_1} a(t)s^*(t)\, dt}{\int_{t_0}^{t_1} a(t)a^*(t)\, dt} \tag{4.3}$$

The above results are fundamental to approximating an arbitrary signal by a linear combination of component signals. A suitable set of component signals $a(t)$ can then be identified, which will be useful in general signals analysis. A particular set of component signals of $\{a_n(t)\}$ of great interest defines the Fourier series approximation to a periodic waveform. The concepts underlying a full understanding of how the Fourier series describes a complete set of orthogonal functions are explored in the following section.

## 4.3 Part One—Orthogonal Signals

A general overview of the concept of orthogonality as related to signals is presented here. Principles of orthogonality as related to signals are compared to vectorial interpretations, which may be more easily visualized. Examples using a limited set of simple orthogonal basis signals serve to illustrate these important principles.

- Introduction to Orthogonality: The Best Approximation
- The Inner Product for Signals
- An Orthogonal Signal Space
- The Signal Inner Product Formulation
- Orthogonal Components of an Arbitrary Signal
- Complete Sets of Orthogonal Signals
- Examples: An Orthogonal Set of Signals
- Aside: Orthogonal Signals and Linearly Independent Equations
- Using MATLAB to Evaluate the Signal Inner Product

## 4.4 Orthogonality

Not all decompositions of a signal into components will necessarily be useful, few will be unique, and fewer still will contribute to the aim of simplifying signals analysis. A suitable decomposition into component signals must satisfy conditions that ensure a unique and useful approximation of any signal $s(t)$ with a linear combination of those components. Such a set of signals is said to be orthogonal.

The word *orthogonal* is defined as meaning *right angled* or *mutually perpendicular*. Orthogonality is perhaps most familiar in terms of vectors. Two vectors that are orthogonal (at right-angles to each other) can be easily identified mathematically because their inner product (dot product) equals zero (e.g., $\hat{x} \cdot \hat{y} = 0$). The same inner product is used to find the $x$- and $y$-components of an arbitrary vector $\vec{s}$ and in doing so, to find a unique linear combination of component vectors $\vec{s} = A\hat{x} + B\hat{y}$. (see Figure 4.7).

In the same way, orthogonality can be defined for signals using a signal's inner product (in this case, an integral). When two signals are orthogonal, this inner product will be zero, and this same formulation can be used to find the components of an unknown signal. The underlying mathematics for signal components parallels the manipulation of orthogonal vector components through the inner product.

Just as $x$- and $y$- and $z$-vectors form an complete set of vectors sufficient to represent any vector in three-space $(x, y, z)$, a complete set of orthogonal signals can be defined that will be sufficient to represent any signal. Like the $(x, y, z)$ vector components, a linear combination of orthogonal signals will necessarily be a unique decomposition of an arbitrary signal. While other orthogonal decompositions exist (e.g., cylindrical or spherical coordinate systems), all must obey the same properties of uniqueness.

The consequences of using the signal inner product formulation to find orthogonal components will now be summarized. It is useful to think of vector principles when in doubt about any of the properties.

### 4.4.1 An Orthogonal Signal Space

The concept of orthogonality is readily understood by analogy to vector space. Signals may be described as a linear combination of orthogonal components. For example, consider the function of time written as

$$s(t) = A_1 a_1(t) + A_2 a_2(t) + A_3 a_3(t) + \cdots + A_N a_N(t)$$

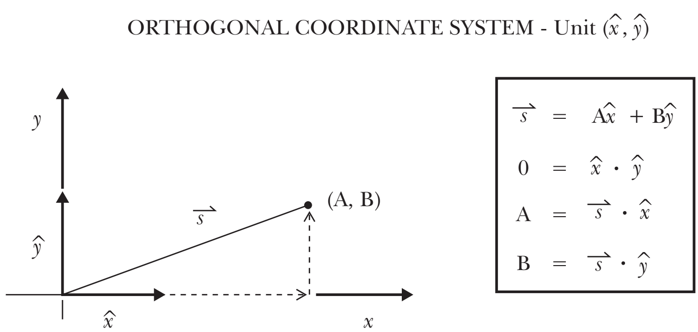



**FIGURE 4.7** A vector $\vec{s}$ in $(x, y)$ space is expressed in terms of $x$- and $y$-components (A, B) using the inner product. A similar formulation will allow a signal to be expressed as components of orthogonal basis signals.

where the set of $N$ functions $\{a_1(t), a_2(t), a_3(t), \cdots a_N(t)\}$ represent basis functions of signal space. The signals in this set are expected to be *mutually perpendicular* or *mutually orthogonal.* Each of the values $\{A_n\}$ are constants.

1. The inner product for signals is an integral over an interval $[t_0, t_1]$.

$$x(t) \cdot y(t) = \int_{t_0}^{t_1} x(t)y^*(t)\, dt \tag{4.4}$$

The signals $x(t)$ and $y(t)$ are *orthogonal* if this inner product equals zero. [5] Orthogonality applies only in the interval $[t_0, t_1]$ for which this integral is defined. Note the complex conjugate form of $y^*(t)$ is of importance only for complex signals.

2. The signals in the set $\{a_1(t), a_2(t), a_3(t), \cdots a_N(t)\}$ are mutually orthogonal if each one is orthogonal to all others, as given by:

$$\int_{t_0}^{t_1} a_n(t)a_n(t)\, dt = 0, \quad \text{for all } n \neq m \tag{4.5}$$

3. A normalized orthogonal or *orthonormal* set of signals is identified when the inner product of each basis signal with itself equals unity:

$$\int_{t_0}^{t_1} a_n(t)a_m(t)\, dt = 1, \quad \text{for all } n = m \tag{4.6}$$

Signal space is not necessarily defined on orthonormal signals. In practice, a constant factor restores a normalized set of orthogonal basis signals (see item 5 below).

4. Once an orthogonal set of $N$ base signals has been defined as:

$$\{a_1(t), a_2(t), a_3(t), \cdots a_N(t)\}$$

then an arbitrary signal $s(t)$ can be described as a linear combination of the components:

$$s(t) = A_1 a_1(t) + A_2 a_2(t) + A_3 a_3(t) + \cdots + A_N a_N(t)$$

and recorded as $\{A_1, A_2, A_3, \cdots A_N\}$.

5. To determine the orthogonal components of an arbitrary signal $s(t)$, the inner product with each of the orthonormal basis signals can be applied:

$$A_n = \int_{t_0}^{t_1} s(t)a_n(t)\, dt \tag{4.7}$$

In general, if the basis signals are not orthonormal:

$$A_n = \frac{\displaystyle\int_{t_0}^{t_1} s(t)a_n(t)\, dt}{\displaystyle\int_{t_0}^{t_1} a_n(t)a_n(t)\, dt} \tag{4.8}$$

6. Because the $\{a_1(t), a_2(t), a_3(t), \cdots a_N(t)\}$ are mutually orthogonal, they cannot be described in terms of each other.

7. Because the $\{a_1(t), a_2(t), a_3(t), \cdots a_N(t)\}$ are mutually orthogonal, the set $\{A_1, A_2, A_3, \cdots A_N\}$ is unique. No other linear combination of these signals can better represent the signal $s(t)$.
8. The orthogonal set $\{A_1, A_2, A_3, \cdots A_N\}$ is called a *complete set* if *any* signal $s(t)$ can be represented as a linear combination of these signals.

**Interpreting the Inner Product** The interpretation normally given to the inner product integral

$$A_n = \int_{t_0}^{t_1} a_n(t)s(t)\ dt$$

as in Eqn. 4.7 is to say that the constant $A_n$ represents the *component of* $a_n(t)$ *present in the signal* $s(t)$. Consequently, if $A_n = 0$, there is no component of $a_n(t)$ present in the signal $s(t)$. Conversely, if $a_n(t)$ is identical to $s(t)$, then $A_n$ should be exactly one, since *there is one component of* $a_n(t)$ *present in the signal* $s(t)$. To ensure this, the denominator term in Eqn. 4.8 is included to normalize the values of $A_n$.

## 4.4.2 The Signal Inner Product Formulation

How does the inner product "extract" orthogonal components from an arbitrary signal? Because the signal inner product of orthogonal terms is zero, applying the formula to a linear combination of orthogonal components will yield only the one term of interest. This effect is illustrated in detail below.

Assume that a signal $s(t)$ is known to a linear combination of $N$ orthogonal components of $\{a_n(t)\}$ defined on the interval $[t_0,t_1]$. Find the $a_m(t)$ component of $s(t)$.

$$s(t) \approx \sum_{n=0}^{\infty} A_n a_n(t)\ dt \tag{4.9}$$

Taking the inner product with $a_m(t)$ of both sides of the equation yields:

$$\int_{t_0}^{t_1} s(t)a_m(t)\ dt \approx \int_{t_0}^{t_1} \left[\sum_{n=0}^{\infty} A_n a_n(t)\right] a_m(t)\ dt \tag{4.10}$$

This can be rearranged to give:

$$\int_{t_0}^{t_1} s(t)a_m(t)\ dt \approx \sum_{n=0}^{\infty} A_n \int_{t_0}^{t_1} a_n(t)a_m(t)\ dt \tag{4.11}$$

Now if and only if $a_n(t)$ is orthogonal to $a_m(t)$, then all the terms in $a_n(t)a_m(t)$ will be zero except when $n = m$, leaving only the desired result:

$$\frac{\displaystyle\int_{t_0}^{t_1} s(t)a_m(t)\ dt}{\displaystyle\int_{t_0}^{t_1} a_m(t)a_m(t)\ dt} = A_m \tag{4.12}$$

This is the same result obtained in Eqn. 4.2 using a *best-fit* approach.

**EXAMPLE 4.1** **(Orthogonality of Odd and Even Signals)**

*Consider an odd signal* $g(t)$ *and an even signal* $s(t)$. *Are these signals orthogonal in the interval* $[-a, +a]$?

**Solution:**
To demonstrate orthogonality in the interval $[-a, +a]$, it must be shown that:

$$\int_{-a}^{+a} g(t)s^*(t)\, dt = 0$$

Now, the product of an odd and an even function is necessarily odd, and for any odd function the above integral over $[-a, +a]$ will be zero. Therefore, by inspection, the signals $g(t)$ and $s(t)$ are orthogonal in the interval $[-a, +a]$. Over some other interval, such as $[0, +a]$, the two signals are not necessarily orthogonal.

This result is intuitively correct, since it has already been observed that no linear combination of odd functions could ever yield an even result, and vice versa.

**EXAMPLE 4.2** **(Orthogonality of Sine and Cosine Signals)**

*Are the signals* $g(t) = \sin(2\pi f_0 t)$ *and* $s(t) = \cos(2\pi f_0 t)$ *orthogonal over the interval* $[-\pi, \pi]$?

**Solution:**
From Example 4.1, any product of odd and even functions will be orthogonal over the interval $[-\pi, +\pi]$. Since $\sin(2\pi f_0 t)$ is odd, and $\cos(2\pi f_0 t)$ is even, it follows that the product of any sine and cosine functions is orthogonal in this interval. This may be confirmed as follows:

$$\int_{-\pi}^{+\pi} \sin(2\pi f_0 t)\cos(2\pi f_0 t)\, dt = \int_{-\pi}^{+\pi} \frac{1}{2}\sin(4\pi f_0 t)\, dt = 0$$

Over some other interval, such as $[0, +\pi/2]$, the two signals are not necessarily orthogonal.

Once a signal is described as a linear combination of orthogonal components, its properties can be described in terms of the known behavior of the component signals. New signals become easily analyzed when expressed in terms of well-known components. Finally, many otherwise complicated computations can be readily performed on signals once they are expressed in component form. In the latter case, graphical results obtained by inspection will build confidence in theory and a better understanding of the practical world of signals.

## 4.4.3 Complete Set of Orthogonal Signals

Given a suitable set of orthogonal signals, other signals can be constructed as a linear combination of orthogonal components. To be useful in constructing arbitrary signals, the orthogonal set must also be a *complete set.* In other words, simply having a set of $N$ signals that are mutually orthogonal does not guarantee that all conceivable waveforms can be constructed using only those $N$ signals (inclusion of additional orthogonal components might provide a better approximation).

### 4.4.4 What If a Complete Set Is Not Present?

Even if a complete set of orthogonal signals is not available, the computation will always yield a "best approximation" using the orthogonal components at hand. The value in using orthogonal basis signals is precisely that individual signal components may be computed separately from each other.

Using a vector space analogy, the set of unit vectors $\{x, y\}$ is a complete set on the $\{x, y\}$ plane, but is not complete in three-dimensional $\{x, y, z\}$ space. Consider the point $ix + jy + kz$, and find the best approximation to this position using only $\{x, y\}$ components. It would seem reasonable to try to find a position on the $xy$-plane located as close as possible to the point described by $(i, j, k)$.

Think of which point on the floor of a room $(x, y)$ is the closest to a ceiling lamp located at $(i, j, k)$, or answer the question *where would be the best place to stand when changing the light bulb?* The closest or "best possible" $(i, j)$ may be easily found without knowing the missing $z$-component. The desired position would be directly below the point, on a line perpendicular to the $xy$-plane. Now, the $(i, j)$ coordinates of this point are, of course, the same as those found in the complete $(i, j, k)$ description. Orthogonal components may be computed separately from each other.

### 4.4.5 An Orthogonal Set of Signals

The practical definition and use of orthogonal signals can now be studied. To guarantee the unique decomposition of signals into components, it is necessary to:

1. Be certain that the basis functions are mutually orthogonal;
2. Check that the functions are orthonormal, or expect to normalize components;
3. Remember that orthogonality applies only in an interval $[t_0, t_1]$.

**Defining Orthogonal Basis Signals** Consider the set of functions based on the piecewise signals defined on the interval $[-2, +2]$ as shown in Figure 4.8. In fact, a similar set of orthogonal signals forming the *Walsh transform* is well suited to digital signal processing applications.

1. These four functions are mutually orthogonal on the interval $[-2, +2]$, which will be confirmed if the signal inner product equals zero for all possible combinations of the signals $\{a_n(t)\}$.
2. Now, another signal g(t) defined on the interval $[-2, +2]$ may be approximated in terms of $\{a_n(t)\}$ by applying the formula to find each component $A_n(t)$:

$$A_n = \frac{\int_{-2}^{+2} a_n(t)\ g(t)\ dt}{\int_{-2}^{+2} a_n(t)\ a_n(t)\ dt} \tag{4.13}$$

Because the signals $\{a_n(t)\}$ are orthogonal, but not orthonormal, the denominator does not equal 1. On the other hand, this denominator will equal

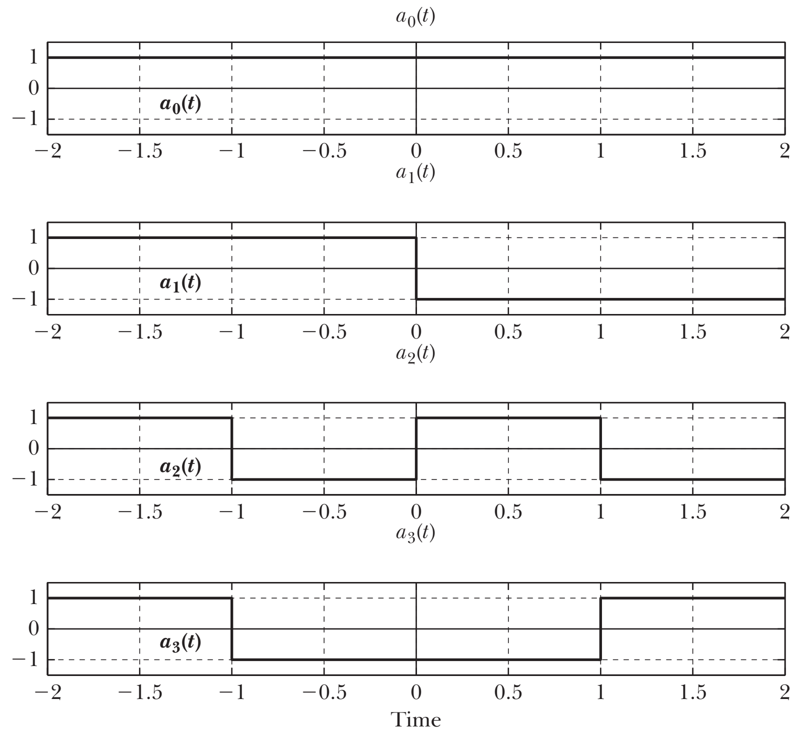



**FIGURE 4.8 Orthogonal Signals** The four signals $\{a_n(t),\, n = \{0, 1, 2\,, 3\}\}$ are mutually orthogonal on the interval $[-2, +2]$.

4 for all of the above signals and the constant term $1/4$ can be used to simplify the previous equation:

$$A_n = \frac{1}{4}\int_{-2}^{+2} a_n(t)g(t)\,dt \tag{4.14}$$

which may be used to compute the linear combination:

$$g(t) = A_0 a_0(t) + A_1 a_1(t) + A_2 a_2(t) + A_3 a_3(t)$$

as the best possible approximation to $g(t)$ using $\{a_n(t)\}$. This set of functions will now be used in several examples illustrating the definition and application of orthogonal signals.

**Confirming Orthogonal Basis Signals** To prove that the signals $\{a_n(t), n = 0, 1, 2, 3\}$ are orthogonal in the interval $[-2, +2]$, it is necessary to show that:

$$\int_{-2}^{+2} a_n(t)\; a_m(t)\; dt = 0, \quad \text{for all } n \neq m$$

A graphical demonstration of the required calculations is seen in Figure 4.9, for six possible combinations of the four basis signals. In each column, the product $a_n(t)a_n(t)$ is obtained graphically and integrated in the bottom row of plots to give $A_{nm}$. By inspection, each integral equals zero when $m \neq n$ and equals 4 when $m = n$. In particular, note the even $\{a_0(t), a_3(t)\}$ and odd $\{a_1(t), a_2(t)\}$. These inner product calculations demonstrate that the signals shown are mutually orthogonal, and the value 4 will appear as a normalizing factor for every non-zero component.

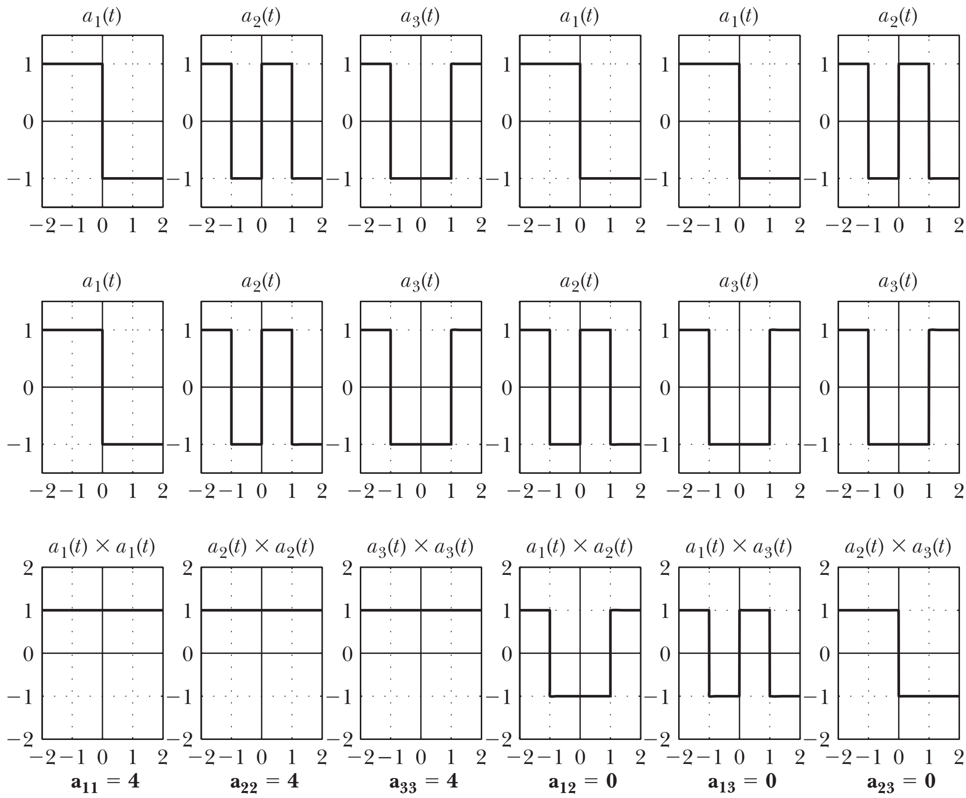



**FIGURE 4.9** The signals $\{a_n(t), n = \{0, 1, 2, 3\}\}$ are mutually orthogonal if the inner product $a_{nm} = a_n(t) \cdot a_m(t)$ is zero for all $n \neq m$. In each column, combinations of these signals are multiplied graphically to compute the inner product. The results $a_{nm}$ shown emerge after integrating the product term in the lower plots and dividing by 4 (see text).

**Finding Orthogonal Components** Once an orthogonal set of basis signals has been defined and verified, the signals can be used to approximate arbitrary signals on the same interval.

### EXAMPLE 4.3 (Orthogonal Components—An Exact Answer)

*Find a best approximation to the staircase signal* s(t) *defined in Figure 4.10 on the interval* $[-2, +2]$ *using components* $\{a_n(t), n = 0, 1, 2, 3\}$.

**Solution:**

Observe that except for the DC-offset ($A_0$ term), the signal $s(t)$ is odd; however, $a_3(t)$ is even, which implies that $A_3$ will be zero.

The four component signals $a_n(t)$ are shown in the top row of the figure, with the target signal sketched underneath each. The product of these component signals with the target signal is shown in the bottom row. The signal dot product for each component signal is the integral of each of these lower plots over the interval $[-2, +2]$, divided by 4 to give the normalized results for each $A_n$.

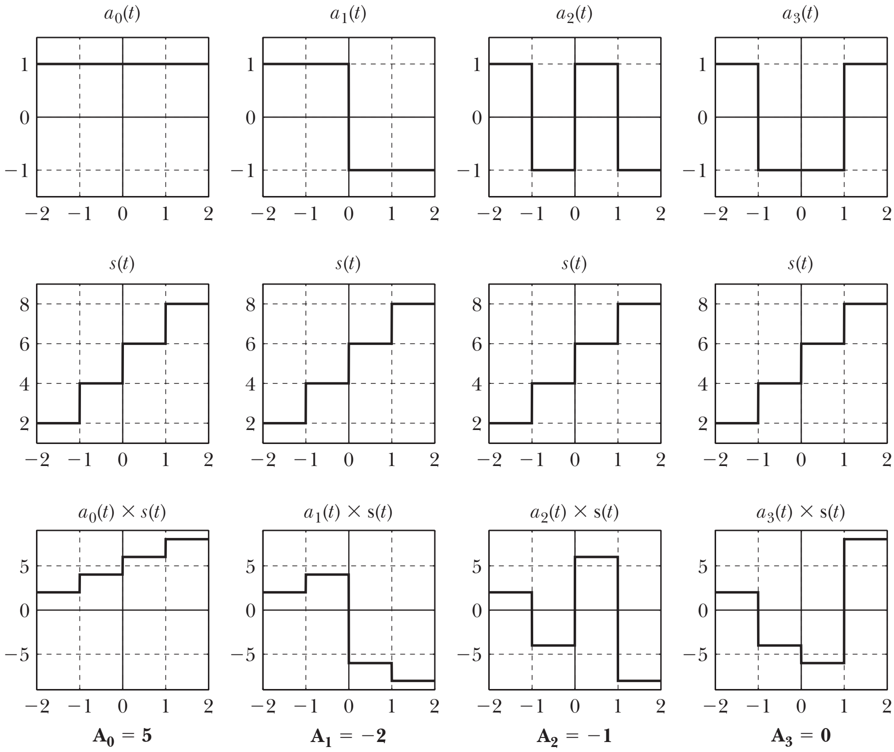



**FIGURE 4.10** The orthogonal signals $\{a_n(t), n = \{0, 1, 2, 3\}\}$ can approximate the signal $s(t)$ by graphically applying Eqn. 4.14 in each column. The components $\{A_n\}$ shown emerge after integrating the product term in the lower plots.

The $A_n$ terms yield the linear combination:

$$g(t) = 5a_0(t) - 2a_1(t) - 1a_2(t) + 0a_3(t)$$

which gives an exact answer $g(t) = s(t)$.

### EXAMPLE 4.4 (Orthogonal Components—A Best Answer)

*Find a best approximation to the signal* $s(t) = \cos((\pi/4)t)$ *defined in Figure 4.11 on the interval* $[-2, +2]$ *using components* $\{a_n(t), n = 0, 1, 2, 3\}$ *as shown.*

**Solution:**

Observe that the signal $s(t)$ is even; however, $a_1(t)$ and $a_2(t)$ are odd. Therefore, $A_1$ and $A_2$ will be zero by inspection.

A graphical application of the inner product yields the linear combination:

$$g(t) \approx 0.636a_0(t) + 0a_1(t) + 0a_2(t) - 0.236a_3(t)$$

By inspection, an exact answer is not possible, but the result shown dashed in Figure 4.12 is the best approximation given the available components.

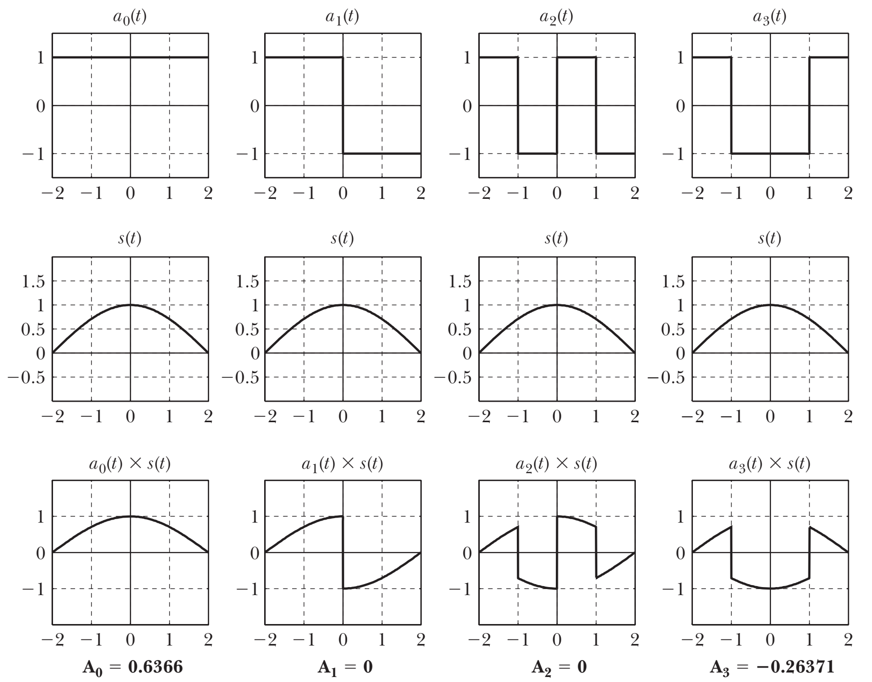



**FIGURE 4.11** The orthogonal signals $\{a_n(t),\ n = \{0, 1, 2, 3\}\}$ can approximate the signal $s(t)$ by graphically applying Eqn. 4.14 in each column. The components $\{A_n\}$ shown emerge after integrating the product term in the lower plots.

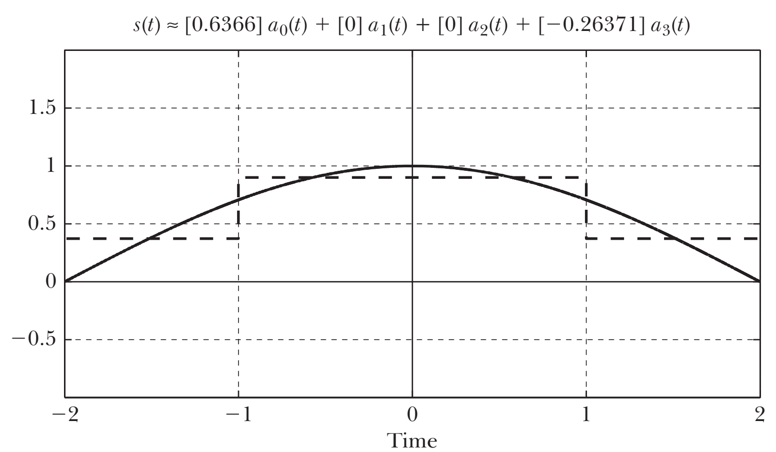



**FIGURE 4.12** The dashed line shows the best approximation to $s(t)$ from Figure 4.11 using orthogonal signals $\{a_n(t),\ n = \{0, 1, 2, 3\}\}$ on the interval $[-2, +2]$.

## 4.4.6 Orthogonal Signals and Linearly Independent Equations

An alternate (matrix) approach to solving for the orthogonal components of a signal can be applied to the staircase example of Figure 4.10. Note that this is not a general approach applicable to all orthogonal signals, but only in specific circumstances such as exist here.

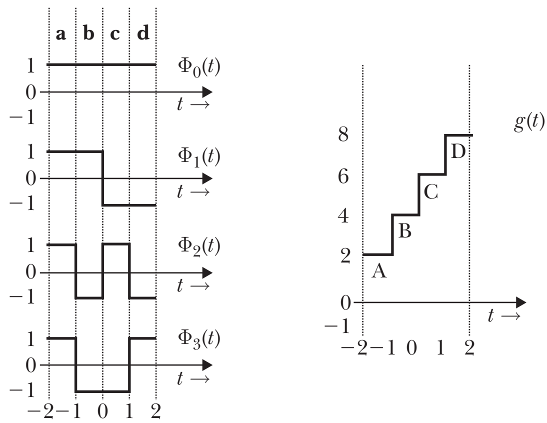



**FIGURE 4.13 Orthogonal Signals** Solve for $\{a, b, c, d\}$ to give $g(t) = a\Phi_0(t) + b\Phi_1(t) + c\Phi_2(t) + d\Phi_3(t)$.

Each of the orthogonal signals $a_n(t)$ has been divided into four regions. These regions correspond to the piecewise functional definitions. Now, the target function $g(t)$ is similarly defined on four regions, which have been labelled $[A, B, C, D]$ as in Figure 4.13, and which have the values [2, 4, 6, 8] in this example.

It can be seen that finding a linear combination of four signals to give $g(t)$ involves finding a solution to the equation;

$$g(t) \approx a\Phi_0(t) + b\Phi_1(t) + c\Phi_2(t) + d\Phi_3(t)$$

Given the four regions of $a_n(t)$, this reduces to:

$$\begin{aligned} a + b + c + d &= A \\ a + b - c - d &= B \\ a - b + c - d &= C \\ a - b - c + d &= D \end{aligned}$$

Or in a familiar matrix form:

$$\begin{bmatrix} +1 & +1 & +1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} \tag{4.15}$$

Now, with four equations and four unknowns, the above system of equations can always be solved if the equations are linearly independent. In other words, a solution exists if no equation can be expressed in terms of another. This condition is met if the matrix has a non-zero determinant. All these observations really mean *orthogonality* as it has been applied to signals.

Furthermore, it can be seen that if the above matrix has a non-zero determinant, then a solution can be found for any $\{A, B, C, D\}$. In other words, this formulation proves that for the limited class of functions described by four regions defined on the interval $[-2, +2]$, the signals $\{a_n(t),\ n = 0, 1, 2, 3\}$ form a complete set, able to represent any signal $g(t)$. This fact was illustrated when an exact answer for the stepped signal $g(t)$ shown here was found in an earlier example. To confirm this result, the matrix solution is readily found using MATLAB.

```
  x = [1 1 1 1 ; 1 1 -1 -1 ; 1 -1 1 -1 ; 1 -1 -1 1 ];
  inv(x) * [2;4;6;8]

ans =

     5
    -2
    -1
     0
```

In other words (as expected):

$$g(t) = 5\Phi_0(t) - 2\Phi_1(t) - 1\Phi_2(t) + 0\Phi_3(t)$$

**MATLAB Exercise 1: Evaluating an Inner Product** In this section, MATLAB is used to find the best approximation to the signal $g(t) = \cos(\pi/4t)$ in terms of the orthogonal set of signals $\{a_n(t)\}$ over the interval $[-2, +2]$.

Let the four basis signals be $\{a0, a1, a2, a3\}$ defined piecewise on the interval $[-2, +2]$. Since the signals are either $+1$ or $-1$ in each smaller interval, begin by defining a one-quarter interval of all ones, using 100 points for accuracy. Each signal $a_n$ will then be 400 points long.

```
a0 = [one  one  one  one ];
a1 = [one  one -one -one ];
a2 = [one -one  one -one ];
a3 = [one -one -one  one ];
t = linspace(-2,2,400); % interval [-2,2] is 400 pts
g = cos(pi/4 * t);      % define target signal g
```

Let the corresponding component of each basis signal be $\{A0, A1, A2, A3\}$, and evaluate the signal inner product for each component $A_n$.

```
A0 = sum(g .* a0) / sum(a0 .* a0); % DC: expect pi/4
A1 = sum(g .* a1) / sum(a1 .* a1); % odd: expect zero
A2 = sum(g .* a2) / sum(a2 .* a2); % odd: expect zero
A3 = sum(g .* a3) / sum(a3 .* a3);

[A0 A1 A2 A3 ]     % see 4 components

ans =
  0.6365   -0.0000   -0.0000   -0.2638

s = A0*a0 + A1*a1 + A2*a2 + A3*a3; % the approximation
plot(t, g, t, s);
```

These results for $\{A0, A1, A2, A3\}$ are consistent with the expected values as previously found by hand in Figure 4.11.

# 4.5 Part Two—The Fourier Series

The Fourier Series—The Orthogonal Signals $\{\sin(2\pi m f_0 t), \cos(2\pi n f_0 t)\}$

- An Orthogonal Set
- Computing Fourier Series Components
- Fourier Series Approximation to a Square Wave
- Fundamental Frequency Component
- Higher-Order Components
- Frequency Spectrum of the Square Wave
- Odd and Even Square Waves
- The Fourier Series Components of an Even Square Wave
- The Fourier Series Components of an Odd Square Wave
- Gibb's Phenomenon
- Setting up the Fourier Series Integration—A Pulse Train
- Some Common Fourier Series

## 4.5.1 A Special Set of Orthogonal Functions

The *Fourier series* is based on the set of orthogonal signals given by:

$$\{\sin(2\pi m f_0 t), \cos(2\pi n f_0 t)\}$$

for all integer $m$ and $n$, where $f_0$ is constant, and the set is defined on an interval corresponding to a complete period $T = 1/f_0$ seconds. These signals are an extension of the "cosine component" concept introduced in Chapter 1; however, cosines alone prove useful only as components of an even signal. It is straightforward to prove that these signals are orthogonal. The properties of orthogonal signals may then be used to derive the *Fourier series components* of an arbitrary periodic waveform.

For any periodic waveform $s(t)$ with period $T$, the first, best, approximation using sinusoidal components will logically begin with a sinusoid having frequency $f_0 = 1/T$ Hz. This frequency $f_0$ will be called the *fundamental frequency* of the waveform $s(t)$. The fundamental frequency also defines the Fourier series as a linear combination of orthogonal sinusoidal components leading to $s(t)$, as:

**DEFINITION 4.1 Fourier Series**

*If* s(t) *is a real periodic signal having fundamental frequency* $f_0$, *then*

$$s(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t)$$

*is called the* Fourier Series *representation of* s(t).

This linear combination involves an infinite series of sinusoidal components having frequencies at integer multiples of $f_0$. In terms of frequency, the cosine terms $A_n$ correspond exactly to components found in the one-sided cosine graph. A similar graph could be used to show the sine components $B_n$.

### 4.5.2 The Fourier Series—An Orthogonal Set?

Before any use can be made of the Fourier series, it should first be demonstrated that this is an orthogonal set of signals. This proof requires that the signal inner product be applied for all possible combinations of signals in the set. For convenience, choose an interval $[-T/2,\ T/2]$, although any interval spanning a full period $T = 1/f_0$ would be suitable.

The proof that the signals $\{\sin(2\pi nf_0 t), \cos(2\pi nf_0 t) \forall \{n,\ m\} \varepsilon I_0^+\}$ are mutually orthogonal in the interval $[-T/2, +T/2]$ can be accomplished in four steps, by proving in turn that:

1. Every sine term is orthogonal to every other sine term;
2. Every cosine term is orthogonal to every other cosine term;
3. Every sine term is orthogonal to every cosine term;
4. Every sine term and every cosine term is orthogonal to the DC term.

In every case, terms are orthogonal if their inner product is zero.

1. Consider the integral:

$$\int_{-T/2}^{+T/2} \sin(2\pi nf_0 t) \sin(2\pi mf_0 t)\, dt \tag{4.16}$$

$$= \int_{-T/2}^{+T/2} \frac{1}{2}[\cos(2\pi[n-m]f_0 t) - \cos(2\pi[n+m]f_0 t)]\, dt \tag{4.17}$$

$$= \frac{1}{2}\int_{-T/2}^{+T/2} \cos(2\pi[n-m]f_0 t)\ dt - \frac{1}{2}\int_{-T/2}^{+T/2} \cos(2\pi[n+m]f_0 t)\, dt \tag{4.18}$$

   Now, for all $n \neq m$, each of the final two integrals represent the area under a cosine over a complete period, and also for integer multiples $[m+n]$, $[n-m]$ of complete periods. By inspection, the integrals equal zero. When $m = n$, the term in $\cos(2\pi[n-m]f_0 t)$ equals 1, and, by inspection, this integral equals $T/2$.

2. The same logic applies to the integral below, and the result also equals zero for $n \neq m$, and $T/2$ when $m = n$.

$$\int_{-T/2}^{+T/2} \cos(2\pi nf_0 t) \cos(2\pi mf_0 t)\ dt \tag{4.19}$$

$$= \int_{-T/2}^{+T/2} \frac{1}{2}[\cos(2\pi[n-m]\, f_0 t) + \cos(2\pi[n+m]\, f_0 t)]\ dt \tag{4.20}$$

$$= \frac{1}{2}\int_{-T/2}^{+T/2} \cos(2\pi[n-m]\, f_0 t)\ dt + \frac{1}{2}\int_{-T/2}^{+T/2} \cos(2\pi[n+m]\, f_0 t)\ dt \tag{4.21}$$

3. The next integral is that of an odd product, and the result is zero, by inspection, for all $\{n,\ m\}$.

$$\int_{-T/2}^{+T/2} \sin(2\pi nf_0 t) \cos(2\pi mf_0 t)\ dt = 0 \tag{4.22}$$

4. Finally, the zero-frequency or DC component associated with the constant $A_0$ term must be shown to be orthogonal to all the sine and cosine terms in the Fourier series.

$$\int_{-T/2}^{+T/2} A_0 \cos(2\pi m f_0 t)\ dt = 0 \tag{4.23}$$

$$\int_{-T/2}^{+T/2} A_0 \sin(2\pi n f_0 t)\ dt = 0 \tag{4.24}$$

By inspection, these integrals are zero for all $n > 0$, as they all represent the area under a complete period of the sinusoids, and integer multiples thereof. When $n = 0$, the first integral equals $T$.

These relationships are of fundamental importance to application of the Fourier series. These results demonstrate that the sines and cosines described above, all integer multiples of a fundamental frequency $f_0$ Hz, are orthogonal over an interval equal to one period $T = 1/f_0$ seconds. This observation justifies the use of the Fourier series, not as some arbitrary choice of component signals, but as the decomposition of a periodic signal with period $1/f_0$ into orthogonal components. It would be satisfying to demonstrate that this is also a complete set of orthogonal signals; unfortunately, such a proof is beyond the scope of this discussion. On the other hand, it has been shown that even if a complete set was not present, the results obtained would nonetheless be a best approximation, given the available components.

Without further proof, the Fourier series is presented as a complete set of orthogonal functions. Consequently, any periodic function $s(t)$ will be expressible (to as much accuracy as is required) by a linear combination of these sines and cosines. It remains to discover exact values for the constants in the linear combination that will best approximate $s(t)$. The orthogonal Fourier series components of an arbitrary periodic waveform $s(t)$ with period $T = 1/f_0$ seconds may now be found by direct application of the signal inner product with each component term.

$$A_n = \frac{\int s(t) a_n(t)\, dt}{\int a_n(t) a_n(t)\, dt}$$

$$B_n = \frac{\int s(t) b_n(t)\, dt}{\int b_n(t) b_n(t)\, dt}$$

It can be seen that the denominator in each of these two equations equals $T/2$ for all $n > 0$. This constant term may simply be substituted in each of the above equations to give:

$$s(t) \approx A_0 + \sum_{n=1}^{\infty} [A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t)]$$

$$A_0 = \frac{1}{T} \int_{-T/2}^{+T/2} s(t)\ dt$$

$$A_n = \frac{2}{T} \int_{-T/2}^{+T/2} s(t) \cos(2\pi n f_0 t)\ dt$$

$$B_n = \frac{2}{T} \int_{-T/2}^{+T/2} s(t) \sin(2\pi n f_0 t)\ dt$$

# 4.6 Computing Fourier Series Components

The Fourier series is a linear combination of orthogonal sines and cosines that are integer multiples, or harmonics, of a fundamental frequency $f_0$. In addition, a zero-frequency (or DC) component $A_0$ acts to "raise or lower" the waveform with respect to the horizontal axis. Significantly, the presence of $A_0$ does not alter the shape or overall appearance of a waveform. Moreover, without the harmonics, the term $A_1$ represents the amplitude of a single cosine with period $T = 1/f_0$ seconds. The presence of harmonics and their relative amplitudes dictate the exact shape of a periodic waveform, as determined by the linear combination of those components. These components may be identified readily, in the distinctive frequency-domain sketch of Figure 4.14.

## 4.6.1 Fourier Series Approximation to an Odd Square Wave

A suitable linear combination of orthogonal sines and cosines will now be developed to approximate the square wave $s(t)$ shown in Figure 4.15, with period $T = 2\pi$ seconds. Which Fourier series components in what proportion will lead to a square wave shape? It is often said that *a square wave contains odd harmonics* (i.e., only odd-numbered multiples of $f_0$, such as $\{\cos(1t), \cos(3t), \cos(5t), \cos(7t)\cdots\}$), so this derivation may confirm that assertion.

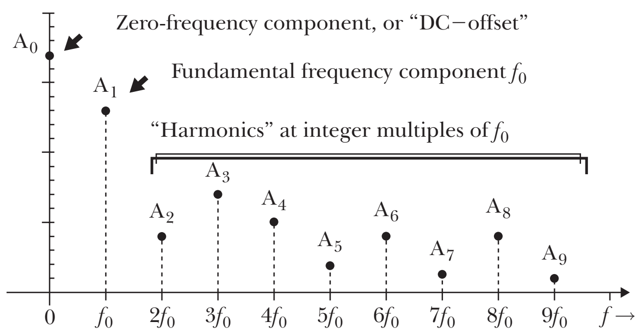



**FIGURE 4.14 Fourier Series Components** Representing a linear combination of a DC component plus non-zero frequency components at integer multiples of a fundamental frequency $f_0$ Hz.

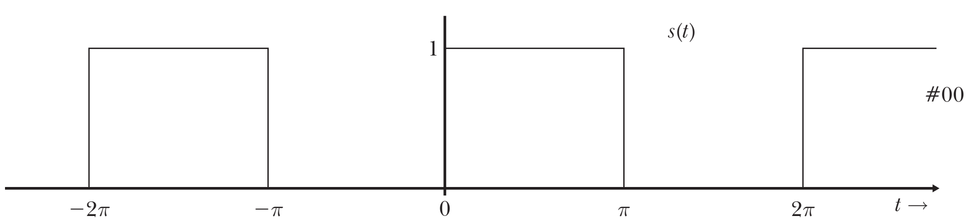



**FIGURE 4.15** What Fourier series components will best approximate this square wave?

## 4.6.2 Zero-Frequency (DC) Component

This square wave is distinguished from a symmetrical square wave varying from $-1/2 to + 1/2$ by being raised above the horizontal axis. "Adding a constant" is the effect of the orthogonal $A_0$ term (zero frequency or DC component) in the Fourier series. In other words, except for the addition of this constant value, the integral of the signal $s(t)$ over one complete period would equal zero. By inspection, $A_0 = 1/2$, in this symmetrical square wave. The signal inner product gives this same result:

$$\begin{aligned} A_0 &= \frac{1}{T}\int_{-T/2}^{+T/2} s(t)\ dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{+\pi} s(t)\ dt \\ &= \frac{1}{2\pi}\int_{o}^{+\pi} 1\ dt \\ &= \frac{1}{2\pi} \times \pi \\ &= \frac{1}{2} \end{aligned}$$

Recall that each orthogonal component in the Fourier series, considered separately, can be thought of as the "best approximation" to the desired waveform, given only that component. If the only component available was this DC component, the best approximation possible would logically be the line through the average value of the signal (the same $A_0 = 1/2$ as shown in Figure 4.16).

If the contribution from the constant $A_0$ was subtracted from $s(t)$, the remaining signal would lie on the horizontal axis, with a zero integral over one period. Furthermore, this signal would be an odd function. The original (neither) square wave signal $s(t)$ shown above was therefore the sum of an even (constant) term, plus other odd terms. Armed with this information, further calculations may be simplified by observing that there will be no cosine (even) components in the remaining signal. Therefore, by inspection:

$$A_n = 0 \text{ for all } n > 0$$

So far, the Fourier series approximation is:

$$s(t) \approx \frac{1}{2} + \text{odd components}$$

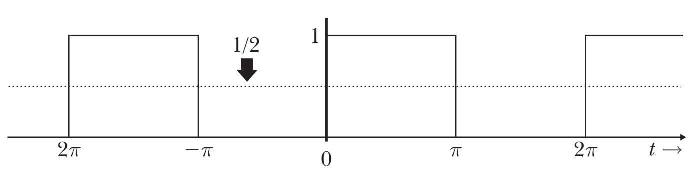



**FIGURE 4.16** The $A_0 = 1/2$ term is the DC component, or the average value of the signal $s(t)$.

# 4.7 Fundamental Frequency Component

The first Fourier series approximation to a periodic function is always a sinusoid ($\sin(2\pi f_0 t)$ and/or $\cos(2\pi f_0 t)$) with the same period $(1/f_0)$ as the signal. In the present example, only sine components are present in this odd square wave with period $2\pi$. (All $A_n = 0$.) The first approximation will use $B_1 \sin(2\pi f_0 t)$, with frequency $f_0 = 1/2\pi$. All the other components of the signal will be integer multiples (harmonics) of this fundamental frequency, leaving terms in $B_n \sin(nt)$. Each orthogonal component $B_n$ makes its own separate contribution to the waveform.

**A *Best-Fit* Approximation** Fitting a Fourier series to a periodic signal begins with the first best approximation using a single sinusoid with the same phase and period as the target waveform, after which successively finer adjustments are made using harmonic components, as required.

The first component of the square wave is given by:

$$
\begin{aligned}
B_1 &= \frac{2}{T}\int_{-T/2}^{+T/2} s(t)\sin(2\pi f_0 t)\,dt \\
&= \frac{2}{2\pi}\int_{-\pi}^{+\pi} s(t)\sin(t)\,dt \\
&= \frac{1}{\pi}\int_{0}^{+\pi} \sin(t)\,dt \\
&= -\frac{1}{\pi}\cos(t)\Big|_{t=0}^{\pi} \\
&= \frac{1}{\pi}(1+1) \\
&= \frac{2}{\pi}
\end{aligned}
$$

This result is seen in Figure 4.17. It appears to be a good fit that, with a few adjustments, will start to look more and more like a square wave. The approximation now stands at:

$$s(t) \approx \frac{1}{2} + \frac{2}{\pi}\sin(t)$$

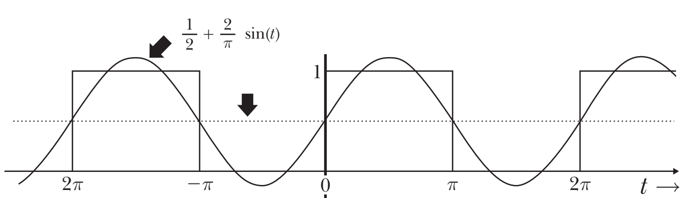



**FIGURE 4.17** The DC component plus the fundamental frequency component gives the first step in the Fourier series approximation to the square wave.

### 4.7.1 Higher-Order Components

Further components of $\sin(nt)$ found in this approximation to an odd square wave may be calculated generally as a function of $n$ for $n > 0$ using the interval $[-\pi, +\pi]$. Observe that $s(t)$ is 0 for $T < 0$ in this interval.

$$
\begin{aligned}
B_n &= \frac{2}{T}\int_{-T/2}^{+T/2} s(t)\sin(2\pi n f_0 t)\, dt \\
&= \frac{2}{2\pi}\int_{-\pi}^{+\pi} s(t)\sin(nt)\, dt \\
&= \frac{1}{\pi}\int_{0}^{+\pi} 1\sin(nt)\, dt \\
&= (1/n\pi)(-\cos(nt)|_0^{+\pi}) \\
&= (1/n\pi)[1-\cos(n\pi)]
\end{aligned}
$$

The resulting $B_n$ are found by substituting $n$ into this formula, as shown below, where the term $\cos(n\pi)$ is either $-1$ or $+1$ for even and odd values of $n$, respectively. Observe that the value for $B_1$ is the same as that found previously.

It is clear that only odd-numbered terms in $n$ will survive. This confirms the notion that square waves include only "odd harmonics." The relative amplitude of each component term varies as $1/n$. The resulting approximation becomes:

$$s(t) \approx \frac{1}{2} + \frac{2}{\pi}\left[\sin(t) + \frac{1}{3}\sin(3t) + \frac{1}{5}\sin(5t) + \frac{1}{7}\sin(7t)\cdots\right] \tag{4.25}$$

The result of adding each term $B_n \sin(nt)$ is shown in Figure 4.18 for $n = \{3, 5, 7\}$. Observe that the relative contribution of each component decreases as $1/n$.

The square wave could never be represented exactly by the Fourier series unless an infinite number of terms was available. As the number of terms grows, each new component included in the approximation brings a contribution decreasing as $1/n$. In practice, only a few components are usually required until the *best fit* obtained is acceptable. Knowing the relative contribution of each harmonic term makes possible certain analyses that are not evident in the time domain. It is clear from the sketches above that after only a few terms the square wave form is quite evident in the Fourier series approximation. The entire process is summarized in the following figure, where the desired square wave shape emerges as progressively more sinusoidal terms are contributed.

**TABLE 4.1**

**Fourier Series Components of an Odd Square Wave**

| $B_n$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ |
|---|---|---|---|---|---|---|---|
| $\frac{1}{n\pi}\cdot[1-\cos(n\pi)]$ | $\frac{2}{\pi}$ | 0 | $\frac{+1}{3}\cdot\frac{2}{\pi}$ | 0 | $\frac{+1}{5}\cdot\frac{2}{\pi}$ | 0 | $\frac{+1}{7}\cdot\frac{2}{\pi}$ |

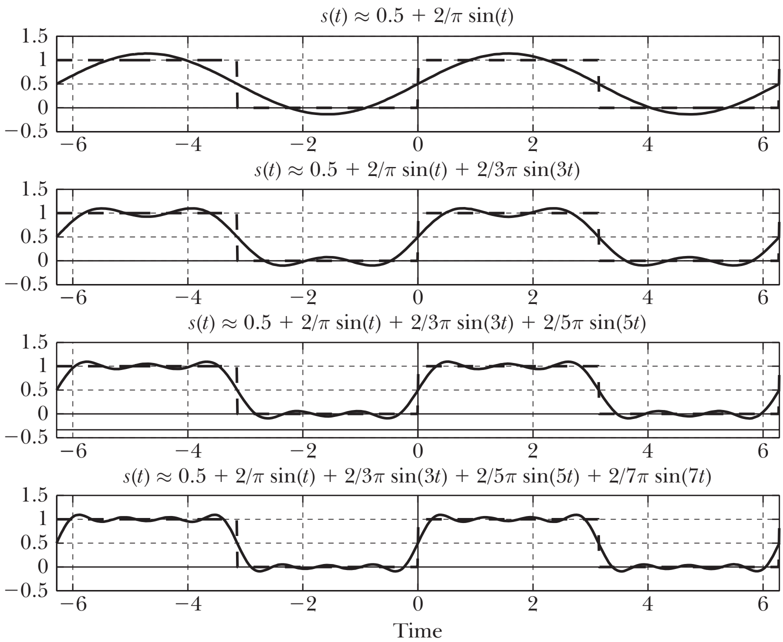



**FIGURE 4.18** Fourier series approximation for the DC component plus the first four frequency terms at ($f_0$, $3f_0$, $5f_0$, $7f_0$).

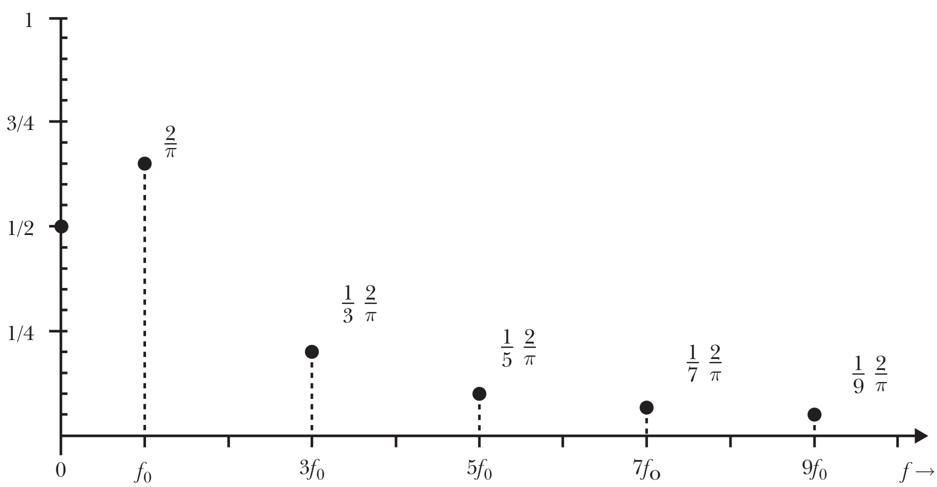



**FIGURE 4.19 One-Sided Sine Graph** Fourier series components of the odd square wave shown in frequency. The same information is presented here as would be found in the time-domain sketch of a square wave. The spectrum analyzer in Figure 4.4 is also showing a square wave.

## 4.7.2 Frequency Spectrum of the Square Wave *s(t)*

The frequency components of the odd square wave above have been calculated, and their addition was shown to give the expected square wave shape in the time domain. The same signal sketched in the frequency domain will show clearly the frequency components and their relative contributions. Figure 4.19 is a one-sided sine amplitude *vs.* frequency representation of the square wave $s(t)$, where the components are those computing using the Fourier series.

Note that the DC-offset, or *zero-frequency* component $A_0$, lies at $f = 0$ on the graph. It was seen in the time domain that this term has no effect on the overall shape of the waveform. However, given this graph alone, without any other knowledge of the time-domain sketches, the DC-offset, or *average value* or *area under one period* of the time-domain function can be obtained directly from this $f = 0$ component. The frequency-domain graph also shows that the largest contributor to this waveform is the fundamental component $f_0 = 1/2\pi$ Hz. The relative contributions of the other components fall off rapidly. Without these other components, the fundamental sine wave at $f_0 = 1/2\pi$ Hz would be just a sine wave. It is the harmonics that alter the form from being just a sine wave. No non-integer multiples of the fundamental component are present in a periodic waveform. (Orthogonality is defined over an interval [one complete period] and the Fourier series is based on integer multiples of $f_0$.)

## 4.8 Practical Harmonics

Harmonics are present whenever there is any deviation from an ideal sine wave; these are the orthogonal Fourier series components present in any non-sinusoidal periodic waveform. Conversely, the smallest amount of deviation from a pure sinusoid is evident by the presence of harmonics in the frequency spectrum. The examples in this section illustrate how harmonics can emerge in practical systems.

### 4.8.1 The 60 Hz Power Line

People in North American are surrounded by 60 Hz electromagnetic fields emanating from AC power lines; however, it is also true that significant energy can be found at 120 Hz, 180 Hz, 240 Hz, and so on. Why not a frequency component at 107 Hz or 99 Hz? First, 60 Hz alternating current is usually generated mechanically (regardless of whether the generators turn under the force of falling water, or from steam heated with coal, oil, or nuclear fuel). This would never produce a perfect sinusoidal waveform. Second, the effects of passing over the distribution network and into various loads inevitably leads to more distortion in the waveform finally reaching homes and businesses. The resulting 60 Hz signal (distorted, yet still periodic), could be represented by a Fourier series made up of a 60 Hz fundamental plus orthogonal harmonic components at integer multiples of 60 Hz as shown in Figure 4.20.

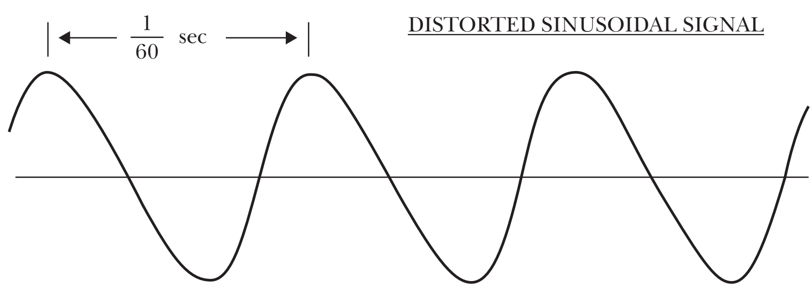



**FIGURE 4.20** A 60 Hz waveform that is not exactly sinusoidal will have a significant component at the fundamental frequency (60 Hz) as well as numerous smaller harmonic components at integer multiples of 60 Hz.

## 4.8.2 Audio Amplifier Specs—Total Harmonic Distortion

When shopping for a home entertainment system, the audio specifications often refer to *total harmonic distortion* (THD). This figure of merit refers to the relative harmonic content present at the output of the amplifier with respect to a pure sinusoidal test signal at the input as shown in Figure 4.21. Any harmonics present in the output signal indicate distortion with respect to the ideal sinusoid (fundamental) at the input. If THD = 0, a pure sinusoid at the input emerges as an amplified pure sinusoid at the output. (Unfortunately, only an ideal linear system can boast such performance.) Note that, in general, this test would have to be repeated for frequencies over the expected range of operation.

To describe exactly how much distortion is present after a sinusoid passes through some nonlinear system, the *distortion factor* is defined as:

$$\text{THD} = \sqrt{\frac{\text{power in harmonics}}{\text{power in original sinusoid}}} \tag{4.26}$$

## 4.8.3 The CB Radio Booster

Truck drivers using Citizen Band radios who hope to boost their signal range by adding an illegal high-powered amplifier may draw attention to themselves in unexpected ways; a study of the Fourier series shows why. Such radio transmitters operate on a fundamental frequency near 27 MHz and, as has been shown, slight distortions in the ideal transmitted frequency mean that a small amount of energy is always transmitted on harmonic frequencies, too. This is akin to transmitting identical (but weaker) copies of a signal on frequencies extending up through the radio spectrum (54, 81, 108, 135 MHz), well beyond any CB channels. Of course, when such a radio is designed to operate at the 4 watt legal limit, steps are taken in the factory so that the inevitable harmonics remain too small to bother anyone else. But, if a hapless operator adds an uncertified high-powered amplifier the amplification of existing (small) harmonics and the introduction of new (large) harmonic content may lead to complaints. For example, the harmonic at 54 MHz is very close to the transmitting frequency of television Channel 2. Similarly, a CB-radio operator operating at excessive power on CB Channel 10 (27.075 MHz) could block air traffic control communications on 135.375 MHz. Such cases of interference occur with surprising regularity.

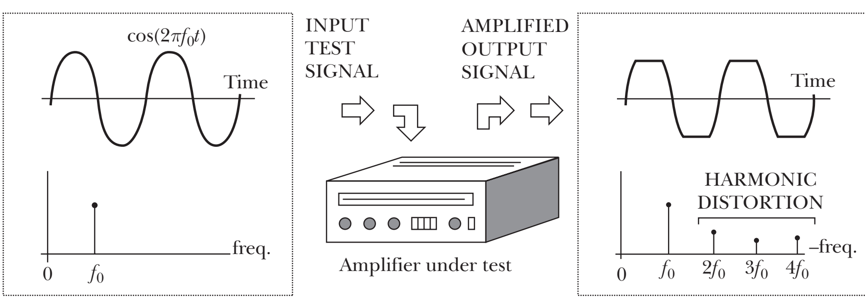



**FIGURE 4.21 Amplifier Test** Harmonics present in the output signal reflection distortion, with respect to a perfect sinusoidal input.

# 4.9 Odd and Even Square Waves

The computation of Fourier series components for an odd square wave signal $s(t)$ with period $T = 2\pi$ seconds reveals a DC offset, plus components of $\sin(nt)$. The presence of odd-numbered harmonics decreasing in amplitude as $1/n$ characterizes the square wave. This particular linear combination of orthogonal components of $\{\sin(2\pi n f_0 t)\}$ is a unique decomposition reflecting the square wave shape and amplitude. In other words, any deviation in the presence or proportion of any of these components will necessarily alter (distort) the waveform.

Now, consider an even square wave $g(t)$ with the same period $T = 2\pi$ as in Figure 4.22. The same sinusoidal components in the same proportion must be present for both the odd square wave and the even square wave. However, the even square components are cosines rather than sines. This difference also leads to a change in sign.

Figure 4.23 shows a square wave and the unique set of frequency components that add to give the square wave shape; the only difference between the odd and even square wave is the choice of origin. When the origin is at **B**, the odd square wave has sine components that add directly. When the origin shifts to **A** or **C** to give an even square wave, the corresponding cosine components *alternate in sign.*

## 4.9.1 The Fourier Series Components of an Even Square Wave

Components of $\cos(nt)$ found in the approximation to an even square wave may be calculated generally as a function of $n$ for all $n > 0$. Let the even square wave be $g(t)$, and compute the Fourier series in the interval $[-\pi, +\pi]$.

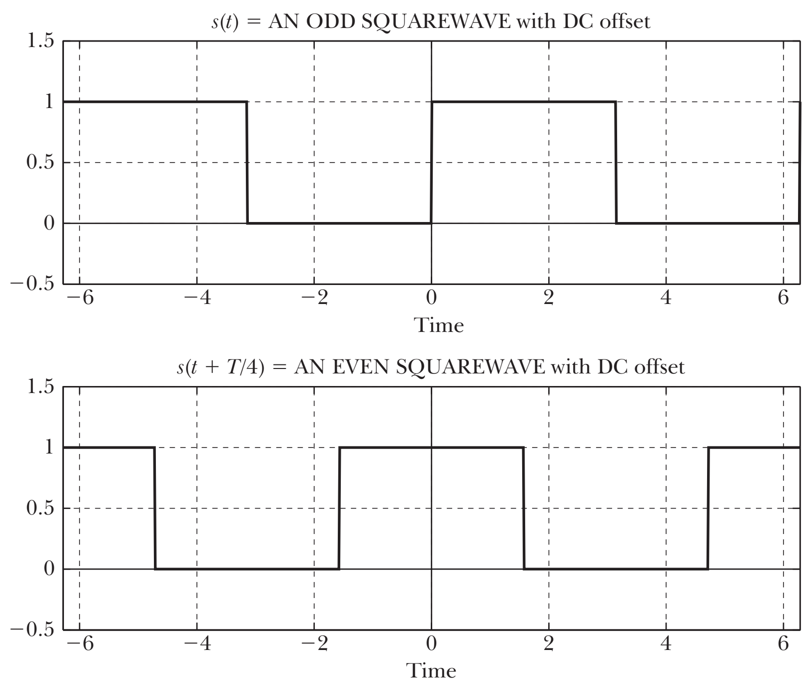



**FIGURE 4.22** Odd and even square waves differ only by a time shift. The period, amplitude, and overall appearance is unchanged. The two Fourier series must be very closely related.

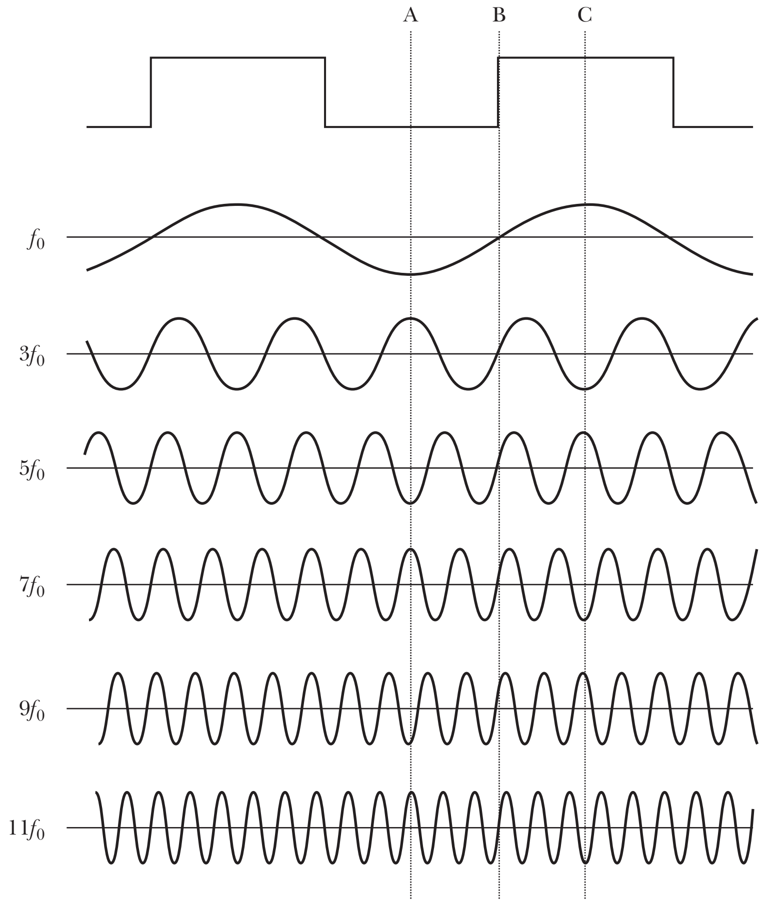



**FIGURE 4.23 Fourier series components of a square wave** For the origin at **B**, the components add as sines to form an odd square wave. For the origin at **C**, the components add as cosines to form an even square wave. Observe that the cosine terms alternate in sign for the even square wave.

$$
\begin{aligned}
A_n &= \frac{2}{T}\int_{-T/2}^{+T/2} g(t)\cos(2\pi n f_0 t)\,dt \\
&= \frac{2}{2\pi}\int_{-\pi}^{+\pi} g(t)\cos(nt)\,dt \\
&= \frac{1}{\pi}\int_{-\pi/2}^{+\pi/2} 1\cos(nt)\,dt \\
&= \frac{2}{\pi}\int_{0}^{+\pi/2} \cos(nt)\,dt \\
&= \frac{1}{n\pi}\sin(nt)\Big|_{t=0}^{+\pi/2} \\
&= \frac{1}{n\pi}\sin(n\pi/2)
\end{aligned}
$$

**TABLE 4.2**

**Fourier Series Components of an Even Square Wave**

| $A_n$ | $A_1$ | $A_2$ | $A_3$ | $A_4$ | $A_5$ | $A_6$ | $A_7$ |
|---|---|---|---|---|---|---|---|
| $\frac{2}{n\pi} \cdot \left[\sin(\frac{n\pi}{2})\right]$ | $\frac{2}{\pi}$ | 0 | $\frac{-1}{3} \cdot \frac{2}{\pi}$ | 0 | $\frac{+1}{5} \cdot \frac{2}{\pi}$ | 0 | $\frac{-1}{7} \cdot \frac{2}{\pi}$ |

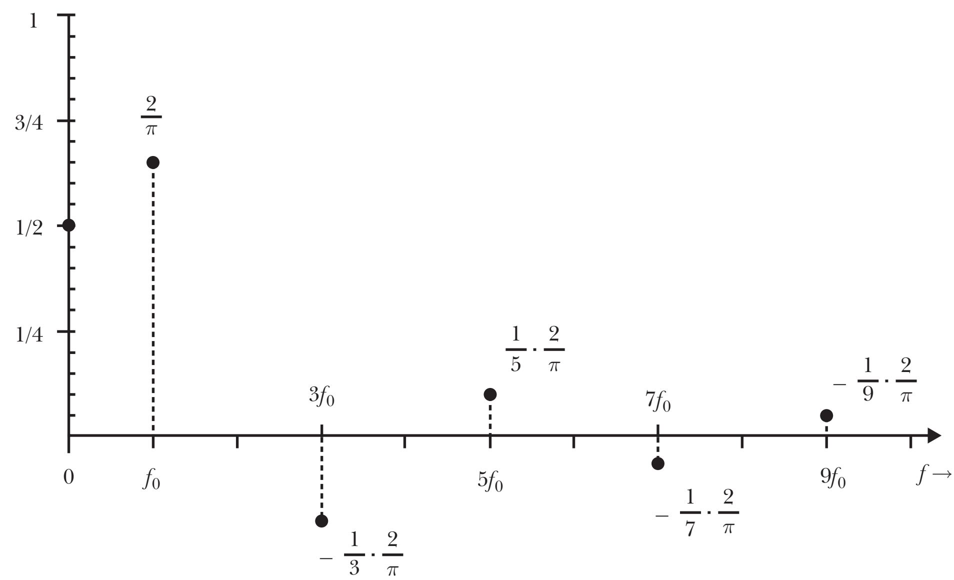



**FIGURE 4.24 One-Sided Cosine Graph** Fourier series components of the even square wave shown in frequency. The same components in the same proportions are present in the odd square wave of Figure 4.19 and differ only in sign.

It is observed that $g(t)$ is zero for $|t| > \pi/2$ in this interval. The inner product equation is further simplified by observing that the product $\cos(nt)$ is even and using the rule from Eqn. 2.1.

Now, $\sin(n\pi/2)$ will be $(0, +1, 0, -1)$ for successive values of integer $n$, leaving only odd-numbered terms, with the each of those terms alternating in sign, as shown in the Table 4.2. These values of $A_n$ confirm the results obtained above by explicitly shifting the waveform.

The corresponding frequency-domain representation of this even square wave is shown in Figure 4.24:

$$s(t) \approx \frac{1}{2} + \frac{2}{\pi}\left[\cos(t) - \frac{1}{3}\cos(3t) + \frac{1}{5}\cos(5t) - \frac{1}{7}\cos(7t) + \cdots\right] \tag{4.27}$$

Note that, with the exception of the alternating sign in adjacent components, the magnitude and position of each component are identical to that of the odd square wave previously examined.[2]

[2] In fact, a spectrum analyzer shows only the magnitude (not the sign) of each component, so there would be no difference at all in the laboratory between an even or odd signal (just as with an oscilloscope).

# 4.10 Gibb's Phenomenon

The derivation of the Fourier series components includes components up to infinity, and an infinite number of components would be necessary to represent a square wave exactly. As each component is added, the "best approximation" grows better although, in practice, it would be impossible to combine terms all the way to infinity. In any case, the relative contribution of each new term grows exceedingly small as the number of components grows (varying as $1/n$).

A close examination of the Fourier series expansion of a square wave reveals some unusual behavior near discontinuities. This effect is shown exaggerated in the sketch below. It happens that small *overshoots* that seem inconsequential for the first few terms never really disappear, even as $n$ goes to infinity. This effect is called *Gibb's phenomenon* and is not limited to square wave signals; the Fourier series approximation to any discontinuous function will exhibit similar behavior. This phenomenon will now be analyzed in detail through examination of the Fourier series expansion of a square wave.

The presence of Gibb's phenomenon is evident even after a few terms, and is found by examining the position and magnitude of the overshoot corresponding to the first maximum in the odd square wave expansion.

Consider a square wave $s(t)$ with period $2\pi$ seconds and amplitude $A = 1$ and no DC offset, then the Fourier series approximation is:

$$s(t) \approx \frac{4}{\pi}\left[\cos(t) - \frac{1}{3}\cos(3t) + \frac{1}{5}\cos(5t)\cdots\right] \tag{4.28}$$

Of course, the oscillating waveform has multiple extrema, all of which would give a zero derivative, but the point of interest here is the one closest to $t = 0$, or the first maximum. The time derivative of the Fourier series can be used to examine the behavior of this overshoot by setting the derivative to zero and, by inspection, selecting the smallest value of $t$ that satisfies the equation.

$$\frac{d}{dt}\frac{4}{\pi}\left[\cos(t) - \frac{1}{3}\cos(3t) + \frac{1}{5}\cos(5t)\ldots\right] = 0 \text{ for a maximum} \tag{4.29}$$

Specifically, after $n$ terms, this derivative is zero for $t = (\pi - \pi/2n)$, where the peak value of the approximation is 1.1813 for $n = 5$ and 1.1789 for $N = 100$ as shown in Figure 4.25.

The essence of Gibb's phenomenon is that this overshoot approaches the square wave edge but never goes to zero. Instead, its peak converges to about 18 percent of the square wave amplitude or $1.1789798\cdots$ as $n$ goes to infinity.

While the position of the overshoot approaches the discontinuity as $n$ grows, its amplitude converges to a constant. Therefore, the total squared area under the overshoot grows smaller with increasing $n$, consistent with the best approximation premise that the error diminishes as more terms are added. So, although Gibb's phenomenon is a real effect, of particular interest if a numerical calculation happens to fall on this point, it does not contradict the Fourier series. This issue is to be examined again in the context of convolution and the Fourier transform.

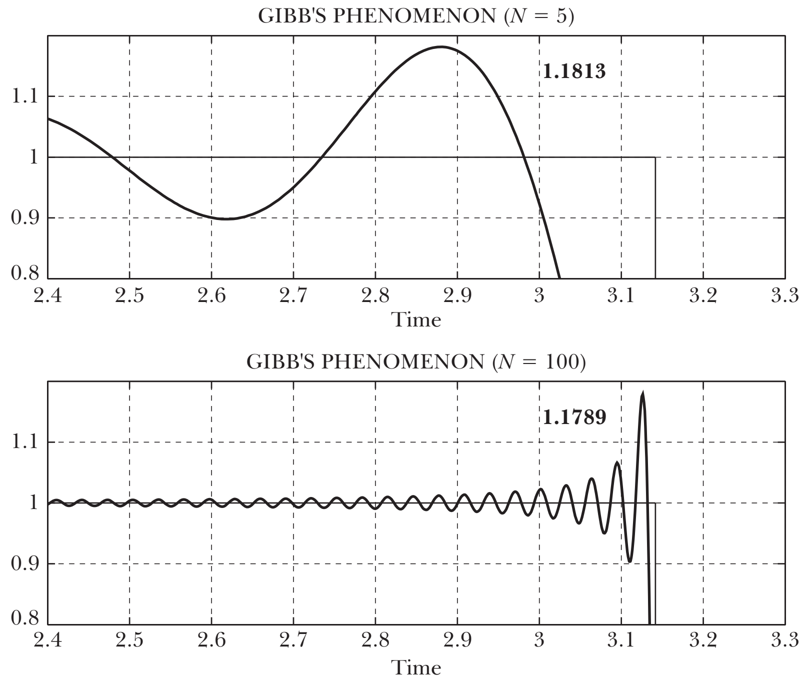



**FIGURE 4.25 Gibb's Phenomenon** While the Fourier series approximation to a square wave gets better with the addition of components, the overshoot at the discontinuity converges to 1.1789798 as it approaches the edge.

## 4.11 Setting Up the Fourier Series Calculation

When calculating the Fourier series components of a periodic signal, it is necessary to set up the inner product integrals that lead to the Fourier series coefficients $\{A_n,B_n\}$. These integrals can be solved with much less effort if some basic assumptions are made about the periodic signal $s(t)$ to be analyzed.

1. Sketch the signal $s(t)$ to be even or odd, if possible—Since only pure sine $\{B_n\}$ or cosine $\{A_n\}$ terms contribute to an odd or an even signal, respectively, this simple expedient will avoid half the mathematics by setting the other terms to zero, by inspection.
2. Select a period of $2\pi$ seconds, centered on the origin—While the Fourier series components can be computed over any complete period $(1/f_0)$ of a periodic signal $s(t)$, it is useful to use a period $2\pi$ seconds, as the terms in $2\pi nf_0t$ will then simply reduce to terms in $nt$. Furthermore, using the period centered on the origin ensures the maximum benefit from using odd or even signals.

**EXAMPLE 4.5 (Fourier Series of a Pulse Train)**

*Find the Fourier series of a 1 V pulse train with a 10 percent duty cycle and a period of 1 ms. In other words, this signal consists of rectangular 1 V pulses 100 $\mu$s long repeating 1000 times per second.*

**Solution:**
A general solution to this problem uses an even pulse width of $2a$ seconds and a period of $2\pi$ seconds, as shown in Figure 4.26. The value of $a$ can later be adjusted for any desired duty cycle. Similarly, the frequency components located at multiples of $f_0 = 1/2\pi$ Hz need only to be re-labelled for the desired $f_0 = 1000$ Hz.

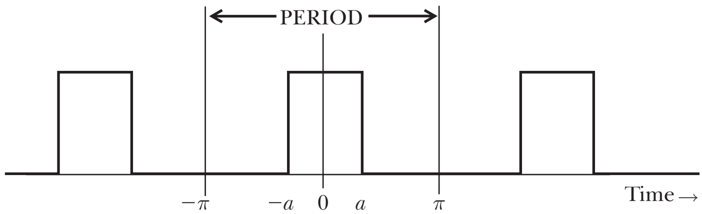



**FIGURE 4.26** A pulse train sketched as an even function with period=2 $\pi$ s and pulse width $2a$ sec.

The odd Fourier series components are found by inspection as:

$$B_n = 0, \quad \text{since } s(t) \text{ is even}$$

The even Fourier series components are given by:

$$A_n = \frac{2}{T} \int_{-T/2}^{+T/2} s(t) \cos(2\pi n f_0 t)\, dt$$

Substituting the period $T = 2\pi$ seconds gives:

$$A_n = \frac{1}{\pi} \int_{-\pi}^{+\pi} s(t) \cos(nt)\, dt$$

Since $s(t)$ is even, the integral simplifies to:

$$A_n = \frac{2}{\pi} \int_{0}^{+\pi} s(t) \cos(nt)\, dt$$

and $s(t) = 1$ only for $|t| < a$:

$$A_n = \frac{2}{\pi} \int_{0}^{+a} \cos(nt)\, dt$$

which can finally be integrated to yield the desired terms:

$$A_n = \frac{2}{\pi} \frac{1}{n} \sin(nt) \Big|_{t=0}^{a} = \frac{2}{\pi} \frac{1}{n} \sin(na)$$

In this final expression, each $A_n$ represents the magnitude of a frequency term $\cos(nt)$ at an integer multiple $nf_0$ of the fundamental frequency. The amplitude of successive terms decrease as $1/n$. This expression is valid for valid for $0 < a \leq \pi$.

## 4.11.1 Appearance of Pulse Train Frequency Components

The previous equation describing the cosine components $A_n$ approximating an even pulse train can be used to explore the behavior of the Fourier series expansion for different pulse widths. Pulse width is expressed as *duty cycle*, or percentage of the overall period.

**Pulse Train with 10 Percent Duty Cycle** For a 10 percent duty cycle, the above pulse train has $a = 0.10\pi$, giving Fourier series components:

$$A_n = \frac{2}{\pi}\frac{1}{n}\sin(n0.10\pi)$$

The sine term is zero whenever the argument equals an integer multiple of $\pi$. Therefore, every tenth frequency term will be zero as seen in Figure 4.27, where the result is shown for each component $n$ up to 20.

**Pulse Train with 20 Percent Duty Cycle** For a 20 percent duty cycle, the pulse train has $a = 0.20\pi$, giving:

$$A_n = \frac{2}{\pi}\frac{1}{n}\sin(n\,0.20\pi)$$

In this case, every fifth frequency term will be zero as seen in Figure 4.28. The spacing between frequency components does not change, as the fundamental frequency is unchanged.

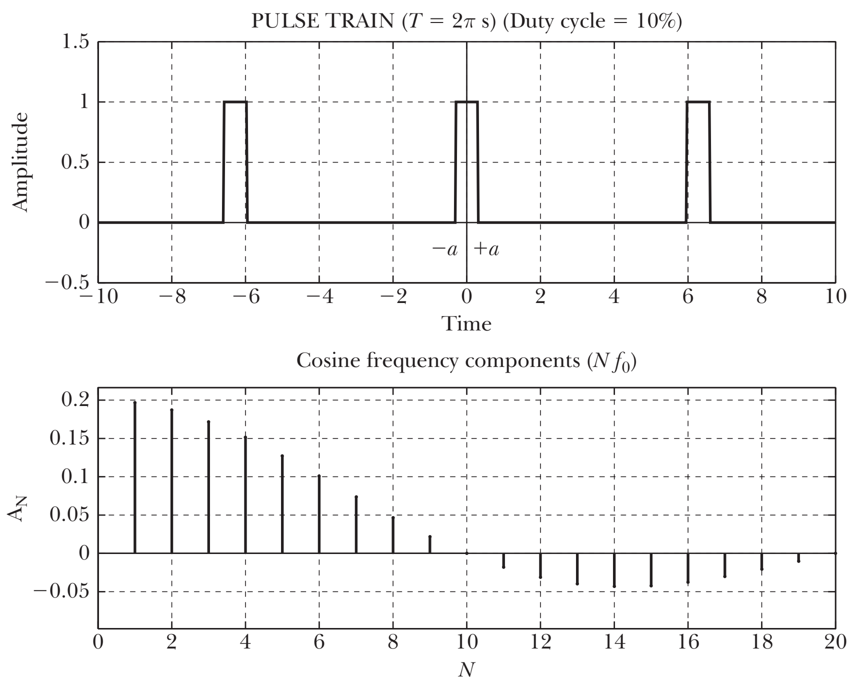



**FIGURE 4.27 Fourier Series Components** An even pulse train with 10 percent duty cycle. Every tenth frequency component is zero. Components are spaced at integer multiples of the fundamental frequency ($f_0 = 1/2\pi$).

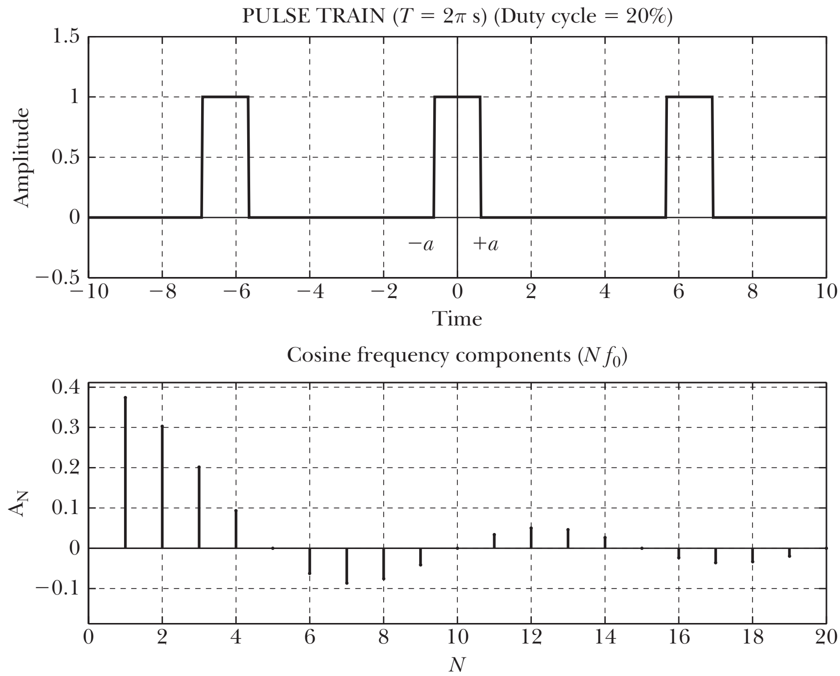



**FIGURE 4.28 Fourier Series Components** An even pulse train with 20 percent duty cycle. Every fifth frequency component is zero. Components are spaced at integer multiples of the fundamental frequency ($f_0 = 1/2\pi$).

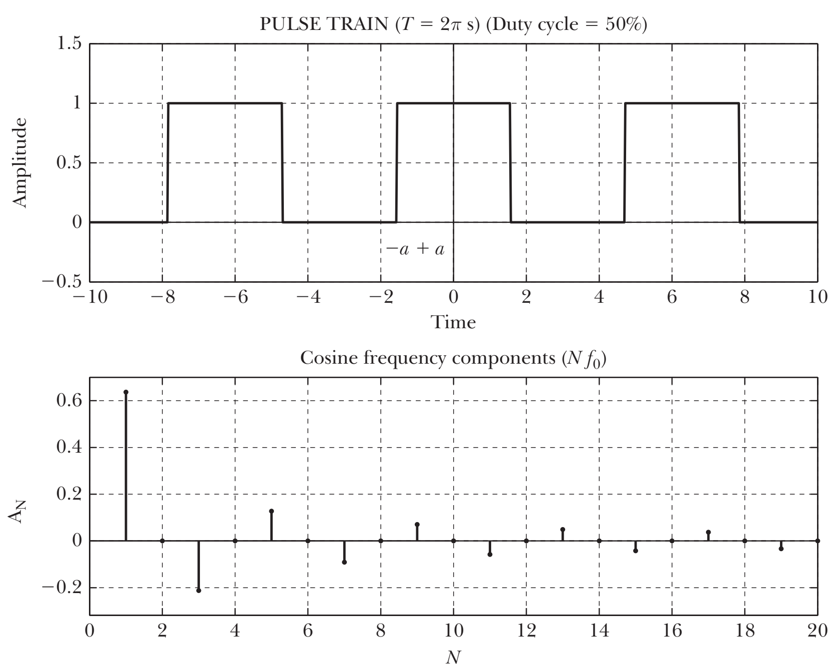



**FIGURE 4.29 Fourier Series Components** The even pulse train with 50 percent duty cycle is a square wave signal. Every second frequency component is zero (only odd-numbered components). Components are spaced at integer multiples of the fundamental frequency ($f_0 = 1/2\pi$). Compare to Figure 4.27.

**Pulse Train with 50 Percent Duty Cycle (Square Wave)** When $a = \pi/2$ (50 percent duty cycle, or pulse width equal to one-half period), the previous result should give the now-familiar Fourier series terms for an even square wave as seen in Figure 4.29. This is a satisfying confirmation of the validity of the pulse train Fourier series approximation. In the first equation of Section 4.12, it can be noted that $\sin(n\pi/2)$ is zero for all even integers $n > 0$, and equals 1 for all odd integers $n > 0$.

This exercise has shown that the Fourier series for any periodic waveform gives the sinusoidal component present at each multiple of the fundamental frequency. In the previous examples, the line spacing in Hz would vary with the overall period of the periodic pulse waveform, but the overall appearance of the frequency components (the zero crossings) is inversely proportional to the pulse width. The square wave emerged as a special case of the pulse train, having a 50 percent duty cycle.

## 4.12 Some Common Fourier Series

It is useful to recognize the commonly encountered Fourier series below. Except as noted, each is shown in simplified form with a period of $2\pi$ seconds, no DC offset, and an amplitude of 1 Volt.

1. **Square Wave (odd):** Only odd-multiple frequency terms, with amplitudes varying as $1/n$:

$$s(t) = \frac{4}{\pi}\left(\frac{\sin(t)}{1} + \frac{\sin(3t)}{3} + \frac{\sin(5t)}{5} + \frac{\sin(7t)}{7} + \cdots\right)$$

2. **Square Wave (even):** Only odd-multiple frequency terms, with amplitudes varying as $1/n$ and alternating in sign:

$$s(t) = \frac{4}{\pi}\left(\frac{\cos(t)}{1} - \frac{\cos(3t)}{3} + \frac{\cos(5t)}{5} - \frac{\cos(7t)}{7} + \cdots\right)$$

3. **Triangle Wave (even):** Only odd-multiple frequency terms, with amplitudes varying as $1/n^2$:

$$s(t) = \frac{8}{\pi^2}\left(\frac{\cos(t)}{1^2} + \frac{\cos(3t)}{3^2} + \frac{\cos(5t)}{5^2} + \frac{\cos(7t)}{7^2} + \cdots\right)$$

4. **Sawtooth (odd):** Every frequency term, with amplitudes varying as $1/n$:

$$s(t) = \frac{2}{\pi}\left(\frac{\sin(t)}{1} + \frac{\sin(2t)}{2} + \frac{\sin(3t)}{3} + \frac{\sin(4t)}{4} + \cdots\right)$$

5. **Comb Function (even):** Every cosine frequency term, equally weighted:

$$s(t) = 1 + 2(\cos(t) + \cos(2t) + \cos(3t) + \cos(4t) + \cos(5t) + \cdots)$$

6. **Half-wave Rectified Cosine (even):** Includes the DC offset associated with rectification:

$$s(t) = \frac{2}{\pi} + \frac{4}{\pi}\left(\frac{\cos(2t)}{1\cdot 3} - \frac{\cos(4t)}{3\cdot 5} + \frac{\cos(6t)}{5\cdot 7} - \frac{\cos(8t)}{7\cdot 9} + \cdots\right)$$

# 4.13 Part Three—The Complex Fourier Series

- The Complex Fourier Series—The Orthogonal Signals $e^{j2\pi nf_0 t}$
  - Must all signals be even or odd?
  - Relationship between the Real and Complex Fourier Series—Relating $(A_n, B_n)$ and $C_n$
  - Magnitude and Phase in Complex Fourier Series Components
  - Computing Complex Fourier Components
  - Real Signals and the Complex Fourier Series
  - Amplifying, Shifting, Scaling $s(t)$
  - Power in Periodic Signals: Parseval's Theorem
  - Review of Properties: The Time and Frequency Domains
- Using the Fourier Series—Study of a DC Power Supply
  - Where's the DC?
  - An AC-DC Converter
  - $V_{rms}$ is always greater than or equal to $V_{dc}$
  - Fourier Series Approximation: The Full-Wave Rectifier
  - Power in the Fundamental Frequency Component
- Finding Complex Fourier Series Components using MATLAB

## 4.13.1 Not All Signals Are Even or Odd

Nearly every example shown so far has included a purely even or a purely odd signal. This approach simplifies computations because odd or even signals have strictly odd or strictly even components. On the other hand, practical signals formed by a linear combination including both even and odd components will be neither even nor odd. It follows that any periodic signal that is not strictly even or strictly odd must have *both* sine and cosine components in the Fourier series.

The general cosine described by $s(t) = \cos(2\pi f_0 t + \Phi)$ incorporates odd and even signals by varying $\Phi$. To see the effect of $\Phi$, the trigonometric identity $\cos(x + y) = \cos x \cos y - \sin x \sin y$ can be applied with $x = 2\pi f_0 t$ and $y = \Phi$ to obtain:

$$s(t) = \cos(2\pi f_0 t + \Phi) = \cos(\Phi)\cos(2\pi f_0 t) - \sin(\Phi)\sin(2\pi f_0 t) \tag{4.30}$$

Note that for $\Phi = 0$, $s(t) = \cos(2\pi f_0 t)$, while for $\Phi = \pi/2$, $s(t) = \sin(2\pi f_0 t)$. For most values of $\Phi$, $s(t)$ will be neither even nor odd.

This expression describes a phase-shifted sinusoid shift by $\Phi$ rad as a linear combination of both sines and cosines of frequency $f_0$. This result can be identified as a Fourier series expansion of the form:

$$s(t) = \cos(2\pi f_0 t + \Phi) \approx \sum_{n=-\infty}^{+\infty} A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t) \tag{4.31}$$

with the non-zero coefficients $A_1 = \cos(\Phi)$, and $B_1 = -\sin(\Phi)$.

This effect is shown in Figure 4.30 for $A_1 = B_1 = 1$. Given only the Fourier series components $A_1$ and $B_1$, the phase shift $\Phi$ may be found directly by observing:

$$\Phi = \tan^{-1}\left[\frac{-B_1}{A_1}\right]$$

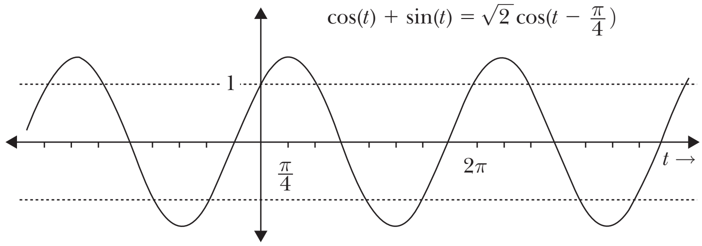



**FIGURE 4.30** A phase-shifted sinusoid has both odd and even sinusoidal components.

and the amplitude is constant independent of $\Phi$, given by:

$$\text{amplitude} = \sqrt{A_1^2 + B_1^2}$$

The amplitude of a shifted sinusoid is unaffected by any phase change.

# 4.14 The Complex Fourier Series

Consider the observation that real-world signals are unlikely to be exactly even or odd signals. Add to this the fact that representation of signals that are neither odd nor even requires computation of both the sine and cosine terms in the Fourier series. Think of the problem of combining cosine and sine representations in a single graphical form, or in a single expression. It is convenient to represent these components in one expression, and the complex exponential will serve to incorporate both as:

$$e^{j2\pi f_0 t} = \cos(2\pi f_0 t) + j\sin(2\pi f_0 t) \tag{4.32}$$

which has the additional advantage of being able to represent complex signals as required. The use of this exponential as a mathematical tool requires that real sine components will be manipulated as imaginary terms; however, the final results (input and output signals) must always be real in the laboratory.

To make use of this form, the *complex Fourier series* is defined to express the Fourier series of either odd or even or neither (or complex) signals $s(t)$ in terms of the above exponential.

**DEFINITION 4.2 Complex Fourier Series**

*A periodic signal* $\mathrm{s}(t)$ *with period* $\mathrm{T} = 1/\mathrm{f}_0$ *can be represented by the terms* $\mathrm{C_n}$ *in:*

$$s(t) = \sum_{n=-\infty}^{\infty} C_n e^{j2\pi n f_0 t}$$

*which is called the* complex Fourier series.

Here, the constants $C_n$ are the (complex) components of the periodic signal $s(t)$. Note that this summation extends for positive and negative integers in the interval $[-\infty, +\infty]$. Consequently, this expression incorporates:

1. for $n < 0$, a linear combination of components of $C_{-n}e^{-j2\pi n f_0 t}$, and;
2. for $n > 0$, a linear combination of components of $C_{+n}e^{+j2\pi n f_0 t}$, and;
3. for $n = 0$, the DC component $C_0 e^0 = C_0$.

## 4.14.1 Complex Fourier Series—The Frequency Domain

A graph showing complex Fourier series components will necessarily include both positive and negative values of $n$. The *two-sided complex exponential* graph in Figure 4.31 shows both positive and negative components ($C_n$) of $e^{+j2\pi nf_0t}$ as a function of frequency. Note that does not mean that the graph shows negative frequencies found in a signal $s(t)$. Rather, component $C_n$ is a single value of $e^{+j2\pi nf_0t}$, for either positive or negative N, as shown in Figure 4.32. This representation of the periodic signal $s(t)$ will simply be called *the frequency domain.*

Taken in pairs, components of the complex exponential can represent either sines or cosines. For a real cosine, use the identity:

$$s(t) = \cos(2\pi f_0 t) = \frac{e^{+j2\pi f_0 t} + e^{-j2\pi f_0 t}}{2}$$

where there are two components, each with amplitude 1/2:

$$s(t) = \cos(2\pi f_0 t) = C_{+1}e^{+j2\pi f_0 t} - C_{-1}e^{-j2\pi f_0 t} = \frac{1}{2}e^{+j2\pi f_0 t} + \frac{1}{2}e^{-j2\pi f_0 t}$$

as shown in Figure 4.33.

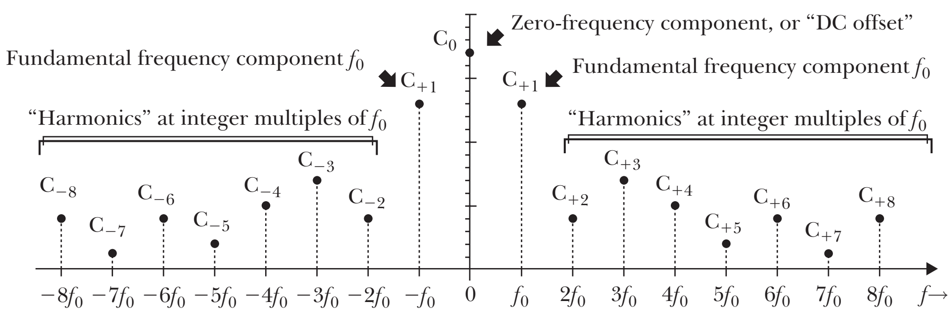



**FIGURE 4.31 The Frequency Domain** A graph of the complex Fourier series shows components of $e^{j2\pi nf_0t}$ for all values of $n$, both positive and negative.

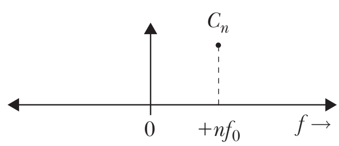



**FIGURE 4.32 The Frequency Domain** A single component $C_n$ of $e^{+j2\pi nf_0t}$, located at $nf_0$. The *two-sided* graph implies that $n$ could be positive or negative.

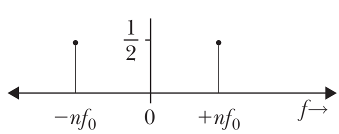



**FIGURE 4.33** The two-sided complex exponential graph of $s(t) = \cos(2\pi nf_0 t)$ is even and real. This distinctive two-component spectrum characterizes a time-domain cosine, where each component is one half the cosine amplitude.

Similarly, for a real sine:

$$s(t) = \sin(2\pi f_0 t) = \frac{e^{+j2\pi f_0 t} - e^{-j2\pi f_0 t}}{2j}$$

where there are two imaginary components, each with amplitude 1/2:

$$s(t) = \sin(2\pi f_0 t) = \frac{1}{2j} e^{+j2\pi f_0 t} - \frac{1}{2j} e^{-j2\pi f_0 t}$$

Recalling the complex identity: $-j = 1/j$, then $C_{-1}$ is positive, while $C_{+1}$ is negative, as:

$$s(t) = \sin(2\pi f_0 t) = C_{+1} e^{+j2\pi f_0 t} - C_{-1} e^{-j2\pi f_0 t} = \frac{-j}{2} e^{+j2\pi f_0 t} + \frac{+j}{2} e^{-j2\pi f_0 t}$$

as shown in Figure 4.34.

Using the complex Fourier series means that there is no longer any need for separate *one-sided cosine* or *one-sided sine* graphs, which are best suited to purely even or purely odd signals, because complete phase information can now be incorporated into one series approximation. Sine and cosine components can be shown on the same graph by superimposing the real and imaginary parts. To accomplish this, imaginary components can be distinguished by being dotted, or in a different color. Figure 4.35 shows the frequency components of the signal $s(t) = \sin(t) + \cos(t)$.

Although this graphical form is frequently encountered, it fails to illustrate the significance of the complex exponential format. Showing orthogonal odd and even components graphically in complex form requires an extra dimension, which is difficult to show on a flat page. Consider the same signal in the three-dimensional illustration of Figure 4.36, where the imaginary plane lies orthogonal to the real plane.

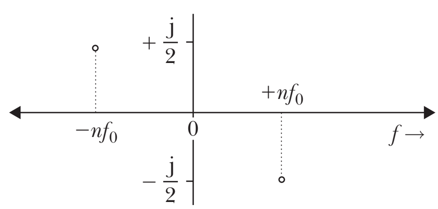



**FIGURE 4.34** The two-sided complex exponential graph of $s(t) = \sin(2\pi n f_0 t)$ is odd and imaginary.

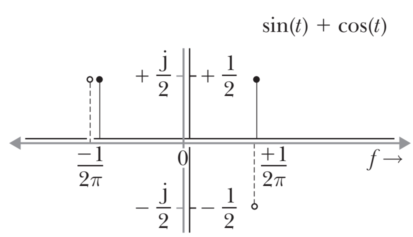



**FIGURE 4.35** Two-sided complex exponential graph of $s(t) = \sin(t) + \cos(t)$. The real components are even, while the imaginary components (dashed lines) are odd. This result is the sum of Figure 4.33 and Figure 4.34, shown here slightly offset for clarity.

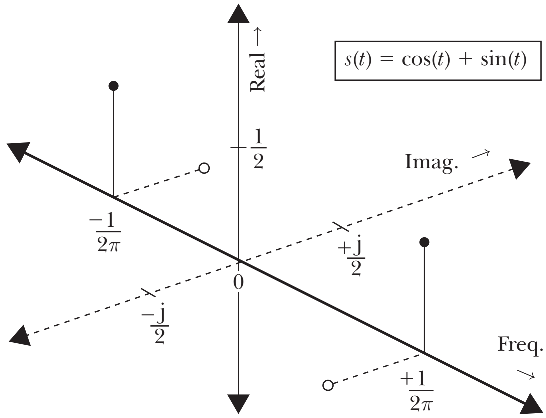



**FIGURE 4.36** Three-dimensional two-sided complex exponential graph of $s(t) = \sin(t) + \cos(t)$.

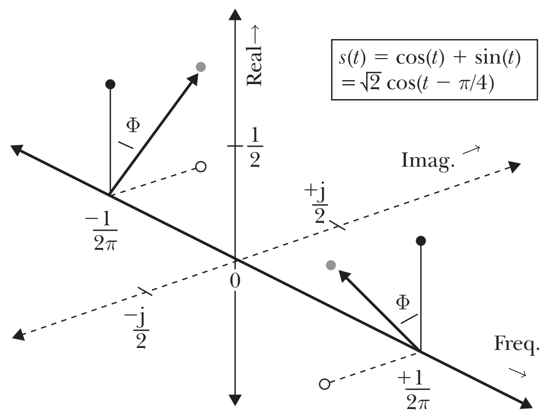



**FIGURE 4.37** Three-dimensional two-sided complex exponential graph of $s(t) = \sin(t) + \cos(t)$. The resulting vector sum is a cosine as $A\cos(t + \Phi)$ where A = $\sqrt{2}$ and $\Phi = -\pi/4$.

Real and imaginary components can be clearly identified in Figure 4.36. Observe that just as $\sin(t)$ is related to $\cos(t)$ by a phase shift of $\pi/2$ rad, then graphically the $\sin(t)$ and $\cos(t)$ components differ by a rotation about the frequency axis by $\pi/2$ rad. Similarly, the linear combination of the orthogonal components of $s(t) = \sin(t) + \cos(t)$ is equivalent to a cosine shifted by $\pi/4$ rad. Moreover, the effective angular phase rotation about the frequency axis describes a cosine shifted by $\pi/4$ rad, as shown in Figure 4.37. The phase variation with respect to a normal cosine may also be seen in this sketch.

The (vector) addition of orthogonal $\sin(t)$ and $\cos(t)$ components exactly illustrates the mathematical relationship: $s(t) = \sin(t) + \cos(t) = \sqrt{2}\cos(t - \pi/4)$. This frequency graph should be compared to the time-domain sketch of Figure 4.30.

Note that the positive and negative components rotate in opposite directions with respect to each other for a given phase shift. Furthermore, observe on the graph that the result found above might also be obtained by starting with the (imaginary) sine components, and then rotating both $\pi/4$ rad in the opposite direction to that shown on the figure; this phase shift operation corresponds to the equivalent expression $s(t) = \sqrt{2}\sin(t + \pi/4)$.

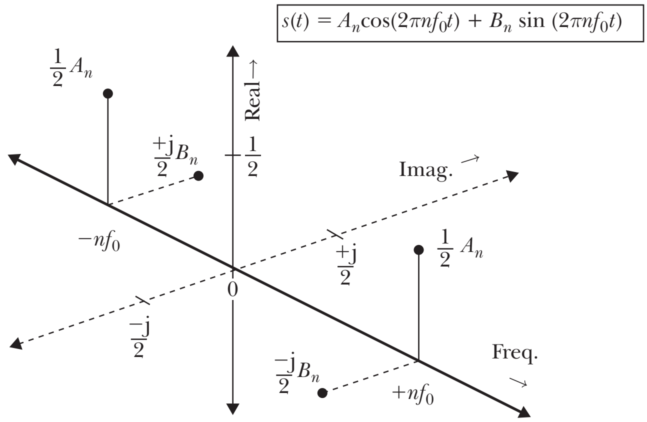



**FIGURE 4.38** A single complex Fourier series component, described as $A_n\cos(2\pi n f_0 t) + B_n\sin(2\pi n f_0 t)$ for the fundamental frequency $f_0$ Hz.

## 4.14.2 Comparing the Real and Complex Fourier Series

The relationship between the real $\{A_n, B_n\}$ Fourier series and the complex $\{C_n\}$ Fourier series can be readily established. Consider a single frequency term at frequency $nf_0$ in the Fourier series expression, as shown in Figure 4.38;

$$s(t) = A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t)$$

By inspection of this graph, and the above equation for $s(t)$ in which $A_n$ and $B_n$ are the amplitudes of the sinusoidal components in the Fourier series, the complex components $C_N$ are:

$$\text{For } n > 0, \quad C_n = A_n - jB_n$$

and

$$\text{For } n < 0, \quad C_n = A_n + jB_n$$

Similarly, by inspection, the squared magnitude of each of the complex components $C_n$ is related to $A_n$ and $B_n$ by:

$$C_n^2 = A_n^2 + B_n^2 \tag{4.33}$$

Finally, the phase $\Phi$ of the shifted sinusoidal component $s(t) = \cos(2\pi n f_0 t + \Phi)$ corresponding to $C_n$ can be found directly on the graph and in terms of $A_n$ and $B_n$ as:

$$\Phi = \tan^{-1}\frac{B_n}{A_n} \tag{4.34}$$

## 4.14.3 Magnitude and Phase

A useful alternative to sketching either the real and imaginary graph or the three-dimensional graph is to sketch separate graphs of Magnitude vs. Frequency and Phase vs. Frequency. It has been shown that the magnitude of a complex component reflects the overall amplitude of the corresponding unshifted sinusoidal

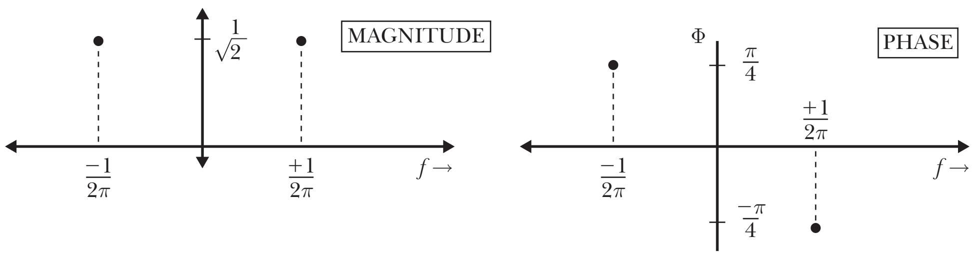



**FIGURE 4.39** The signal $s(t) = \sin(t) + \cos(t) = \sqrt{2}\cos(t - \pi/4)$, as Magnitude vs. Frequency and as Phase vs. Frequency. A phase change (time shift) of a signal has no effect on its magnitude.

component. The magnitude graph is very useful, for example, in computing the power in a signal component, since the power in a sinusoid is independent of any phase shift that might be present. In Figure 4.39, the signal $s(t) = \sin(t) + \cos(t) = \sqrt{2}\cos(t - \pi/4)$ is shown as Magnitude vs. Frequency and Phase vs. Frequency.

Note that the overall magnitude shown graphically can be interpreted as a cosine with no apparent phase shift. The magnitude (always real and even) is independent of any phase shift or time shift in a signal. Because the magnitude is even, there is really no need to see both positive and negative sides of the frequency axis, just as a spectrum analyzer displays the magnitude of an input signal showing only the positive frequency axis.

## 4.15 Complex Fourier Series Components

The complex Fourier series is presented as a complete set of orthogonal periodic signals:

$$s(t) \approx \sum_{n=-\infty}^{\infty} C_n e^{+j2\pi n f_0 t} \tag{4.35}$$

Components $C_n$ of a periodic signal $s(t)$ with period $T = 1/f_0$ can be found by application of the signal inner product integral. Recall that the inner product definition for complex signals requires that the complex conjugate of the second term be used in both the numerator and denominator.

$$C_n = \frac{\displaystyle\int_{-T/2}^{+T/2} s(t) e^{-j2\pi n f_0 t}\, dt}{\displaystyle\int_{-T/2}^{+T/2} e^{+j2\pi n f_0 t} e^{-j2\pi n f_0 t}\, dt} \tag{4.36}$$

and the denominator equals $T$, leaving:

$$C_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(t) e^{-j2\pi n f_0 t}\, dt \tag{4.37}$$

The complex Fourier series is a linear combination of orthogonal complex exponential terms based on complex components of sines and cosines at integer multiples, or harmonics, of a fundamental frequency $f_0$. In addition, a zero-frequency (or DC) component found at $C_0$ acts to "raise or lower" the waveform with respect to the horizontal axis. Significantly, the presence of $C_0$ does not alter the shape or overall appearance of a waveform.

Although integration of an exponential is generally straightforward, there may be times when it will be easier to break the previous complex exponential into sine and cosine terms before integration, leaving:

$$C_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(t)[\cos(2\pi nf_0 t) - j\sin(2\pi nf_0 t)]\,dt \tag{4.38}$$

In particular, if the signal $s(t)$ is known to be odd or even, then one of the above integrals will be zero. This approach may simplify the necessary computation. Given the above equation, it can be seen that even or odd signals $s(t)$ will have only real or imaginary components, respectively, in the complex Fourier series. This observation will now be explored.

### 4.15.1 Real Signals and the Complex Fourier Series

Consider a real-valued periodic signal that is to be approximated by the complex Fourier series. Any real and odd signal components will be expressed as a linear combination of orthogonal sines, each corresponding to an odd pair of imaginary complex exponential components. Likewise, any real and even signal components will be expressed as even pairs of real complex exponential components. These relationships are summarized in Table 4.3. For completeness, the same relationships are shown for imaginary signals.

It follows that the *magnitude* of each component in a real signal $s(t)$, whether real or imaginary or odd or even, will be real and even.

As a general rule, the complex Fourier series of any real signal $s(t)$ has real part even and imaginary part odd. The Fourier series is easily sketched given the sine and cosine components, as shown in Figure 4.40.

### 4.15.2 Stretching and Squeezing: Time vs. Frequency

The complex Fourier series representation of a signal $s(t)$ with period $T = 1/f_0$ is distinguished by components $C_n$ corresponding to integer multiples of $f_0$. If the period $T$ of $s(t)$ is increased, the spacing between the lines $n/f_0$ will decrease, and vice versa. The values $C_n$ corresponding to each frequency $nf_0$ will not change. In other words, the only difference between square waves with different periods is the

**TABLE 4.3**

**Time-Domain and Frequency-Domain Properties**

| Time Domain | Complex Fourier Series |
|---|---|
| Real & even | Real & even |
| Real & odd | Imaginary & odd |
| Real & neither | Real part even & imaginary part odd |
| Imaginary & even | Imaginary & even |
| Imaginary & odd | Real & odd |
| Imaginary & neither | Real part odd & imaginary part even |

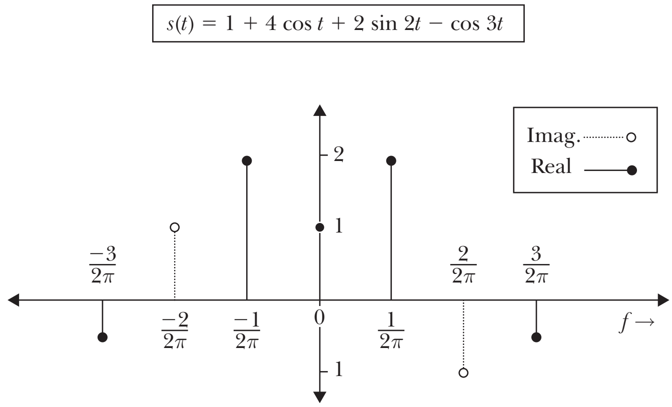



**FIGURE 4.40** The complex Fourier series of any real signal $s(t)$ has real part even and imaginary part odd.

spacing of components in the frequency domain (always at integer multiples of the fundamental frequency). Substituting $s(kt)$ into the defining equation for the complex Fourier series yields a direct result:

$$C_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(kt)e^{-j2\pi n f_0 t}\,dt \tag{4.39}$$

Let $x = kt$, then $t = x/k$ and $dt = 1/kdx$. The new period will be $X = kT$, or $T = X/k$ and the fundamental frequency is $1/X = 1/kT$ Hz.

$$= \frac{k}{X}\int_{\frac{-X/k}{2}}^{\frac{+X/k}{2}} s(x)e^{-j2\pi(1/k)nf_0 x}\frac{1}{k}\,dx \tag{4.40}$$

$$= \frac{1}{X}\int_{\frac{-X/k}{2}}^{\frac{+X/k}{2}} s(x)e^{-j2\pi(1/k)nf_0 x}\,dx \tag{4.41}$$

This is the same expression for $C_n$ with a new period $X = kT$. The values of $C_n$ for a given frequency $nf_0$ do not change; however, the spacing between components in the frequency domain varies inversely with the period of the signal $s(kt)$. In effect, *stretching* a signal in one domain implies *squeezing* in the other. This is the effect described by the tape recording example described early in Chapter 1. When a tape is played slowly, time stretches out, but the frequency content is lowered, and vice versa. This property will be formalized in the next chapter, when the Fourier series is generalized to nonperiodic signals.

## 4.15.3 Shift in Time

Shifting a waveform in time does not change its appearance. The shape and amplitude remain constant, while the contribution of each component $C_n$ varies only in phase.

Let $\{C_n\}$ be the complex Fourier series components of $s(t)$, and let $\{S_n\}$ be the complex Fourier series components of $s(t - t_0)$. To find each $S_n$, substitute $s(t - t_0)$ into the defining equation for the complex Fourier series:

$$S_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(t - t_0)e^{-j2\pi n f_0 t}\,dt$$

Let $x = t - t_0$, then $t = x + t_0$ and $dx = dt$ and the period does not change.

$$S_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(x)\, e^{-j2\pi n f_0[x+t_0]}\, dx$$

$$S_n = e^{-j2\pi n f_0 t_0}\,\frac{1}{T}\int_{-T/2}^{+T/2} s(x)\, e^{-j2\pi n f_0 x}\, dx$$

therefore,

$$S_n = e^{-j2\pi f_0 n t_0}\, C_n$$

Where the same components $C_n$ each undergo a phase change as a function of $nf_0 t_0$ (the time shift $t_0$ and the frequency $nf_0$ of the component $C_n$). Moreover, the magnitude of each component $|S_n| = |C_n|$ is unaffected by a phase shift.

**THEOREM 4.1** *(Time Shift)*

*If a periodic signal* $\mathrm{s(t)}$ *with period* $\mathrm{T = 1/f_0}$ *is expressed as complex Fourier series components* $\{\mathrm{C_n}\}$*, then the time shifted signal* $\mathrm{s(t - t_0)}$*, has complex Fourier series terms:*

$$\{e^{-j2\pi f_0 n t_0}\, C_n\}$$

Check: If the time shift is exactly one period, then $s(t - t_0) = s(t)$, and there will be no shift. Let $t_0 = 1/f_0$, giving:

$$\{e^{-j2\pi n}\, C_n\} = \{C_n\}$$

and there is no change since

$$e^{-j2\pi n} = 1 \text{ for all integer } n.$$

## 4.15.4 Change in Amplitude

Amplifying (attenuating) a signal by a factor $k$ simply multiplies the frequency components $C_n$ by the same constant $k$.

## 4.15.5 Power in Periodic Signals

Recall that the power in a periodic signal with period $T$ is given by:

$$\text{Power} = \frac{1}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} |s(t)|^2\, dt$$

The power in any sinusoidal signal $a(t) = A\sin(2\pi f_0 t)$ is given directly by $A^2/2$. For the periodic signal $c(t) = Ce^{-j2\pi f_0 t}$, the power is given directly by $|C|^2$, since the magnitude of the complex exponential equals 1:

$$\text{Power} = |C|^2 = \frac{1}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} C^2 e^{+j2\pi f_0 t} e^{-j2\pi f_0 t}\, dt$$

Through application of the Fourier series, any periodic signal can be represented as a linear combination of orthogonal sine and cosine terms. It will be shown that the total power in $s(t)$ can be found by adding the power from each of the orthogonal component terms individually. This should not be immediately obvious because, given either of these linear combinations of components, computing $|s(t)|^2$ implies that the entire summation must be squared, resulting in a large number of squared terms and cross terms. To suggest that the same result is obtained by adding the computed power in individual terms (square) or by squaring the entire signal seems to contradict the arithemetic rule $[x^2 + y^2] \neq [x + y]^2$. A sample computation will resolve this dilemma.

**Find the Total Power in $s(t) = A\cos(t) + B\sin(t)$** The signal $s(t)$ is the linear combination of two orthogonal Fourier series components.

The total power in $s(t)$ is given by:

$$\text{Power} = \frac{1}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} [A\cos(t) + B\sin(t)]^2\, dt$$

Expanding the square of the sum reveals three terms:

$$= \frac{1}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} [A^2\cos^2(t) + 2AB\sin(t)\cos(t) + B^2\sin^2(t)]\, dt$$

$$= \frac{1}{T}\left[ A^2\int_{-\frac{T}{2}}^{+\frac{T}{2}} \cos^2(t)\ dt \ + \ 2AB\int_{-\frac{T}{2}}^{+\frac{T}{2}} \sin(t)\ \cos(t)\ dt \ + \ B^2\int_{-\frac{T}{2}}^{+\frac{T}{2}} \sin^2(t)\ dt \right]$$

Now, because $\cos(t)$ and $\sin(t)$ are orthogonal over this interval, the second integral containing these cross terms equals zero; it is the inner product between the two signals. Consequently, the overall power is simply the sum of the power found in each sinusoidal component.

This simplification is possible only because the component terms are orthogonal. This result is significant as it illustrates that, for signals expressed as orthogonal components, the power in individual components can be computed separately, because the cross-product terms will all give zero. The above observation, when generalized to a complete orthogonal set of signals, is known as *Parseval's theorem.* Its application implies that the Fourier series coefficients alone, or a frequency-domain graph, can be used directly to determine the relative power in each frequency component of a signal. In fact, the squared Magnitude vs. Frequency graph is often referred to as the *power spectrum* of a signal, since it illustrates the relative power-per-frequency in each component.

## 4.15.6 Parseval's Theorem for Periodic Signals

Because its Fourier series components are mutually orthogonal, the power in a periodic $s(t)$ can be computed in time domain or the frequency domain.

**THEOREM 4.2** *(Parseval's Theorem)*

*If a periodic signal* s(t) *is expressed as complex Fourier series components* $\{C_n\}$, *then the power in* s(t) *is given by:*

$$\text{power} = \int_{-\infty}^{\infty} |s(t)|^2 \, dt = \sum_{n=-\infty}^{\infty} |C_n|^2$$

**EXAMPLE 4.6 (Power in a Cosine)**

*Find the power in* $s(t) = A\cos(2\pi f_0 t)$ *in both time and frequency domains.*

**Answer:**

1. In the time domain, $V_{rms} = A/\sqrt{2}$ for a sinusoidal signal. The power is $A^2/2$.
2. In the frequency domain, the power in $s(t)$ can be computed as the sum of the power found in each of two complex components $C_n$ corresponding to the cosine, as shown in Figure 4.41. The components $C_n$ and $C_{-n}$ each equal $A/2$, giving the same total power, as shown below.

$$Power = C_{-1}^2 + C_{+1}^2 = \left[\frac{A}{2}\right]^2 + \left[\frac{A}{2}\right]^2 = \frac{A^2}{2}$$

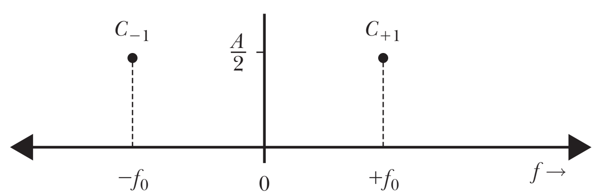



**FIGURE 4.41** The power in this time-domain cosine signal $s(t) = A\cos(2\pi f_0 t)$ can be computed in the frequency domain by adding the power in each orthogonal component $C_n$.

**EXAMPLE 4.7 (Power in Orthogonal Cosines)**

*Find the total power and* $V_{rms}$ *voltage of* $s(t) = \cos(t) - 1/3\cos(3t)$ V, *as seen in Figure* 4.42.

**Answer:**

This signal is actually the first two terms in the Fourier series expansion of an even square wave with period $T = 2\pi$ seconds, and amplitude $A = \pi/4$ Volts. Calculation of the total power in $s(t)$ proves to be much easier in the frequency domain. By inspection, the linear combination of two cosines $s(t) = \cos(t) - 1/3\cos(3t)$ leads to the complex Fourier series components in Figure 4.43.

The total power in $s(t)$ may be computed by adding the squared magnitude of each frequency-domain component $C_n$ Figure 4.43:

$$\text{Power} = \left[\frac{-1}{6}\right]^2 + \left[\frac{+1}{2}\right]^2 + \left[\frac{+1}{2}\right]^2 + \left[\frac{-1}{6}\right]^2 = \frac{5}{9}$$

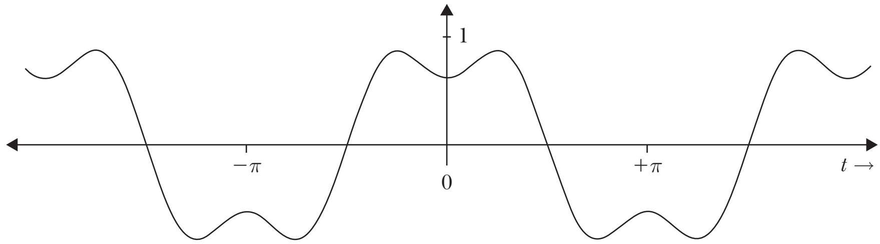



**FIGURE 4.42** $s(t) = \cos(t) - 1/3\cos(3t)$ Volts, as Amplitude vs. Time: The power in this signal can be computed in either the time or frequency domain. See Figure 4.43.

The rms voltage is then given by the square root of the power:

$$V_{rms} = \sqrt{\text{Power}} = \sqrt{\frac{5}{9}} = 0.745\ \text{V}$$

This root mean squared voltage is slightly greater than the $1/\sqrt{2} = 0.707$, which is contributed from the unit amplitude cosine term considered alone. The addition of an orthogonal component has increased the overall signal power, and the $V_{rms}$ voltage. The resulting $V_{rms}$ is in Figure 4.44.

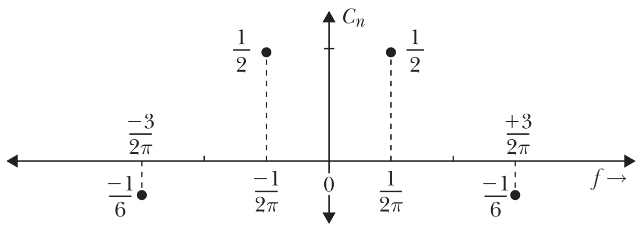



**FIGURE 4.43** $s(t) = \cos(t) - 1/3\cos(3t)$ Volts, as Amplitude vs. Frequency. The power in this time-domain cosine signal can be computed in the frequency domain by adding the power in each orthogonal component $C_n$. See Figure 4.42.

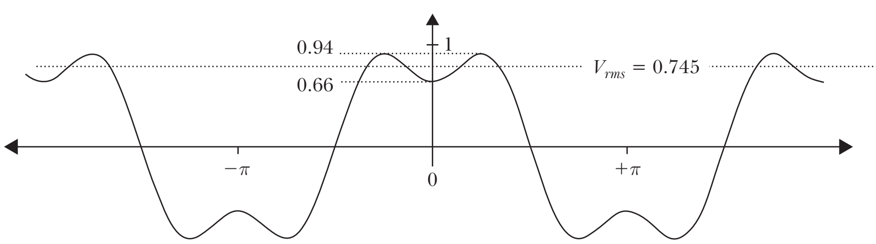



**FIGURE 4.44** Voltage of $s(t) = \cos(t) - 1/3\cos(3t)$ Volts, as Amplitude vs. Time, showing computed $V_{rms}$ = 0.745 V.

The time-domain and frequency-domain graphs are different ways of representing the same signal. As in this example, it is often possible to perform computations in either domain, choosing one or the other as appropriate, to facilitate computations and to better understand the behavior and properties of a signal. Properties of a signal in either domain can often be ascertained in the other, without direct observation. As another example of such an analytical approach, consider the definition of the complex Fourier series, and let $t = 0$:

$$s(t) = \sum_{n=-\infty}^{\infty} C_n e^{j2\pi n f_0 t}$$

$$s(0) = \sum_{n=-\infty}^{\infty} C_n$$

In other words, the value of $s(0)$ in the time domain equals the arithmetic sum of all the components $C_n$ in the frequency domain. Such properties can be extremely valuable. This assertion may be confirmed by inspection of the example $s(t)$ above.

$$\text{Let } RHS = s(0) = \cos(0) - \frac{1}{3}\cos(0) = \frac{2}{3}$$

From Figure 4.43:

$$\text{Let } LHS = \sum_{n=-\infty}^{\infty} C_n = \frac{-1}{6} + \frac{1}{2} + \frac{1}{2} + \frac{-1}{6} = \frac{2}{3}$$

Since $RHS = LHS$, the property is confirmed.

**Example:** Consider the unit-amplitude square wave Fourier series coefficients $B_n$ as computed in Section 4.12. The total power in this odd square wave equals 1.000. The power in each of the sinusoidal components $B_n$ is given by $(B_n)^2/2$. The cumulative sum of the power in each component should tend to 1.000. Compute the power in each component and the total power up to $N = 7$.

The total power found in components up to $n = 7$ is:

$$\frac{1}{2}\frac{16}{\pi^2}\left[1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49}\right] = 0.9496$$

or nearly 95 percent of the total square wave power.

Because the components are mutually orthogonal, each contributes in a unique and separate way to the overall signal. The signal is fully approximated only when all the non-zero components have been added. The power added by each component contributes to the overall power in the signal, and the relative power in each component is a measure of its overall contribution to the signal. Analyses such as the above permit an accurate determination of how many components are necessary to produce an adequate approximation to a signal $s(t)$. The same results can be used

**TABLE 4.4**

**Fourier Series Components of an Odd Square Wave**

| $B_n$ | $B_1$ | $B_2$ | $B_3$ | $B_4$ | $B_5$ | $B_6$ | $B_7$ |
|---|---|---|---|---|---|---|---|
| $\frac{4}{n\pi}[1 - \cos(n\pi)]$ | $\frac{4}{\pi}$ | 0 | $\frac{+1}{3}\cdot\frac{4}{\pi}$ | 0 | $\frac{+1}{5}\cdot\frac{4}{\pi}$ | 0 | $\frac{+1}{7}\cdot\frac{4}{\pi}$ |
| Power | $\frac{1}{2}\frac{16}{\pi^2}$ | 0 | $\frac{1}{2}\frac{1}{9}\frac{16}{\pi^2}$ | 0 | $\frac{1}{2}\frac{1}{25}\frac{16}{\pi^2}$ | 0 | $\frac{1}{2}\frac{1}{49}\frac{16}{\pi^2}$ |

to assess what range of frequency components must be handled by a communications channel before accurate waveforms can be transmitted.

In this limit, the sum of the power in all the non-zero components in the above signal forms an infinite series that should converge to the total power = 1.000. In other words:

$$1 = \frac{1}{2}\frac{16}{\pi^2}\sum_{n=0}^{\infty}\frac{+1}{[2n+1]^2}$$

This is true if the summation alone converges to $\pi^2/8$, or :

$$\frac{\pi^2}{8} = \sum_{n=0}^{\infty}\frac{1}{[2n+1]^2}$$

This result can be confirmed in MATLAB. Note that this infinite series does not converge very rapidly.

**EXAMPLE 4.8 (Infinite Series in MATLAB)**

*Show that:*

$$\frac{\pi^2}{8} = \sum_{n=0}^{\infty}\frac{1}{[2n+1]^2}$$

**Solution:**

```
n = 0:10000;   % sum the first 10000 terms
RHS = sum(1./([2*n+1].^2))

RHS =
    1.2337

LHS = pi^2 / 8

LHS =
    1.2337
```

Since *RHS* = *LHS*, the series appears to converge to the expected value.

# 4.16 Properties of the Complex Fourier Series

$$s(t) \approx \sum_{n=-\infty}^{\infty} C_n e^{j2\pi n f_0 t}$$

The table on page 152 illustrates some important relationships between the time domain and the frequency domain for various values and manipulations.

| **Periodic Signal *s(t)*** | $\Leftrightarrow$ | **Complex Fourier Series** |
|---|---|---|
| $s(t)$ | $\Leftrightarrow$ | $\{C_n\}$ |
| Period = $T$ seconds | $\Leftrightarrow$ | Discrete values at $nf_0$ Hz |
| Real | $\Leftrightarrow$ | Real & even, imaginary & odd |
| DC Offset $= \dfrac{1}{T}\displaystyle\int_{-T/2}^{+T/2} s(t)\,dt$ | $\Leftrightarrow$ | $C_0$ |
| $s(0)$ | $\Leftrightarrow$ | Total Area $= \displaystyle\sum_{n=-\infty}^{+\infty} C_n$ |
| Power $= \dfrac{1}{T}\displaystyle\int_{-T/2}^{+T/2} \lvert s(t)\rvert^2\,dt$ | $\Leftrightarrow$ | Power $= \displaystyle\sum_{n=-\infty}^{+\infty} \lvert C_n\rvert^2$ |
| $ks(t)$ | $\Leftrightarrow$ | $k\ C_n$ |
| $s(kt)$ | $\Leftrightarrow$ | Discrete values at $n/kf_0$ Hz |
| $s(t - t_0)$ | $\Leftrightarrow$ | $\{e^{-j2\pi nf_0 t_0}\, C_n\}$ |

## 4.17 Analysis of a DC Power Supply

In this section, the Fourier series is used to analyze a simple DC power supply. Although other techniques could serve in this analysis, the goal here is to illustrate a practical application of the Fourier series while fully exploring the characteristics and behavior of the power supply.

Electric utilities deliver electricity to homes in the form of sinusoidally varying voltage waveforms. In North America, a 60 Hz, 120 V (rms) waveform is generally delivered to homes, while a 50 Hz, 220 V (rms) waveform is employed in much of the rest of the world. Although the average voltage of a cosine waveform is zero, its overall power is finite and non-zero, and sinusoids are classified as being power signals. Because the standard signal level is quoted as an RMS voltage, the cosine entering North American homes and businesses actually has an amplitude of $\sqrt{2} \times 120 = 170\,\mathrm{V}$, or a peak-to-peak amplitude of almost 340 V.

### 4.17.1 The DC Component

The $V_{rms}$ voltage of a waveform is not the same as the DC component. In particular, for a sinusoid $V_{rms} = A/\sqrt{2}\,\mathrm{V}$ where $A$ is the amplitude, while the DC component is defined as the average voltage level of the waveform, and here $\mathrm{V_{DC}} = 0\,\mathrm{V}$.

If there is no DC component in a cosine signal, where does the DC come from when needed for supplying battery-powered home appliances? A simple transformer can bring the voltage down to useful levels, but what is inside the familiar plug-in *DC power supply* or *AC-DC adapter* that acts to *convert* AC to DC? Looking at this problem from the perspective of signals makes these distinctions clear.

As there is no DC component in a pure cosine waveform, an AC-DC adapter must make one. To this end, such devices generally make use of nonlinear circuit elements (diodes or rectifiers) to effectively introduce a DC component. At the same

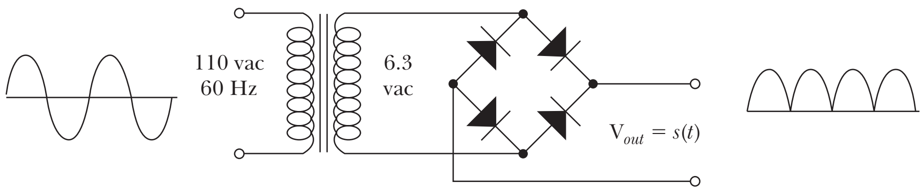



**FIGURE 4.45** **A DC Power Supply** The sinusoidal input signal has zero DC component; however, the rectified output signal has a significant DC component, which, when isolated, provides the DC output.

time, as may be expected of nonlinear circuit elements, many other harmonic components are generated. To be useful, a significant fraction of the power in the AC input signal should emerge in the DC component produced by the power supply. Ideally, 100 percent efficiency might be anticipated, and different circuit designs achieve different efficiencies (not surprisingly, always less than 100 percent). It would be helpful if the efficiency of a given design approach could be evaluated.

## 4.17.2 An AC-DC Converter

The circuit diagram of a simple DC power supply is shown in Figure 4.45. The first step in this conventional power supply design is a *step down transformer* that brings the amplitude of the incoming 60 Hz sinusoid down to some reasonable level, close to the desired DC voltage. Next, a *full-wave rectifier* is used in which the cosine signal essentially passes through an ideal absolute value circuit (as $g(t) = |t|$, assuming no voltage drop through the diodes). This ideal output signal will be studied. Inspection of the output waveform shows that the goal of producing some DC component has been achieved. On the other hand, many non-DC components that were not previously present will also be found. In particular, the full-wave rectified periodic signal has a fundamental frequency of 120 Hz that is double that of the original 60 Hz cosine.[3] It is useful to examine the relationships between the DC component and the overall signal (consisting of all frequency components, including DC).

The following questions may be asked of the waveform output from Figure 4.45:

1. What is the Fourier series expansion of this waveform?
2. What is the DC component of this signal?
3. What is the $V_{rms}$ value of this signal?
4. How is the $V_{rms}$ value related to the DC component?
5. What fraction of the total power is contained in the DC component?
6. How can the DC component be usefully extracted from this signal?
7. What happens if the DC component cannot be extracted cleanly?

## 4.17.3 $V_{rms}$ Is Always Greater Than or Equal to $V_{dc}$

It should now be clear what the AC-DC adapter does to "convert AC to DC." The original cosine input signal has no DC component; however, the full-wave rectified output waveform is a periodic signal with a DC component, plus harmonic

[3] This is not to say that this circuit has doubled the frequency of the cosine, but has led to a signal with double its fundamental frequency. Nonetheless, such a circuit configuration can also be applied usefully as a frequency doubler in various applications.

components at integer multiples of its fundamental frequency. The DC component is therefore just one of many orthogonal components making up the rectified waveform. By comparison, the rms voltage (non-zero even for the original cosine) is the square root of the total power in all the components, including DC. By inspection, therefore, $V_{rms} \geqslant V_{dc}$. (The equality is true only for a pure DC signal.) Computation of the power in the individual Fourier series components can be used to determine what fraction of the overall waveform power is actually contained in the DC component.

### 4.17.4 Fourier Series: The Full-Wave Rectifier

The first component to be computed in the Fourier series is the zero-frequency or DC component, given by:

$$A_0 = \frac{1}{T}\int_{-T/2}^{+T/2} s(t)\ dt$$

Before continuing, it is necessary to find a suitable expression for $s(t)$, the full-wave rectified sinusoid. By inspection, the full-wave rectified signal consists of positive half periods of a 60 Hz cosine. These half cycles define the overall period of the output signal as $T = 1/120$ s, $f_0 = 1/\mathrm{T} = 120$ Hz, even though $s(t)$ is a 60 Hz cosine with frequency 1/2T Hz or $s(t) = \cos(2\pi(1/2T)t)$.

$$A_0 = \frac{1}{T}\int_{-T/2}^{+T/2} A\cos(2\pi(1/2T)t)\ dt$$

The DC component may now be computed. Recognizing that $s(t)$ is an even function, this integral immediately reduces to:[4]

$$\begin{aligned}
A_0 &= \frac{2}{T}\int_{0}^{+T/2} A\cos(2\pi(1/2T)t)\ dt \\
&= \frac{2A}{T}\int_{0}^{+T/2} \cos(\pi(1/T)t)\ dt \\
&= \frac{2A}{\pi}\ \sin(\pi(1/T)t)\Big|_{t=0}^{T/2} \\
&= \frac{2A}{\pi}\ \sin\left(\frac{\pi}{2}\right) \\
&= \frac{2A}{\pi}
\end{aligned}$$

As seen in Figure 4.46, DC component in the full-wave rectified cosine equals $2A/\pi$. This is slightly less than the $V_{rms}$ $(A/\sqrt{2})$, as expected.

[4] In general, it is faster and more instructive to approach problems such as this by substituting numerical values such as $2\pi 60$ and $1/120$ only after completing the integration, and only if necessary. Always complete a thorough mathematical analysis before substituting numerical values. Not only will constant terms frequently cancel out, but the mathematics will be much clearer, and errors will be easier to spot. A calculator should normally be employed only as a last step, if numerical results are required. In this case, the result is independent of period, so no numerical substitution is involved at all.

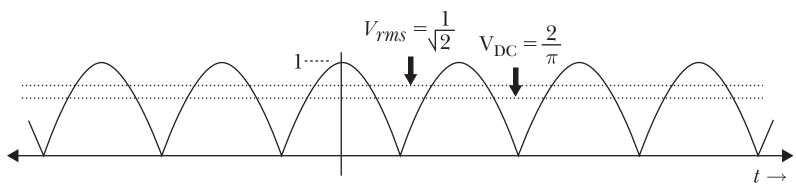



**FIGURE 4.46** $V_{rms}$ compared to $V_{dc}$ for the full-wave rectified cosine $s(t) = |A\cos(2\pi f_0 t)|$, for A = 1V. In general, $V_{rms}$ is always greater than or equal to $V_{dc}$.

The total power in this waveform is given by $(V_{rms})^2 = 1/2$. The power in the DC component alone is given by $(2/\pi)^2 = 4/\pi^2$. The fraction of total signal power found in the DC component is therefore:

$$\frac{\text{Power in DC}}{\text{Total Power}} = \frac{4/\pi^2}{1/2} = 0.811$$

In other words, about 81 percent of the total power is in the DC component. This is quite good, considering that the goal of this circuit was to convert a sinusoidal voltage to a DC level. The remaining 19 percent is due to non-zero frequency components in the waveform. On the other hand, if the desired *pure DC* output signal is to be obtained, those other components must be suppressed. The practical feasibility of eliminating those components depends on their frequency (120 Hz, 240 Hz, 360 Hz, 480 Hz $\cdots$) and on their relative (power) contributions to the waveform. This requires computation of the Fourier series for the non-zero frequency components.

## 4.17.5 Complex Fourier Series Components $C_n$

Since the full-wave rectified signal is even, the simple Fourier series components $\{A_n, B_n\}$ will be found.

**Odd Fourier series components $B_n$**—Since the full-wave rectified cosine waveform is an even function, there will be no odd components in the series. By inspection, $B_n = 0$, for all n.

**Even Fourier series components $A_n$**—The fundamental frequency $f_0 = 120\text{ Hz} = 1/T$, and $s(t) = A\cos(\pi t/T)$ as when computing the DC component. This leaves:

$$A_n = \frac{2}{T}\int_{-T/2}^{+T/2} s(t)\cos(2\pi n(1/T)t)\,dt$$

$$A_n = \frac{2}{T}\int_{-T/2}^{+T/2} A\cos(2\pi(1/2T)t)\cos(2\pi n(1/T)t)\,dt$$

Recognizing that the integral is an even function, this simplifies to:

$$A_n = \frac{4}{T}\int_{0}^{+T/2} A\cos(2\pi(1/2T)t)\cos(2\pi n(1/T)t)\,dt$$

Next, recall the identity 2 cosA cosB = $\cos(A - B) + \cos(A + B)$, leaving:

$$A_n = \frac{2A}{T}\int_{0}^{+T/2} \cos\left(2\pi\frac{1-2n}{2T}t\right) + \cos\left(2\pi\frac{1+2n}{2T}t\right)\,dt$$

$$A_n = \frac{2A}{T}\left[\int_0^{+T/2} \cos\left(\pi \frac{1-2n}{T} t\right) dt + \int_0^{+T/2} \cos\left(\pi \frac{1+2n}{T} t\right) dt\right]$$

$$A_n = \frac{2A}{T}\left[\frac{T}{\pi(1-2n)} \sin\left(\pi \frac{1-2n}{T} t\right)\Big|_{t=0}^{T/2} + \frac{T}{\pi(1+2n)} \sin\left(\pi \frac{1+2n}{T} t\right)\Big|_{t=0}^{T/2}\right]$$

$$A_n = \frac{2A}{\pi}\left[\frac{1}{1-2n} \sin\left(\pi \frac{1-2n}{T} t\right)\Big|_{t=0}^{T/2} + \frac{1}{1+2n} \sin\left(\pi \frac{1+2n}{T} t\right)\Big|_{t=0}^{T/2}\right]$$

$$A_n = \frac{2A}{\pi}\left[\frac{1}{1-2n} \sin\left(\pi \frac{1-2n}{2}\right) + \frac{1}{1+2n} \sin\left(\pi \frac{1+2n}{2}\right)\right]$$

$$A_n = \frac{2A}{\pi}\left[\frac{1}{1-2n} \sin\left(\frac{\pi}{2}(1-2n)\right) + \frac{1}{1+2n} \sin\left(\frac{\pi}{2}(1+2n)\right)\right]$$

Both the sine expressions above simply alternate sign with successive values of $n$; both can be replaced by $(-1)^n$ found by substitution in this equation. Let the amplitude $A = 1$, then:

$$A_1 = \frac{2A}{\pi}\left[\frac{1}{1-2} \sin\left(\frac{\pi}{2}(1-2)\right) + \frac{1}{1+2} \sin\left(\frac{\pi}{2}(1+2)\right)\right]$$
$$= \frac{2A}{\pi}\left[\frac{1}{-1} \sin\left(-1\frac{\pi}{2}\right) + \frac{1}{3} \sin\left(3\frac{\pi}{2}\right)\right] = \frac{4}{\pi} \cdot \frac{1}{3}$$

Other terms follow as:

$$A_1 = +\frac{4}{\pi} \cdot \frac{1}{1 \cdot 3}$$

$$A_2 = -\frac{4}{\pi} \cdot \frac{1}{3 \cdot 5}$$

$$A_3 = +\frac{4}{\pi} \cdot \frac{1}{5 \cdot 7}$$

$$A_4 = -\frac{4}{\pi} \cdot \frac{1}{7 \cdot 9}$$

From these coefficients, the frequency-domain representation of this signal includes components of cosine spaced at 120 Hz intervals, alternating sign, with a DC component of $2/\pi$ as shown in Figure 4.47.

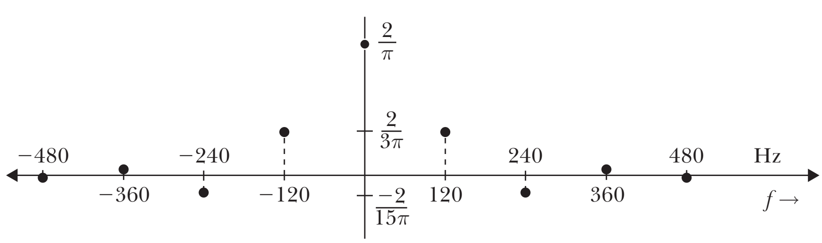



**FIGURE 4.47** Fourier Series Expansion of the full-wave rectified cosine $s(t) = |\cos(2\pi f_0 t)|$.

**MATLAB Exercise 2: Plotting Fourier Series Components** MATLAB may be used to plot the first two terms (DC + fundamental frequency) of the previous approximation.

```
t = -5 : 0.01 : 5;
a = abs(cos(t));                % full-wave rectified cosine
f0= 2/pi;                       % define DC component
f1= (1/3)*(4/pi)*cos(2*t);      % define fundamental component
plot(t,a,t,f0+f1);              % plot DC and the 1st term
grid on;
```

**Power in the Fundamental Frequency Component at 120 Hz** Recall that the power in the half-wave rectified signal with amplitude $A$ is:

$$\text{Power} = A^2/2 = 0.5A^2$$

It has been shown that the power in the DC component ($A_0 = 2A/\pi$) is

$$A_0^2 = 4A^2/\pi^2 = 0.405A^2$$

By inspection, the power in the non-zero frequency components will fall off rapidly with frequency. Each cosine component $A_n$ has power equal to $A_n^2/2$. Now, from the previous calculation, $A_1 = 4A/(3\pi)$. Therefore, the power in $A_1$ is given by:

$$A_1^2/2 = 8A^2/(9\pi^2) = 0.090A^2$$

It can be seen that the power in the DC plus only the fundamental frequency equals $0.495A^2$, or over 99 percent of the total signal power. All the remaining terms carried to infinity will, combined, contribute the remaining (less than 1 percent) of the output signal power. Such a result implies that the linear combination $s(t) = A_0 + A_1\cos(2\pi f_0 t)$ must already be a very good approximation to this waveform shape. Figure 4.48 confirms this observation.

If the goal of producing a pure DC output signal is to be achieved, the importance of eliminating the component at 120 Hz is evident from Figure 4.48.

Isolating the DC component from the above signal requires a circuit capable of rejecting high frequencies and letting low frequencies pass. In practice, a large capacitor would be added across the circuit output. From the frequency-domain graph, it can be seen that the ideal filter circuit should cleanly cut off all frequency components at and above 120 Hz, leaving only the DC component. Now, if this lowpass circuit was perfect, the aim of achieving an AC-DC converter will have been accomplished. On the other hand, as no lowpass filter circuit is ideal, a fraction of

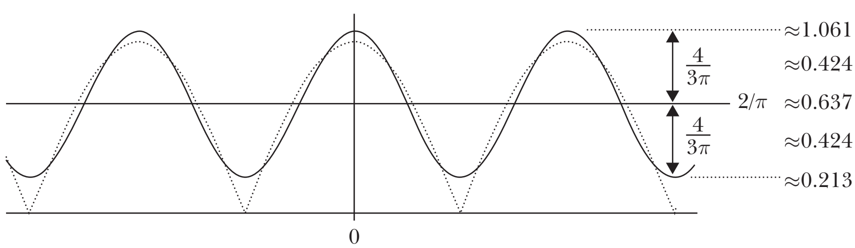



**FIGURE 4.48** The DC and fundamental frequency components contain 99 percent of the power in the full-wave cosine and consequently, nearly all its shape. The original full rectified cosine signal is superimposed for comparison.

the harmonic power will pass through, and the output of the DC power supply will be a DC component, plus some (small) components at 120 Hz. In practice, the constant voltage level exhibits some "ripple" as seen in the time domain (on an oscilloscope). Provided these components are small compared to DC, the circuit being powered by the supply will not be affected. Poor filtering is most evident in audio applications where a (120 Hz) hum may be heard in speakers or headphones powered by such a power supply. Sensitive circuits will, in any case, require some additional voltage regulation to ensure constant voltage under expected loads; this will effectively eliminate any small harmonic content in the output.

The circuit above (full-wave rectifier plus capacitor filter) is commonly found in inexpensive plug-in power supplies. A more thorough analysis would include the properties of real components and any output load. Nonetheless, this example serves to demonstrate the utility of signals analysis in understanding the behavior of electronic devices and periodic signals. In the chapters to follow, this approach will be extended to all manner of signals in an ever-widening variety of applications.

## 4.18 The Fourier Series with MATLAB

The Fourier series components of a periodic signal can be found in MATLAB using the `fft()` or *fast Fourier transform* function, so named because it implements an efficient (fast) routine to approximate a periodic signal using orthogonal complex Fourier series components. The essential features and limitations of the `fft()` function are outlined here using a simple cosine input signal, followed by some practical examples of using this versatile MATLAB function. The focus of this section is specifically in finding Fourier series components; the `fft()` function is re-examined in more detail in Chapter 10.

### 4.18.1 Essential Features of the fft() in MATLAB

**1. Periodic Signals Are Defined on a Period of $2^N$ Points** In the fast Fourier transform, periodic signals are generally defined on an interval defined by exactly $2^N$ points. One complete period of the signal to be studied will be defined for positive time ranging over 0 to $2^N - 1$ points.[5]

For example, using $N = 4$, then a cosine with a period $(T = 1/f_0)$ spanning 16 points can be defined over integer time points ranging from 0 to 15, as shown below.

```
t = 0:15; % define 16 points
cosine = cos(2*pi*t/16);    % define one period of cosine
plot(t,cosine)
grid on;
```

[5] The MATLAB `fft()` function does not demand $2^N$ points, but efficient computation of the transform depends on this constraint.

Although cosine is an even function, it has been defined here only for positive time starting at zero. The Fourier series computation can take place over any complete period; it is not necessary to be centered on the origin.

**2. The Fourier Series Is Defined on $2^{N-1} - 1$ Frequency Components** This constraint is a direct consequence of the first feature. Given a time signal defined on a period of $2^N$ points, the `fft()` function returns $2^N$ frequency components, including $2^{N-1} - 1$ orthogonal sinusoids in complex exponential form (and both positive and negative frequencies). If the interval of $2^N$ points is defined as one second, then Fourier series components from 0 through $2^{N-1} - 1$ Hz are output. For the same signal and a different period, for example, a cosine at 1 Hz, 100 Hz, or 1 MHz, the above calculation does not change; however, the components along the frequency axis are scaled (by establishing a value for $f_0$) to match the actual period.

For example, with $N = 4$, there will be $2^N = 16$ points in both the input signal and the output signal. The input signal has one period $(T = 1/f_0)$ spanning 16 points. In turn, the 16 point output signal describes paired complex frequency components $\pm nf_0$ components for $n = 1$ through 7, plus a DC component. If that single period of a cosine is established on 16 points and the cosine is said to be 8 Hz, then the 16 time points are spaced every 1/8 second, and the 16 frequency components in the output signal represent integer multiples of 8 Hz. For the cosine in this example, the Fourier series should give a single pair of non-zero components where $C_{-1} = 0.5$ and $C_{+1} = 0.5$.

The vertical axis can be scaled in MATLAB by dividing the output of `fft()` by $2^N$. The output of the `fft()` function on $2^N$ points is itself $2^N$ points, as shown below:

```
  fcos = fft(cosine) / 16; % compute 16-point FFT
  freal = real(fcos)       % view only real parts

ans =

  Columns 1 through 4
   0.0000     0.5000     0.0000     0.0000

  Columns 5 through 8
   0.0000     0.0000     0.0000     0.0000

  Columns 9 through 12
   0.0000     0.0000     0.0000     0.0000

  Columns 13 through 16
   0.0000     0.0000     0.0000     0.5000
```

There are $2^N = 16$ output values representing both positive and negative complex Fourier series components up to $2^{N-1} - 1$. The real values are shown above, where there are only two non-zero values at `freal(2)` and `freal(16)`. All imaginary components are zero (as expected for an even input signal), which may be confirmed using `imag(fs)`. However, to see the output values in a meaningful way requires some reordering.

The first 8 complex values in the `fft()` output vector (columns 1 to 8) represent positive frequency components of the complex exponential starting with the DC or $C_0$ component. The remaining 8 values (columns 9 to 16) represent the negative-frequency components. To plot the complete components as a two-sided spectrum order requires extracting the negative half of the components and arranging them

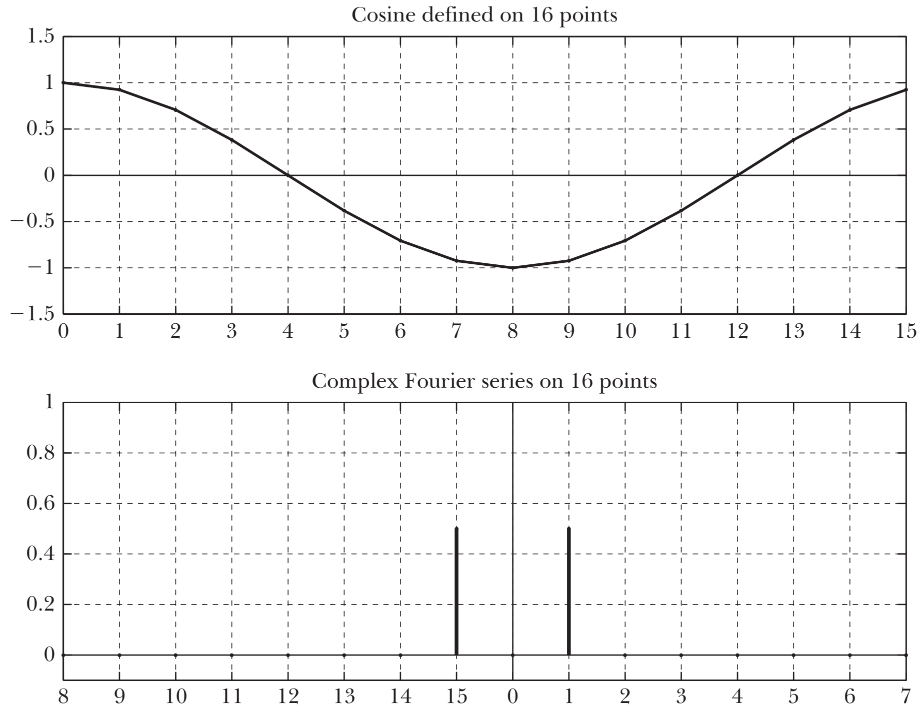



**FIGURE 4.49 MATLAB Fourier Series** A single period of cosine defined on 16 points is transformed into the frequency domain using the `fft()` function. When reordered as shown, the output matches the expected even pair of Fourier series components $C_1 = C_{-1} = 1/2$.

to the left of the positive-frequency components. The required reordering would redefine the frequency terms as `freal(9:16 1:8)`; this rearrangement task can be accomplished directly using `fftshift(freal)`. For the 16 components shown above:

```
f = 0:15;                          % define 16 frequency components
reorder = fftshift(freal);         % same as freal(9:16 1:8)
stem(f, reorder);                  % plot reordered fs from above
set(gca, 'XTick', f);              % define frequency axis
set(gca, 'XTickLabel', fftshift(f)); % new freq labels
grid on;
```

The input and (rearranged) output signals for this example are shown in Figure 4.49. In practice, only the positive half frequencies are required to determine the relative frequency components present in a signal and such rearrangement may not always be necessary. This exercise illustrates the essential features of the MATLAB `fft()` function. It is now possible to analyze a full-wave rectified cosine.

### 4.18.2 Full-Wave Rectified Cosine (60 Hz)

Analysis of a full-wave rectified cosine may be derived from the previous example by simply repeating the entire exercise using the absolute value function on the defined cosine. The cosine frequency (60 Hz) is of no consequence to this computation except when labelling the frequency components.

```
t =0:15;                              % define 16 points
cosine = abs(cos(2*pi*t/16));         % define full-wave cosine
fs = fft(cosine)/16;                  % compute 16-point FFT
real(fs)                              % view only real parts

ans =

  Columns 1 through 4
   0.6284   -0.0000    0.2207    0.0000

  Columns 5 through 8
  -0.0518    0.0000    0.0293    0.0000

  Columns 9 through 12
  -0.0249    0.0000    0.0293    0.0000

  Columns 13 through 16
  -0.0518    0.0000    0.2207   -0.0000
```

The imaginary components in the output are all zero. These values are shown graphically in Figure 4.50. Note the presence of a significant DC component ($f = 0$). The fundamental frequency component is at $C_2$ corresponding to $2f_0$ or 120 Hz when the known information about the input signal is considered.

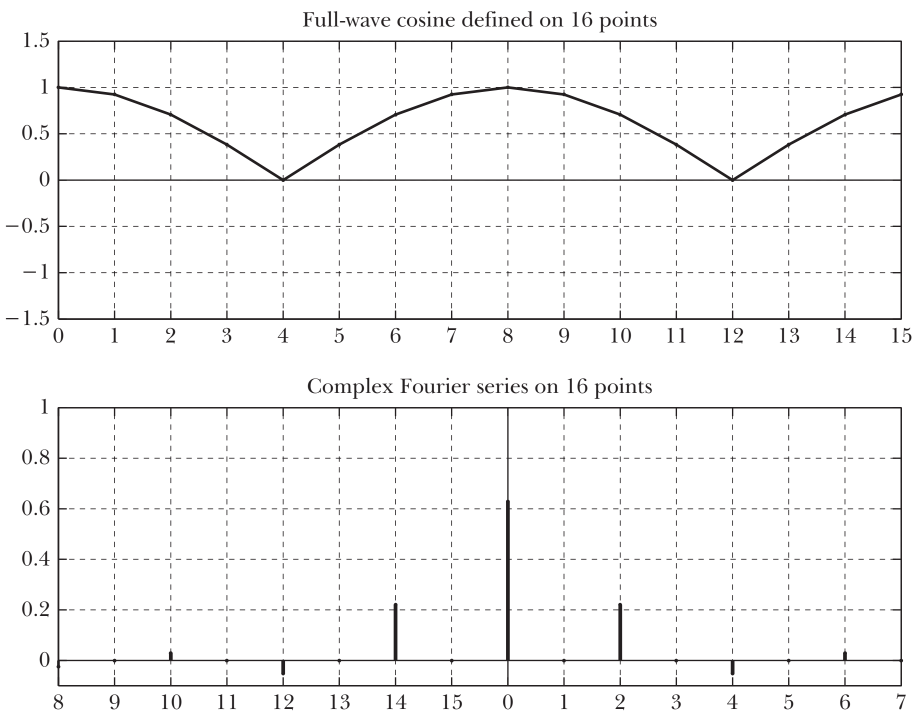



**FIGURE 4.50 MATLAB Fourier Series** The cosine of Figure 4.49 now full-wave rectified, and defined on 16 points. When reordered as shown, the output matches the expected even pairs, including a DC component, and multiple harmonics at twice the frequency of the original cosine.

The above result can be compared directly with the values calculated for the DC power supply in Section 4.17. The first few Fourier series components may be examined directly from the MATLAB result vector above, and seen on the figure. Only the real components need to be examined. Note the odd multiples of 60 Hz are all zero.

| $C_n$ | Freq. (Hz) | Value of $C_n$ |
|---|---|---|
| $C_0$ | 0 = DC | $0.6366 = 2/\pi$ |
| $C_1$ | 60 | 0 |
| $C_2$ | 120 | $0.2122 = 2/3\pi$ |
| $C_3$ | 180 | 0 |
| $C_4$ | 240 | $-0.0424 = -2/15\pi$ |

For more accuracy, the signal should be defined with more than 16 points. If these results are to be used in computations, it is important to remember that they represent only the positive frequency components $C_n$ in the complex Fourier series. Each value is only one half the amplitude of the corresponding cosine component (except DC).

### 4.18.3 Useful Hints and Help with MATLAB

*The intricacies of the fast Fourier transform* (FFT) *described below are not unique to MATLAB.*

1. **Defining odd or even signals for the Fourier series**
   When the FFT function is used, an input signal is always defined on a single period starting at the origin. This can be confusing when specifying periodic signals having symmetry around the origin. The signal in Figure 4.51 is one period of an even pulse train (a single even rectangle), yet this is not immediately obvious until the overall signal is considered. The steps used in generating this signal are shown below. Since the origin ($t = 0$) is included in the width of the leftmost region (0 through a), the total pulse width is $1 + 2a$ (where $a = 102$). The result is an even pulse train, as evidenced by the fact that there are no imaginary components in the Fourier series. Alternatively, the same periodic pulse may be placed anywhere in the defined interval and the magnitude of the Fourier series components will be unchanged, as shown in Figure 4.52.

```
t = 0:1023;                        % define 1024 points
pulse(1:1024) = 0;                 % define all = 0
pulse([1:103,923:1024]) = 1;       % define interval(s) = 1
fourier = fft(pulse) / 1024;
```

2. **Defining complete periods of a signal**
   When the FFT function is used, a signal is defined on a single period starting at the origin. This fact is easily overlooked when specifying implicitly periodic signals such as sinusoids.

   Consider the definition of a cosine with frequency 0.6 Hz on a time interval from 0 to 4 seconds. To ensure compatibility with the `fft()` function, the signal is defined on 2048 points.

```
t = linspace(0,4,2048); % define interval on 2048 points
c = cos(2*pi*0.6*t);    % define cosine (0.6 Hz)
```

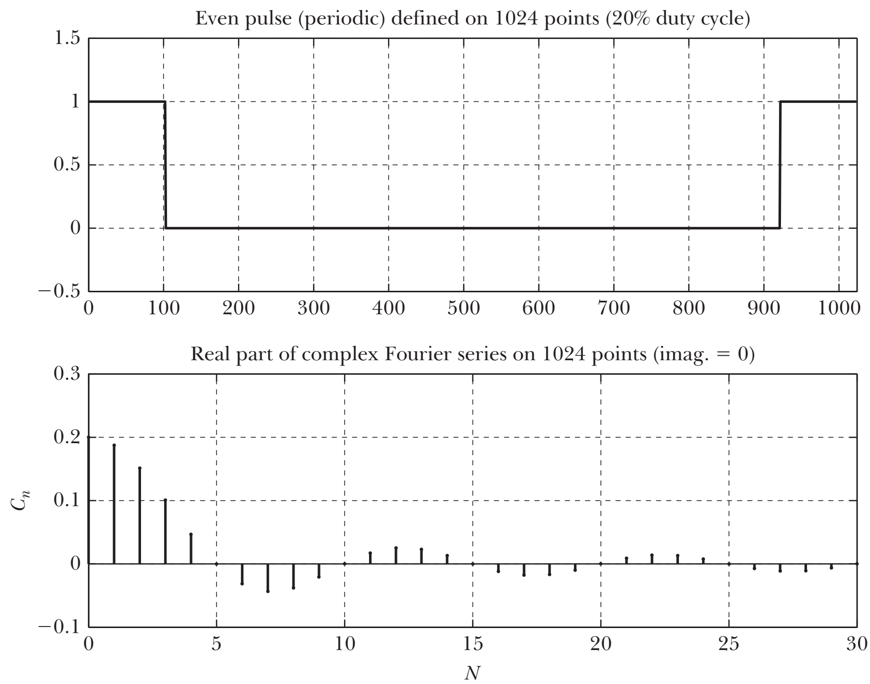



**FIGURE 4.51 An Even Pulse Train** An even pulse train with 20 percent duty cycle is defined on 1024 points in MATLAB, as one period starting at origin. The Fourier series for this signal has no imaginary components.

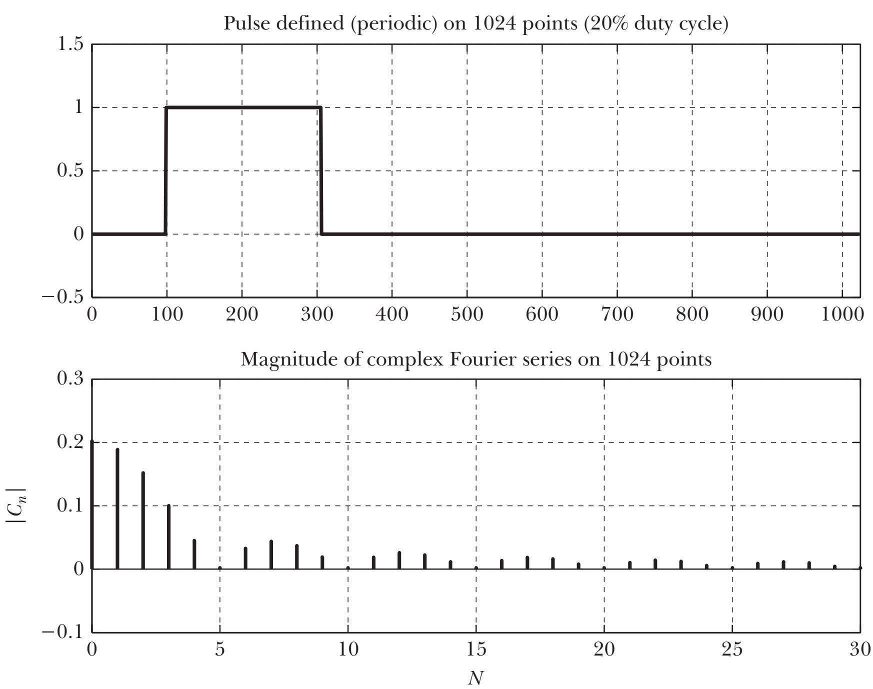



**FIGURE 4.52 Not an Even Pulse Train** A pulse train with 20 percent duty cycle is defined on 1024 points in MATLAB, as one period starting at the origin. Although this pulse train is not an even function of time, the magnitude of the Fourier series components of this signal would be the same for Figure 4.51.

This simple definition appears complete. The problem here is that this signal has been defined over a complete period equal to four seconds, and this does not define complete periods of the desired 0.6 Hz cosine. The actual periodic signal that has been specified is shown in Figure 4.53 to cover several periods. It is the overall periodic signal and not a pure 0.6 Hz cosine that will be approximated when the Fourier series components are computed, since the base period is over the interval [0, 4]. This effect can be subtle if the periodic waveform almost spans complete periods over the time in which is was defined.

This effect is most often encountered only when the frequency components are examined. In Figure 4.53, the complex Fourier series components of the specified waveform are computed. The expected pure cosine components at 0.6 Hz are only vaguely visible.

Unless complete periods of a signal are specifically defined as in the preceding examples, the above effect is generally present whenever the FFT is used.

3. **Defining signals on $2^k$ points**

Finding a specific Fourier series component $C_n$ simply involves computing the defining integral using a specific value of $n$. There is no mathematical reason to demand that an exact power-of-two number of computations be performed, or that the integral be computed on an exact power-of-two number of points.

In general, the efficiencies of the fast Fourier transform algorithm require a power-of-two number of points. The examples and applications of the complex Fourier series in this chapter have stressed the use of $2^k$ points as a fundamental constraint in using the FFT algorithm.

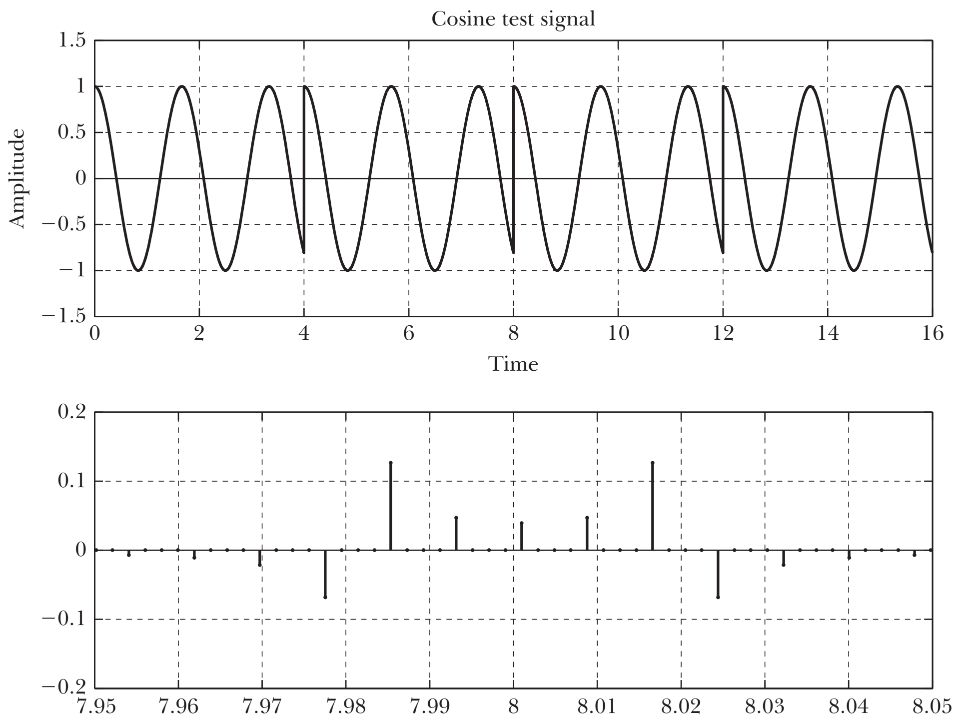



**FIGURE 4.53 Not a Sinusoid Over All Time** A 0.6 Hz cosine defined on [0, 4] seconds does not span a full period and is not sinusoidal over all time. Harmonics are evident in the Fourier series output.

In fact, the MATLAB `fft()` routine does not fail when arbitrary length vectors are used; however, it can be slower. When the MATLAB `fft()` function encounters a vector of arbitrary length, each complex Fourier series component is laboriously computed.

# 4.19 Conclusions

Any set of mutually orthogonal signals can readily be used to approximate another signal $s(t)$. Orthogonality is defined using the signal inner product formula. Like vectors, two signals are orthogonal when the signal inner product is zero; this implies that no component of one signal is present in another. The components of $s(t)$ in terms of orthogonal components can be found using the same signal inner product formula. A most useful complete set of orthogonal signals is based on the sinusoidal components defining the Fourier series. These components translate a signal $s(t)$ to a unique frequency-domain representation, where many computations prove to be greatly simplified. This technique will be extended to continuous signals in the next chapter, where a generalized approach to signals analysis based on the complex Fourier transform is developed.

# End-of-Chapter Exercises

## Orthogonal Signals

**4.1** Consider the four signals in Figure 4.54.

**(a)** Express each signal in terms of the unit rectangle.

**(b)** For each possible pairing of the four signals, identify those that are *orthogonal* on the interval $[-2, +2]$ seconds.

**(c)** Identify a linear combination of signals $\{A, C, D\}$ that gives signal $B$. Is your answer unique?

**4.2** Consider the set of orthogonal basis signals $\{a_n(t)\}$ in Figure 4.55 and the four signals $y(t)$ in Figure 4.56 defined on the interval $[-2, +2]$. In this question, each of this signals $y(t)$ is to be estimated as a linear combination of the signals $\{a_n(t)\}$.

**(a)** For each $y(t)$ determine the DC component by inspection. This is the component of $a_0(t)$ that is required.

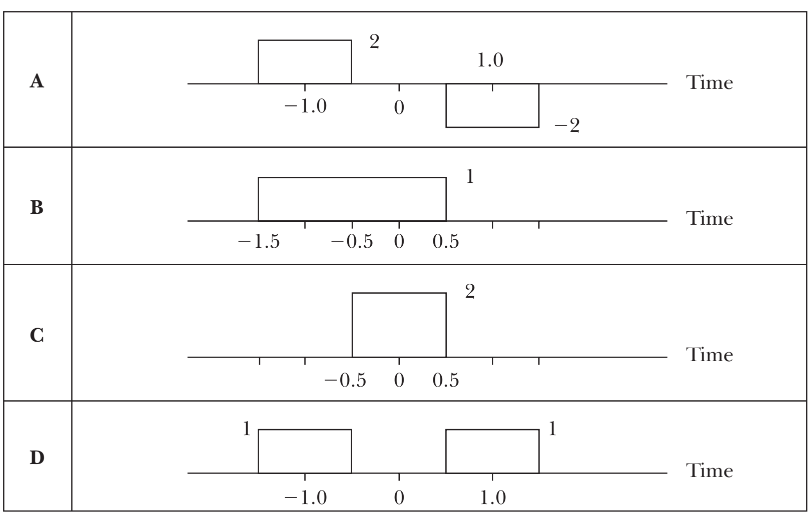



**FIGURE 4.54** Figure for Question 4.1.

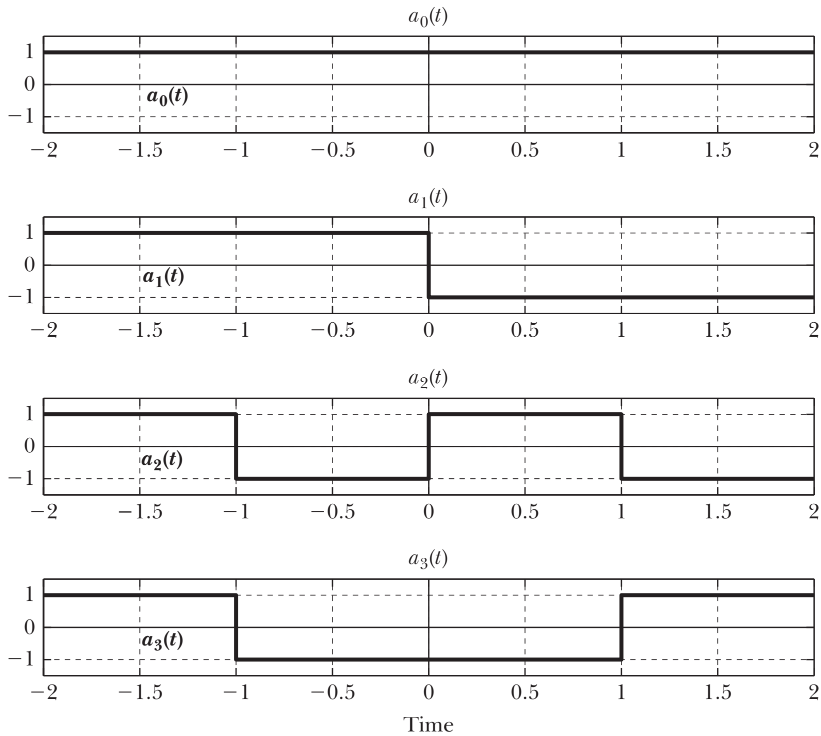



**FIGURE 4.55** Figure for Question 4.2.

**(b)** Identify each $y(t)$ as being odd or even, and eliminate by inspection any $a_n(t)$ that could not usefully contribute to $y(t)$. These terms are zero.

**(c)** Use the signal inner product as required to find the best approximation to the signal $y(t)$ using the orthogonal basis signals $\{a_n(t)\}$ in Figure 4.55. In other words, find $\{A_0, A_1, A_2, A_3\}$ and express your answer in the form of a linear combination of the signals $\{a_n(t)\}$ as:

$$y(t) \approx A_0 a_0(t) + A_1 a_1(t) + A_2 a_2(t) + A_3 a_3(t)$$

**(d)** Check your answer by reconstructing your estimate of $y(t)$ from the linear combination and sketch your approximation.

**4.3** Consider the signal $a(t) = \cos(t)$, and the signal $b(t)$, which is an even square wave with period $2\pi$ seconds and amplitude going $\pm 1V$:

**(a)** Use the signal inner product to find the best approximation to $a(t)$ in terms of $b(t)$. In other words, find B to optimize the approximation $a(t) \approx Bb(t)$.

**(b)** Use the signal inner product to find the best approximation to $b(t)$ in terms of $a(t)$. In other words find $A$ to optimize the approximation $b(t) \approx Aa(t)$.

**(c)** Should you expect $A = 1/B$? Why or why not? Discuss the difference between the values $A$ and $B$ found above.

**4.4** Sketch each of the following signal pairs on the same graph. Determine if the pair of signals are orthogonal to each other on the specified interval. Explain your answers.

**1.** $c(t) = \cos(\pi t)$ $d(t) = t^3$ $-1 \leq t \leq +1$
**2.** $c(t) = \cos(\pi t)$ $d(t) = \cos(3t)$ $-1 \leq t \leq +1$
**3.** $c(t) = \cos(t)$ $d(t) = \sin(t)$ $0 \leq t \leq +\pi/2$

## The Fourier Series

**4.5** Consider the time-shifted cosine $s(t) = \cos(t + \pi/4)$.

**(a)** Find the Fourier series representation of the signal $s(t)$.

**(b)** Sketch the magnitude of the Fourier series components of $s(t)$.

**(c)** Find the constants $A$ and $B$ such that $s(t) = A\cos(t) + B\sin(t)$.

**(d)** Sketch the the Fourier series components of $s(t)$ as real and imaginary parts.

**4.6** A periodic signal $s(t)$ can be expressed as a Fourier series of the form:

$$s(t) \approx A_0 + \sum_{n=1}^{+\infty}[A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t)]$$

Derive an expression for the total power in this signal $s(t)$ in terms of the $A_n$ and $B_n$. (Include $A_0$.)

**4.7** Determine a simplified complex Fourier series representation for the signal $s(t) = \cos^2(3t)$. Sketch this signal in the time and frequency domains. From the frequency-domain sketch, find $V_{rms}$ and identify $V_{dc}$.

**4.8** Two signals $m(t) = \cos(2\pi t)$ and $c(t) = \cos(200\pi t)$ are multiplied together. Accurately sketch the resulting signal $w(t) = m(t) \times c(t)$ in the frequency domain (i.e., as components in the complex Fourier series).

**4.9** Find the $V_{rms}$ value of the periodic voltage signal $g(t) = 1 + \cos(t) + \sin(3t)$. Sketch this signal in the frequency domain.

**4.10** Repeat the analysis of the DC power supply, but use a half-wave rectifier. Compare each answer to the full-wave rectifier studied in the text. Check your results using MATLAB.

**(a)** What is the DC component of this signal?

**(b)** What is the $V_{rms}$ value of this signal?

**(c)** How is the $V_{rms}$ value related to the DC component?

**(d)** What is the Fourier series expansion of this waveform?

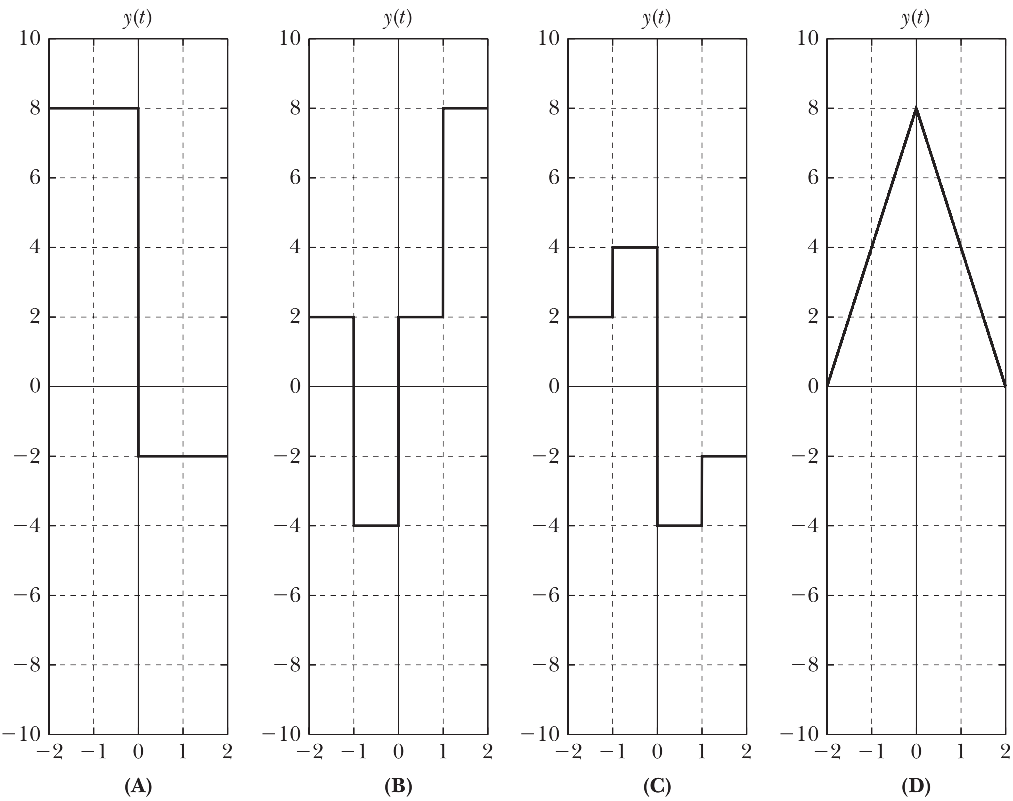



**FIGURE 4.56** Figure for Question 4.2.

**(e)** What fraction of the total power is contained in the DC component?

**(f)** Would this signal be more difficult to filter than the full-wave rectified case? Explain.

**4.11** Following the example of Section 4.18.2, use MATLAB to find the Fourier series components of the same cosine, half-wave rectified.

**(a)** What is the DC component of this signal?

**(b)** What element in the output from `fft()` corresponds to the DC component?

**(c)** Plot the frequency components.

**4.12** Following the example of Section 4.18.2, use MATLAB to find the Fourier series components of a sine wave.

**(a)** Describe the expected components as odd or even, real or imaginary.

**(b)** Plot the frequency components.

**4.13** Consider a sawtooth wave (no DC offset) with Period = 1 s, and answer the following questions.

**(a)** Compute $V_{rms}$ and the total power in the original signal.

**(b)** Find the Fourier series approximation for the signal as a function of N.

**(c)** Accurately sketch the signal and the fundamental frequency component on the same graph.

**(d)** Compute the power in each of the first three sinusoidal components.

**(e)** State, as a percentage of total power, the goodness of the Fourier series approximation for one, two, and three components.

**4.14** Complete Question 4.13 with a triangle wave. Observe that a properly sketched sinusoid is actually quite triangular in form.

**4.15** Complete Question 4.13 with a pulse train with 10 percent duty cycle (pulses go 0–1 Volts).

**4.16** Discuss the statement "*A sawtooth wave cannot be approximated by a Fourier series consisting only of cosine components.*"

**4.17** A digital system is used to generate a crude cosine using the pulses as shown in Figure 4.57 alternating ±1 Volt. The overall signal $s(t)$ has the appearance of a stepped waveform roughly approximating a sinusoid.

**(a)** What is $V_{rms}$ value for $s(t)$?

**(b)** What is the total power in $s(t)$?

**(c)** Compute the Fourier series component of the fundamental frequency only.

**(d)** Sketch $s(t)$ and the fundamental frequency component on the same graph.

**(e)** Compare the power in the fundamental component to the total power in the harmonics.

**4.18** Use MATLAB to analyze the waveform in Figure 4.57 as an approximation to a cosine.

**(a)** Define $s(t)$ piecewise on 1024 points (one period starting at the origin, one second is 128 points).

**(b)** Use the `fft()` function to find the Fourier series components.

**(c)** Print frequency components up to $N = 10$ as real and imaginary terms.

**(d)** Relate the real and imaginary components to the sinusoidal Fourier series components.

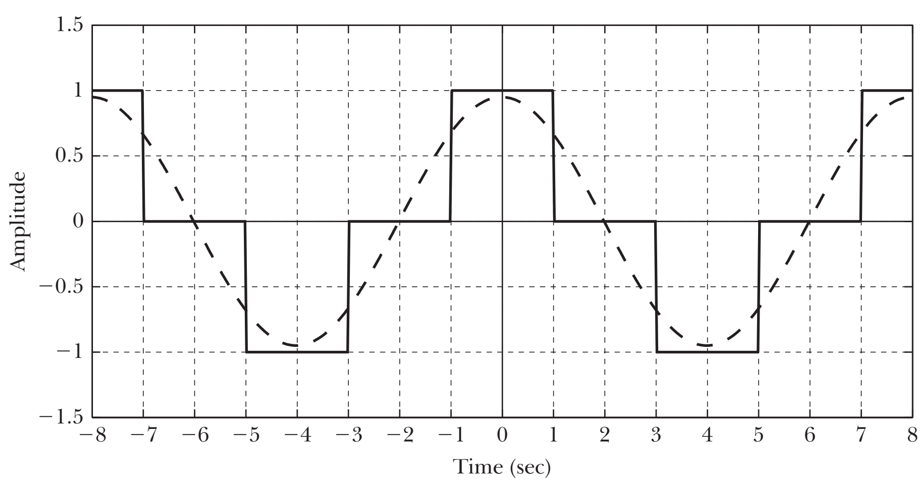



**FIGURE 4.57** Figure for Questions 4.17 and 4.18.

Find the phase and magnitude of each of the first 10 terms.

**(e)** Generate a graph of $s(t)$ with the fundamental frequency component on the same graph.

**(f)** Generate a graph of $s(t)$ with the sum of components up to $N = 3$ on the same graph.

**(g)** Generate a graph of $s(t)$ with the sum of components up to $N = 5$ on the same graph.

**(h)** Generate a graph of $s(t)$ with the sum of components up to $N = 7$ on the same graph.

**4.19** A 1 kHz square wave having amplitude 1 Volt passes through a linear system capable of passing any frequency up to 2 kHz. Does the square wave pass through unaffected? Explain. If the system is an ideal lowpass filter, give an exact expression for the output signal.

**4.20** A signal $m(t)$ has the frequency spectrum of Figure 4.58.

**(a)** Describe the time-domain signal $m(t)$ as (real, imaginary, both).

**(b)** Describe the time-domain signal $m(t)$ as (odd, even, neither).

**(c)** What is the period of $m(t)$?

**(d)** What is the power in $m(t)$?

**(e)** What is the DC offset in $m(t)$?

**(f)** What is $V_{rms}$ for $m(t)$?

**(g)** Give a simplified expression for $m(t)$ in the time domain.

**(h)** The signal $m(t)$ passes through a linear system that only lets through frequencies up to 2 Hz (lowpass filter). Sketch the output signal as seen on an oscilloscope.

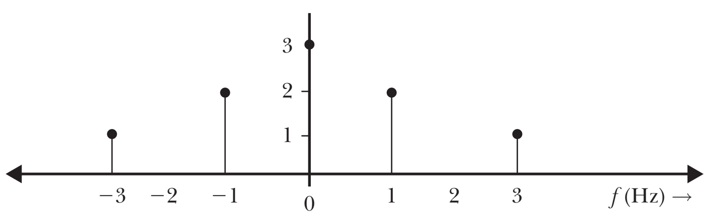



**FIGURE 4.58** Figure for Question 4.20.

**4.21** A periodic signal $m(t)$ with a fundamental frequency of 1000 Hz has the following Fourier series components:

| $C_{-3}$ | $C_{-1}$ | $C_0$ | $C_1$ | $C_3$ |
|---|---|---|---|---|
| 0.5j | 0.5 | 1 | 0.5 | −0.5j |

**(a)** Sketch this signal in the frequency domain.

**(b)** Describe the time-domain signal $m(t)$ as (real, imaginary, both).

**(c)** Describe the time-domain signal $m(t)$ as (odd, even, neither).

**(d)** What is the power in $m(t)$?

**(e)** What is the DC offset in $m(t)$?

**(f)** What is $V_{rms}$ for $m(t)$?

**(g)** Give a simplified expression for $m(t)$ in the time domain.

**(h)** The signal $m(t)$ passes through a linear system that only lets through frequencies up to 2000 Hz (lowpass filter). Sketch the output signal as seen on an oscilloscope.

**4.22** Consider the voltage signal $a(t) = 20\cos(t)$ and its Fourier series representation $A(f)$.

**(a)** Complete the following table for the signal $a(t)$:

| Parameter | Value |
|---|---|
| Amplitude | |
| Frequency | |
| Period | |
| Power | |
| DC Component | |
| $V_{rms}$ | |
| Energy | |

**(b)** By inspection, make a neat labelled sketch of the Fourier series $A(f)$.

**(c)** Sketch the magnitude of the Fourier series for the signal $a(t-2)$.

**4.23** Consider the voltage signal $a(t) = 10 + 10\cos(100\pi t) - 10\sin(200\pi t)$ and its Fourier series representation $A(f)$.

**(a)** Complete the following table:

| Parameter | Value |
|---|---|
| DC component of $a(t)$ | |
| Value of $a(t)$ at $t = 0$ | |
| Period of $a(t)$ | |
| Fundamental frequency of $A(f)$ | |
| Value of $A(f)$ at $f = 0$ | |
| Magnitude of $A(f)$ at $f = -100$ Hz | |
| Power in $A(f)$ | |

**(b)** By inspection, make a neat labelled sketch of the Fourier series $A(f)$.

**(c)** Use MATLAB to plot the signal $a(t)$ and confirm this result against the above answers.

**(d)** Use MATLAB to plot the Four series $A(f)$ and confirm this result against the above answers.

**4.24** A low voltage cosine signal is amplified to have amplitude 1 V but is *clipped* as shown in Figure 4.59 because the amplifier cannot produce voltages exceeding $\pm 0.707$ V. The figure shows the clipped cosine $s(t)$ with the expected undistorted output (dashed).

**(a)** What will be the odd frequency components $B_n$ of the Fourier series?

**(b)** What is the DC component of the Fourier series?

**(c)** From the graph, give an expression for one period of the clipped cosine $s(t)$ on five intervals.

**(d)** Find the Fourier Series components of the clipped cosine $s(t)$.
*Hint*: The computation of the Fourier series components can be greatly simplified for this even signal defined on the intervals as shown.

**4.25** A low voltage cosine signal is amplified to have amplitude 1V but is *clipped* as shown in Figure 4.59 because the amplifier cannot produce voltages exceeding $\pm 0.707$ V. The figure shows the clipped cosine $s(t)$ with the expected undistorted output (dashed).

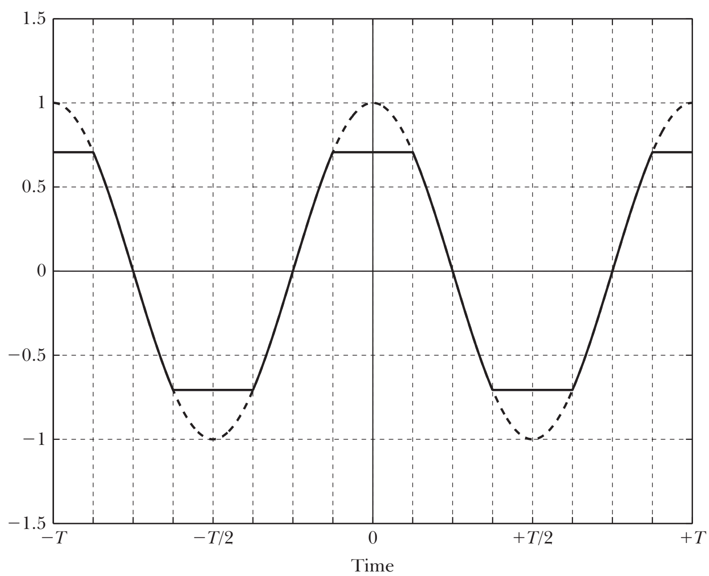



**FIGURE 4.59** Figure for Questions 4.24 and 4.25.

**(a)** Define the signal $s(t)$ on 2048 points using MATLAB. The commands `min(0.707,s)` and `max(-0.707,s)` may be useful to accomplish the clipping by constraining the signal `s` to the upper and lower voltage limits.

**(b)** Find the Fourier series components of the clipped signal using the `fft()` command.

**(c)** Which elements of the `fft()` output hold the fundamental frequency components? Note that the signal as defined spans two periods.

**(d)** What is the amplitude and power of the fundamental frequency component?

**(e)** What is the power in the signal for all time? The command `sum()` can be used to add all the elements of a vector.

CHAPTER 5

# The Fourier Transform

In this chapter, other orthogonal sets of signals are set aside to concentrate on the complex exponential $e^{+j2\pi ft}$. The complex Fourier series for periodic signals is logically extended into a continuous function of frequency suitable for both periodic and nonperiodic signals. A thorough study of the continuous Fourier transform reveals the properties that make this transform technique a classic tool in signals analysis. By relating the time-domain and frequency-domain behavior of signals, the Fourier transform provides a unique perspective on the behavior of signals and systems. The Fourier transform also forms the basis for the Laplace transform in Chapter 7 and the *z*-transform of Chapter 9.

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Explain how the Fourier series extends logically into nonperiodic signals
- Define the orthogonal basis for a Fourier transform representation
- Write the equations for the Fourier transform and its inverse
- Compute the Fourier transform of common time-domain functions
- Recognize common Fourier transform pairs
- Identify graphically the links between a signal and its Fourier transform pair
- Explain how a Fourier transform is affected by time-domain variations (shifting, scaling)
- Apply the rules of the Fourier transform to predict the Fourier transform of an unknown signal

## 5.1 Introduction

In Chapter 4, the concept of orthogonal functions was introduced, from which it was shown that an arbitrary signal could be represented by a unique linear combination of orthogonal component signals defined over a given time interval. The *best approximation* to arbitrary signals defined over a fixed interval was demonstrated for various sets of orthogonal components. Finally, the Fourier series was studied as a complete set of periodic orthogonal signals $\{\sin(2\pi nf_0 t), \cos(2\pi nf_0 t)\}|n\in I+\}$ defined over a single period $T = 1/f_0$ s. By extension, the properties of the complex Fourier series, based on the complete orthogonal set $\{e^{+j2\pi nf_0 t}|n\in I\}$ proved to offer significant advantages and to provide valuable insights into the representation of periodic signals by components. In this chapter, the Fourier series is extended to nonperiodic signals, resulting in the continuous function of frequency known as the Fourier transform.

### 5.1.1 A Fresh Look at the Fourier Series

A periodic signal $s(t)$ with period $T = 1/f_0$ can be approximated by a unique linear combination of the orthogonal Fourier series components. In effect, this period $T = 1/f_0$ defines an orthogonal set $\{e^{+j2\pi n f_0 t} | n \in I\}$ upon which to base the approximation to $s(t)$, and this period also defines the interval over which the above set is orthogonal. Recall that the definition of orthogonality was strictly limited to the interval for which it is defined. The above set of sinusoids proves to be orthogonal over one period $T$ and over all integer multiples $nT$ of the period. Furthermore, for periodic signals, computation of the Fourier series components carried out over a single period of the signal implicitly describes the periodic waveform $s(t)$ for all time.

**Periodic and Nonperiodic Signals** Because the Fourier series components $\{\sin(2\pi n f_0 t), \cos(2\pi n f_0 t)\} | n \in I+\}$, with $T = 1/f_0$ are orthogonal on the interval $[-T/2, +T/2]$, it follows that any signal can be approximated *within that interval* even if the signal is not periodic. For periodic signals with period $T$, the use of sinusoidal components assures that the same approximation describes the periodic waveform for all time. For nonperiodic signals, the linear combination of these components that gives a best approximation within the interval may not resemble the original signal at all outside of that interval. This statement is demonstrated in the next example.

**EXAMPLE 5.1 (Fourier Series: The First Two Terms)**

*Use the first two terms in the Fourier series to find an approximation to the signal* $s(t) = t^2$, *which is valid within the interval* $[-\pi, +\pi]$.

**Solution:**

The signal $s(t)$ is even, and computation of the Fourier series coefficients leads to the best approximation within the interval $[-\pi, +\pi]$ for the DC and fundamental frequency component:

$$A_0 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} t^2 \; dt = \frac{\pi^2}{3} = 3.29$$

and

$$A_1 = \frac{2}{2\pi} \int_{-\pi}^{+\pi} t^2 \cos(t) \; dt = -4$$

giving the first approximation:

$$s(t) = t^2 \approx \frac{\pi^2}{3} - 4\cos(t)$$

This signal is shown graphically in Figure 5.1. The nonperiodic signal $s(t)$ is well modelled within the interval $[-\pi, +\pi]$. It is clear that this approximation to $s(t)$ will apply only within the interval $[-\pi, +\pi]$ as the two signals are quite different otherwise.

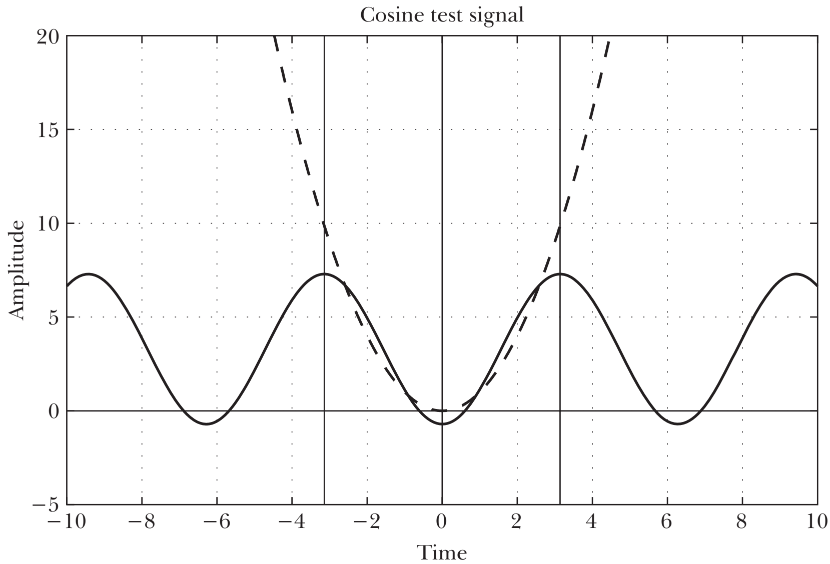



**FIGURE 5.1** Fourier series components of a nonperiodic signal $s(t) = t^2$. The approximation is valid on the interval $[-\pi, +\pi]$ for which it was defined.

This example shows that approximation of a signal $s(t)$ using the Fourier series over a given interval of width $T$ presupposes that the signal $s(t)$ is periodic with the same period $T$. On the other hand, suppose that:

- A signal $s(t)$ exists only within a certain interval; or
- Any study of the behavior of $s(t)$ is to be limited to that interval; and
- Approximation to $s(t)$ by orthogonal components is never expected to apply outside of the interval.

Under these conditions, it is acceptable to apply the periodic Fourier series approximation to a nonperiodic signal. Of course, once this decision is reached, the above assumptions and limitations must be clearly recognized and always respected if results are to have any useful meaning. These same underlying conditions may also explain properties of such an approximation that might otherwise defy explanation.

The fast Fourier transform (FFT) is based on the complex Fourier series, yet, as in the above example, this transform technique is routinely applied to nonperiodic signals. Unusual or unexpected results can often be traced to this important distinction.

### 5.1.2 Approximating a Nonperiodic Signal over All Time

In this section, the Fourier series will be used to represent a single rectangle centered on the origin. As in the previous example, the approximation will apply only over a fixed interval of integration corresponding to one period in a pulse train. The key concept in determining the Fourier components of a nonperiodic signal is to repeat this computation while letting the period go to infinity.[1]

[1] Conversely, an even pulse train includes a rectangle centered at the origin, but if its period is made very long, that single rectangle becomes the only signal of interest.

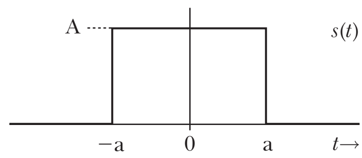



**FIGURE 5.2** This real and even rectangle $s(t)$ has width $= 2a$, height $= A$, and area $= 2Aa$.

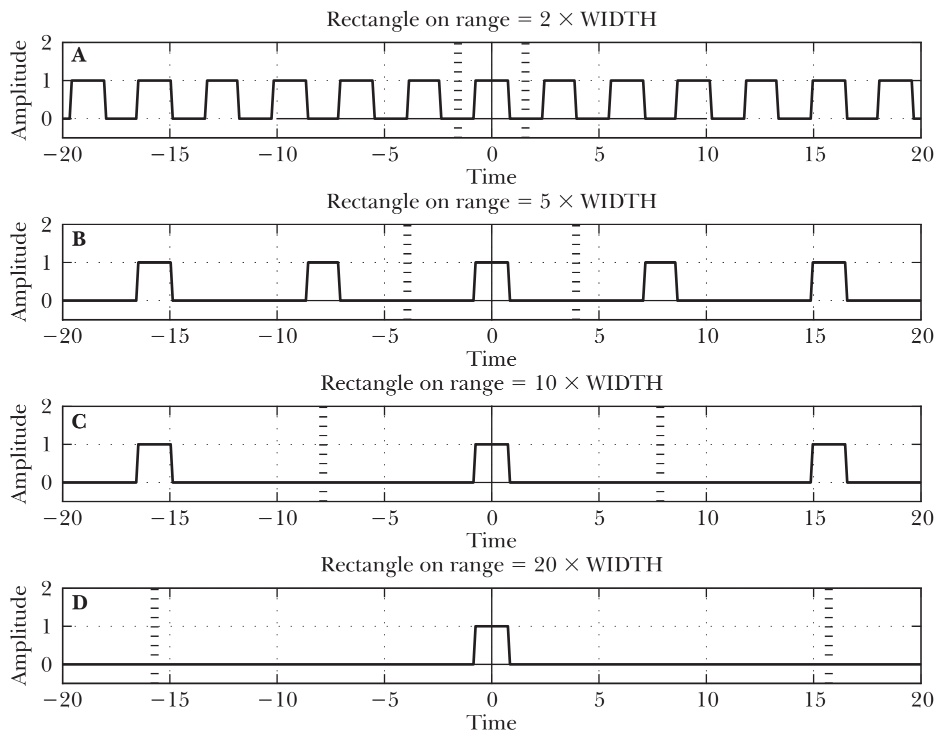



**FIGURE 5.3** For the same even rectangle, the Fourier series components can be recalculated as the interval of integration (period) increases. In the limit, a single rectangle lies at the origin. See Figure 5.4.

Consider the even rectangle $s(t)$ in Figure 5.2 having width $2a$, height $A$, and area $2Aa$. Define an interval $[-T/2, +T/2]$ seconds, where $|T/2| > a$. A Fourier series approximation valid on this interval will be developed. Note that $s(t) = A$ in the interval $[-a, +a]$, and zero otherwise. The Fourier series components (including only even coefficients $A_n$) can be expressed as (with $T = 1/f_0$):

$$A_n = \frac{2}{T} \int_{-a}^{+a} A \cos(2\pi n f_0 t) \; dt$$

$$A_n = \frac{4}{T} \int_{0}^{+a} A \cos(2\pi n f_0 t) \; dt$$

$$A_n = \frac{4}{T} \frac{A}{2\pi n f_0} \sin(2\pi n (a/T))$$

leaving:

$$A_n = \frac{2A}{n\pi} \sin(2\pi n (a/T)) \tag{5.1}$$

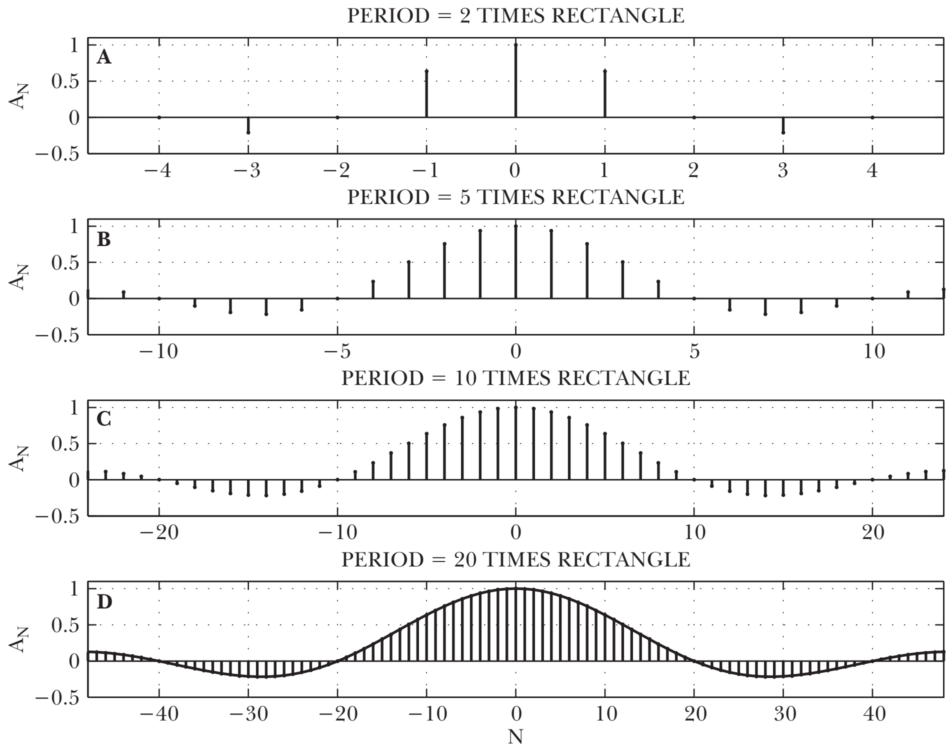



**FIGURE 5.4** For the same even rectangle, the Fourier series components lie closer together as the interval of integration (period) increase. In the limit, a continuous function of frequency describes a distinctive shape known as a *sinc function*. See Figure 5.3.

The terms $A_n$ in this Fourier series expansion depend only on the ratio $a/T$. If the period $T$ is twice the width of the rectangle $T = 4a$, the interval $[-2a, +2a]$ exactly defines the computation of the Fourier series representation for a square wave with period $4a$. The resulting Fourier series components reflect the now-familiar odd-numbered integer multiples of $\cos(2\pi n f_0 t)$ with alternating sign, as seen in Figure 5.4A.

While the above approximation closely matches the single rectangle $s(t)$ over the given interval, it also (implicitly) matches a complete square wave with period $T = 4a$, as shown in Figure 5.4A. If the goal was to represent only the one rectangle, this approximation is wildly wrong outside the interval $-2a < t < +2a$, where the function describing the single rectangle is zero. To obtain a better approximation to a nonperiodic rectangle, the same Fourier series may be applied using a wider interval of integration. By increasing the interval of integration without changing the width of the rectangle, the resulting approximation will apply over a wider interval of time. The same formula defines the Fourier series components $A_n$, but their spacing decreases as multiples of $f_0 = 1/T$, as shown in Figure 5.4B for a width five times the rectangle. Note especially the zero crossings on the figure, which do not change position (the null occurs whenever the sine term is zero) even as the Fourier series components appear more closely spaced in frequency. The amplitude of each component is normalized in the figure with a factor $T$. In Figures 5.4C and 5.4D, the interval is increased again, while keeping the same rectangle at the origin. An important trend is evident in these figures, and the final figure shows components that are clearly tending to a continuous function of frequency (scaled by $T$), just as the time-domain signal appears more like an isolated rectangle.

As the interval goes to infinity, the Fourier series components tend to a continuous function of frequency to be labelled $S(f)$ and called the *Fourier transform* of the rectangle signal. For the rectangle $s(t)$ in this example, that function of frequency $S(f)$ will take the distinctive form known as a *sinc* function. Furthermore, this same $S(f)$ could be worked back to obtain the Fourier series components directly. The sinc function is to be studied in some detail in Section 5.4.

It has been demonstrated that when the interval of integration doubles, each of the Fourier series components falls by one half. This effect was corrected in the graphs in Figure 5.4 by multiplying by a factor varying as $T$. Still, the continuous Fourier transform resulting when $T$ tends to infinity would tend to zero without this correction. Investigation of this factor will reveal a fundamental difference between the Fourier series and the Fourier transform.

Consider the time-domain pulse train described by a fixed-width rectangle repeating with period $T$. When the interval of integration doubles, the total power is halved. (Even though there are twice as many terms $A_n$ in any fixed interval along the horizontal axis, the power contributed by each falls to $1/4$.) The total power in this periodic signal varies as $1/T$. The RMS value of $s(t)$ also falls. Furthermore, the DC component tends to zero as $T$ goes to infinity. In the limit, the power in this solitary rectangle, lying alone in a vast interval stretching to infinity in both directions, approaches zero.

A periodic square wave or pulse train is defined as a power signal, having infinite energy. On the other hand, the single rectangle $s(t)$ is by definition an energy signal, having zero power. Consequently, Fourier transform computations involving non-periodic signals will address total energy, whereas power considerations were always invoked when studying the Fourier series.

### 5.1.3 Definition of the Fourier Transform

The complex Fourier series defines a signal $s(t)$ with period $T = 1/f_0$ as complex components $\{C_n\}$, where:

$$C_n = \frac{1}{T}\int_{-T/2}^{+T/2} s(t)e^{-j2\pi n f_0 t}\,dt$$

Likewise, if the spacing between components at $nf_0$ is made arbitrarily small, the discrete coefficients $C_n$ merge into the continuous function of frequency $S(f)$, as:

**DEFINITION 5.1 Fourier Transform**

*Let* s(t) *be a function of* t, *then*

$$S(f) = \int_{-\infty}^{+\infty} s(t)e^{-j2\pi ft}\,dt$$

*is called the* Fourier transform *of* s(t).

The Fourier transform is often defined in terms of $\omega = 2\pi f$ as:

$$S(\omega) = \int_{-\infty}^{+\infty} s(t)e^{-j\omega t}\,dt \tag{5.2}$$

Like the coefficients $C_n$, the function $S(f)$ is a function of frequency that uniquely represents $s(t)$; the Fourier transformation is a one-to-one mapping, and the two functions $s(t)$ and $S(f)$ are often called a *Fourier transform pair.* As a convention, time-domain signals such as $s(t)$ will be written using lowercase letters, and frequency-domain signals such as $S(f)$ will use the corresponding uppercase letters. The notation

$$s(t) \overset{\mathcal{F}}{\longleftrightarrow} S(f)$$

will be used to indicate that $s(t)$ and $S(f)$ are related through the Fourier transform integral.[2] Where there is no chance of confusion, the simple notation $s(t) \longleftrightarrow S(f)$ can be used. Note that this is not an equality; the double arrow indicates the unique relationship of a Fourier transform pair. The form

$$\mathcal{F}[s(t)] = S(f)$$

is also seen.

## 5.1.4 Existence of the Fourier Transform

For the Fourier transform of a signal $s(t)$ to exist, it is sufficient that $s(t)$ be absolutely integrable, or that $s(t)$ should satisfy the condition:

$$\int_{-\infty}^{+\infty} |s(t)| \; dt < \infty$$

In general, real-world signals of interest to engineers and scientists will satisfy the general conditions allowing their integration (e.g., that they possess a finite number of discontinuities); however, there are a number of signals that by nature will not converge (such as $s(t) = t^2$), and in practice the issue of convergence is of most interest. Functions that do not have a Fourier transform may nonetheless have a Laplace transform as will be seen in Chapter 7.

It can be noted that the above condition is violated by a simple constant $s(t) = A$ and by periodic functions such as $s(t) = \cos(t)$; however, such signals do not really exist *for all time* as their definitions suggest, and these Fourier transforms would include the delta function $\delta(t)$, which itself has an implausibly infinite amplitude. It is safe to say that such periodic signals do possess a frequency spectrum and that the existence of the Fourier transform *in the limit* justifies its use.

## 5.1.5 The Inverse Fourier Transform

The complex Fourier series defines a signal $s(t)$ with period $T = 1/f_0$ as complex components $\{C_n\}$, where the signal is constructed as the addition of all components over all frequencies:

$$s(t) = \sum_{n=-\infty}^{+\infty} C_n e^{+j2\pi n f_0 t}$$

[2] The letters $\mathcal{F}$, $\mathcal{L}$, and $\mathcal{Z}$ will specify the Fourier, Laplace, and *z*-transforms, respectively.

Likewise, the frequency components of the Fourier transform $S(f)$ of a signal $s(t)$ combine as the integration over all frequencies to give:

$$s(t) = \int_{-\infty}^{+\infty} S(f)e^{+j2\pi ft}\,df \tag{5.3}$$

which is called the *inverse Fourier transform* of the signal $S(f)$. When the inverse Fourier transform is expressed in terms of $\omega = 2\pi f$, a factor of $1/2\pi$ is introduced, as:

$$s(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} S(\omega)e^{+j\omega t}\,d\omega \tag{5.4}$$

The use of $\omega$ rather than $2\pi f$ is commonplace in some application areas, and while $2\pi f$ will continue to be used in this chapter, the use of $\omega$ will be preferred for the Laplace transform and the *z*-transform in later chapters.

## 5.2 Properties of the Fourier Transform

The properties of the Fourier transform follow directly from those of the complex Fourier series that were explored in Chapter 4. For example, it is reasonable to expect that the Fourier transform of a real even function will be real and even (a demonstrated property of the complex Fourier series). The Fourier transform may be considered as the more general form, of which the Fourier series is a special case applying specifically to periodic signals. On the other hand, the Fourier transform has many more useful properties, which must be carefully examined in order to fully understand the Fourier transform and its applications. Many properties of the Fourier transform and its inverse can be revealed by manipulating the defining equations. This procedure establishes a firm mathematical basis for each of the properties and all the rules relating the behavior of signals expressed in one domain or the other. To illustrate this approach, some useful properties will be introduced immediately, after which the Fourier transform of the real nonperiodic rectangle signal will be computed.

### 5.2.1 Linearity of the Fourier Transform

The Fourier transform is called a linear transformation since the Fourier transform of a linear combination of signals equals the linear combination of their respective Fourier transforms. This observation may be written as:

**THEOREM 5.1** *(Linearity)*

*If*

$$s_1(t) \xleftrightarrow{\mathcal{F}} S_1(f)$$

*and*

$$s_2(t) \xleftrightarrow{\mathcal{F}} S_2(f)$$

*then for constants* $k_1$ *and* $k_2$

$$[k_1 s_1(t) + k_2 s_2(t)] \xleftrightarrow{\mathcal{F}} [k_1 S_1(f) + k_2 S_2(f)]$$

**Proof:**

This result follows directly from the definition of the Fourier transform and the inherent linearity of integration:

$$
\begin{aligned}
LHS &= \int_{-\infty}^{+\infty} [k_1 s_1(t) + k_2 s_2(t)] e^{-j2\pi ft} \, dt \\
&= k_1 \int_{-\infty}^{+\infty} s_1(t) e^{-j2\pi ft} \, dt + k_2 \int_{-\infty}^{+\infty} s_2(t) e^{-j2\pi ft} \, dt \\
&= k_1 S_1(f) + k_2 S_2(f) = RHS
\end{aligned}
$$

The same approach can be applied to the inverse Fourier transform. The linearity of the Fourier transform is fundamental to its use in signals analysis. It implies that if a signal $g(t)$ can be represented as some linear combination of known signals $\{s_n(t)\}$, then the Fourier transform $G(f)$ can be determined directly from the $\{S_n(f)\}$ without further computation. This technique can be applied either graphically or directly by computation.

## 5.2.2 Value of the Fourier Transform at the Origin

Consider the defining formula for the Fourier transform $S(f)$:

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft} \, dt$$

and let $f = 0$:

$$S(0) = \int_{-\infty}^{+\infty} s(t) \, dt$$

In other words, the total area under the time-domain function $s(t)$ is given by the value of the frequency-domain function $S(f)$ at the origin $(f = 0)$.

Similarly, for the inverse Fourier transform, computed for $t = 0$:

$$s(0) = \int_{-\infty}^{+\infty} S(f) \, df$$

The total area under the frequency-domain function $S(f)$ is given by the value of the time-domain function $s(t)$ at the origin $(t = 0)$.

These results are summarized in Figure 5.5. The significance of this observation also has wide-ranging consequences. In this example, nothing was said about what particular signal $s(t)$ represented, nor was it necessary to know what $S(f)$ actually

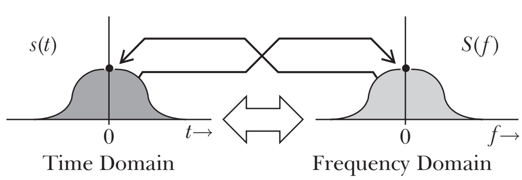



**FIGURE 5.5** The value at the origin in one domain equals the area under the corresponding signal in the other domain.

looked like. Still, it is clear that, given a signal $s(t)$, both the value of its transform at the origin and the integral of that transform (the area under its graph), can readily be determined, often by inspection.

### 5.2.3 Odd and Even Functions and the Fourier Transform

The complex Fourier series for a real periodic signal $s(t)$ has been shown to consist of components having real part even and imaginary part odd. The same behavior will characterize the continuous Fourier transform. This observation comes directly from expansion of the Fourier transform integral:

$$\begin{aligned} S(f) &= \int_{-\infty}^{+\infty} s(t)e^{-j2\pi ft}\,dt \\ &= \int_{-\infty}^{+\infty} s(t)\cos(2\pi ft)\,dt - j\int_{-\infty}^{+\infty} s(t)\sin(2\pi ft)\,dt \end{aligned}$$

If the Fourier transform of a real and even $s(t)$ is evaluated, the product term in $s(t)\sin(2\pi ft)$ will be odd, and the imaginary part of $S(f)$ given by that integral will be zero. The resulting $S(f)$ is purely real, and by inspection, $S(f) = S(-f)$; therefore, $S(f)$ is even. Conversely, if the signal $s(t)$ is real and odd, only the second integral will survive, leaving a purely imaginary $S(f)$, which is readily shown to be odd. These properties are summarized in Table 5.1. Recall that all signals $s(t)$ measurable in the laboratory are real. Furthermore, for any real signal $s(t)$, the magnitude of $S(f)$ will be real and even.

The properties in Table 5.1 demonstrate that given a signal in either domain, $s(t)$ or $S(f)$, it will be possible to predict and to determine various properties in the other domain without necessarily calculating the (inverse) Fourier transform. With such foresight, the actual Fourier transform could be computed with some confidence, knowing what result was expected. A basic knowledge of such properties is key to understanding the Fourier transform.

Note that the properties of odd and even signals and their Fourier transforms apply symmetrically to time-domain or frequency-domain signals, relating $s(t)$ to $S(f)$ and vice versa. Sometimes it more natural to think of signals in the time domain, although certain properties may be easier to appreciate from a frequency perspective. In either case, the corresponding property can be readily applied to either domain. Often, the above relationships are generalized to refer to the

**TABLE 5.1**

**Properties of the Fourier Transform**

| Time Domain: $s(t)$ | Fourier Transform: $S(f)$ |
|---|---|
| Real & even | Real & even |
| Real & odd | Imaginary & odd |
| Real & neither | Real part even & imaginary part odd |
| Imaginary & even | Imaginary & even |
| Imaginary & odd | Real & odd |
| Imaginary & neither | Imaginary part even & real part odd |

*transform domain*, another way of saying *the other domain*. These *mirror image* properties are not too surprising, as the defining equations for the Fourier transform and its inverse are nearly identical. It can be stated that:

> *In general, properties relating the time-to-frequency-domain representations will also apply to the inverse frequency-to-time-domain relationships. At most, there will be a change of sign in the inverse formulation.*

Where there is a difference in the forward or reverse property, it is found with odd or neither functions because of the sign change in the complex exponential. Recall that the forward and inverse Fourier transform equations differ only in the sign of the exponential.

## 5.3 The Rectangle Signal

While periodic signals necessarily have components at integral multiples of a fundamental frequency (as in the Fourier series), nonperiodic signals will have continuous Fourier transforms. The rectangle signal can be expected to have a continuous transform. It can also be predicted that an even rectangle $s(t)$ will possess a Fourier transform $S(f)$ that is real and even.

Consider the time-domain rectangle $s(t)$ from Figure 5.6. This is a general rectangle, and it can be noted that $s(t) = A\text{rect}(t/2a)$. Before computing the Fourier transform of any signal, several observations can (and should) be made relating $s(t)$ to the expected result $S(f)$. In this case:

1. This real and even rectangle $s(t)$ should possess a real and even Fourier transform $S(f)$;
2. The area of the rectangle $= 2Aa$; this should be the value of $S(0)$;
3. The value of the rectangle at the origin $s(0) = A$; this should equal the area under $S(f)$.

The Fourier transform of the above rectangle $s(t)$ is now calculated as:

$$S(f) = \int_{-\infty}^{+\infty} s(t)e^{-j2\pi ft}\, dt$$

$$\begin{aligned}
S(f) &= \int_{-a}^{+a} Ae^{-j2\pi ft}\, dt \\
&= \frac{A}{-j2\pi f} e^{-j2\pi ft}\Big|_{t=-a}^{+a} \\
&= \frac{A}{-j2\pi f}\left[e^{-j2\pi fa} - e^{+j2\pi fa}\right] \\
&= \frac{A}{j2\pi f}\left[e^{+j2\pi fa} - e^{-j2\pi fa}\right] \\
&= \frac{A}{\pi f}[\sin(2\pi fa)]
\end{aligned}$$

giving:

$$S(f) = 2Aa\left[\frac{\sin(2\pi fa)}{2\pi fa}\right] \tag{5.5}$$

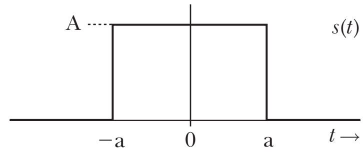



**FIGURE 5.6** This real and even rectangle $a(t)$ has width $= 2a$, height $= A$, and area $= 2Aa$. By inspection, its Fourier transform $S(f)$ should be even and real with the value $S(0) = 2Aa =$ the area under $s(t)$.

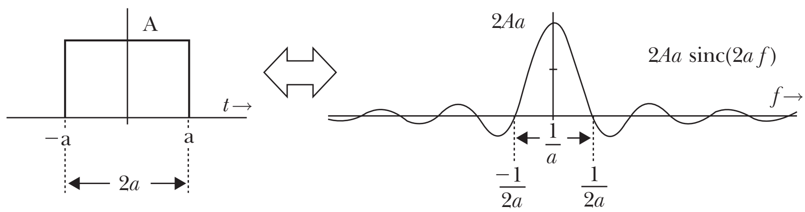



**FIGURE 5.7 A Fourier Transform Pair** The time and frequency domains could be switched without loss of generality.

**Alternate Solution** By recognizing that $s(t)$ is even, only the cosine part of the Fourier transform is required. Furthermore, this even integral equals twice the integral over $[0, +a]$. In this case, the answer is found directly.

$$\begin{aligned} S(f) &= 2\int_0^{+a} A\cos(2\pi ft)\, dt \\ &= 2A \left. \frac{\sin(2\pi ft)}{2\pi f} \right|_{t=0}^{a} \\ &= 2Aa\left[\frac{\sin(2\pi fa)}{2\pi fa}\right] \end{aligned}$$

This distinctive Fourier transform pair can be sketched as in Figure 5.7, with a double arrow linking the time-domain and frequency-domain sketches. Further study of this remarkable function continues in the next section.

## 5.4 The Sinc Function

The sinc (pronounced *sink*) function is defined as:

**DEFINITION 5.2 sinc($x$)**

*The function*

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

*is called the* sinc *function.*[3]

[3] A variation as $\text{sinc}(x) = \frac{\sin(x)}{x}$ may be encountered in mathematics and physics.

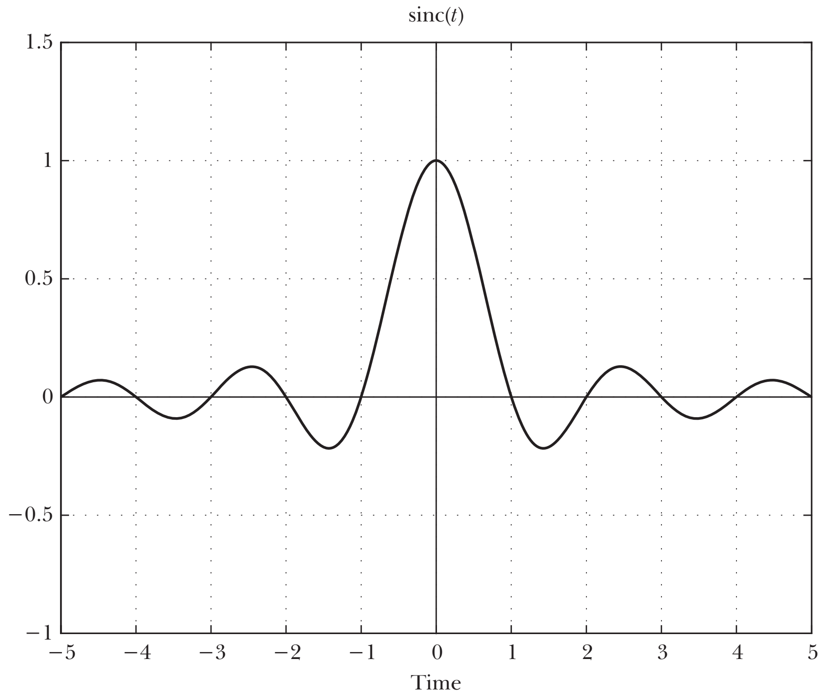



**FIGURE 5.8** The sinc function is frequently encountered in signals analysis as it is the Fourier transform of the unit rectangle. This even function equals 1 at the origin, and zero crossings occur at every integer value of time.

This function plays an important role in signals theory because it is the Fourier transform of a unit rectangle (i.e., $\text{rect}(t) \xleftrightarrow{\mathcal{F}} \text{sinc}(f)$). By inspection, the function $\text{sinc}(x)$ appears to be an odd $\sin(\pi x)$ with an exponentially decreasing amplitude; however, the following observations characterize this signal and make it somewhat different from either $\sin(\pi x)$ or $1/\pi x$. The sinc function is now studied with reference to Figure 5.8.

1. $\text{sinc}(x)$ is an even function
   **Proof:** The function $\text{sinc}(x)$ is the product of two odd functions: $\sin(\pi x) \times (1/\pi x)$.
2. $\text{sinc}(0) = 1$
   **Proof:** Direct substitution of $x = 0$ yields the nondescript result $0/0$. Examination of the limit of $\text{sinc}(x)$ as $x \to 0$, through application of L'Hopital's rule shows that:

$$\frac{\frac{d}{dx}\sin(\pi x)}{\frac{d}{dx}\pi x} = \frac{\pi\cos(\pi x)}{\pi} = 1 \text{ for } x = 0$$

These observations confirm the first two predictions concerning the Fourier transform of the rectangle. Unfortunately, the function $\text{sinc}(x)$ is not readily integrated to verify the third prediction (it is left as an exercise to confirm this result using numerical integration). On the other hand, the amplitude $(A = 1)$ of the transform domain unit rectangle evaluated at the origin can be used find the area under the sinc function, and to conclude that:

$$\int_{-\infty}^{+\infty} \text{sinc}(x)\, dx = 1$$

This result is especially interesting because the integral of sinc($x$) could not be computed directly. Instead, the desired integral could immediately be determined by inspection in the transform domain.

3. integral spacing between zero crossings
**Proof:** Zero crossings in the graph of the sinc function as shown in Figure 5.9 occur at regularly spaced intervals whenever the sine function in the numerator equals zero (except, of course, at the origin). In this case, $\text{sinc}(x) = 0$ for all integer $x \neq 0$.

**EXAMPLE 5.2 (MATLAB: Area Under sinc(*t*))**

*Use MATLAB to determine the integral of* $s(t) = sinc(t)$. *The result is expected to be unity.*

**Solution:**
The `trapz()` function computes an integral numerically by adding terms over an interval. The following interval spans $[-5000, 5000]$ with 10 points per unit time. The function `sinc()` is predefined in MATLAB.

```
t = -5000:0.1:5000;      % define the integral limits
trapz(t, sinc(t))        % compute the integral
ans =
  1.0000
```

This confirms the expected result. This integral cannot be solved directly by hand.

## 5.4.1 Expressing a Function in Terms of sinc(*t*)

Results are often encountered that resemble the sinc function, and it can be useful to make those equations fully equivalent. Creating a sinc function from a result of the form $\sin(ax)/bx$ requires that the argument of the sine function and the denominator be manipulated to match the form $\sin(\pi x)/\pi x$ including the $\pi$ at all times.

Consider the terms $A_n$ in the Fourier series approximation to a square wave with amplitude $A$ from Eqn. 5.1:

$$A_n = \frac{2A}{n\pi}\sin(2\pi n(a/T))$$

This result resembles sinc($x$), or:

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

and it can be expressed directly in terms of the sinc function.

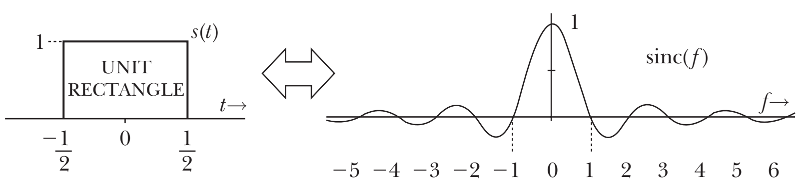



**FIGURE 5.9 A Fourier Transform Pair** $S(f) = \text{sinc}(f)$ is the Fourier transform of a unit rectangle $s(t) = \text{rect}(t)$.

Let $x = 2n(a/T)$ so that the above numerator can be expressed as $\sin(\pi x)$. The term $\pi x$ also appears in the denominator of $\text{sinc}(x)$, so multiply by $\frac{\pi x}{\pi x}$ to give:

$$\begin{aligned} A_n &= \frac{2A}{n\pi}\, 2\pi n(a/T) \left[\frac{\sin(2\pi n(a/T))}{2\pi n(a/T)}\right] \\ &= \frac{1}{T}\, 4Aa \,\text{sinc}\,(2n(a/T)) \end{aligned}$$

where it can be seen that the value of the components $A_n$ will vary as $1/T$. For the square wave case, with period $T = 1/f_0 = 4a$, the above equation reduces neatly to:

$$A_n = \frac{1}{4a}\, 4Aa \,\text{sinc}\,(2n(a/4a))$$

or

$$A_n = A \,\text{sinc}\left(\frac{n}{2}\right)$$

giving the Fourier series components of an even square wave, expressed in terms of the sinc function.

**Check:** Try $n = 1$, then $A_1 = A\pi/2$ as expected for the fundamental frequency component of a square wave with amplitude $A/2$, DC offset $= 1/2$.

As seen in Figure 5.4, in the Fourier series of any periodic rectangle the overall shape or *envelope* described by the frequency components is determined by the sinc function, while the period of the repeating rectangle determines the spacing between components.

## 5.4.2 The Fourier Transform of a General Rectangle

The result obtained in Eqn. 5.5 for the Fourier transform of a general rectangle can now be rewritten as a sinc function, and the two signals may be compared in detail as seen in Figure 5.10:

$$2Aa\left[\frac{\sin(2\pi af)}{2\pi af}\right] = 2Aa\,\text{sinc}(2af) \tag{5.6}$$

The sinc function may be readily sketched knowing its value at the origin and the first zero crossing.

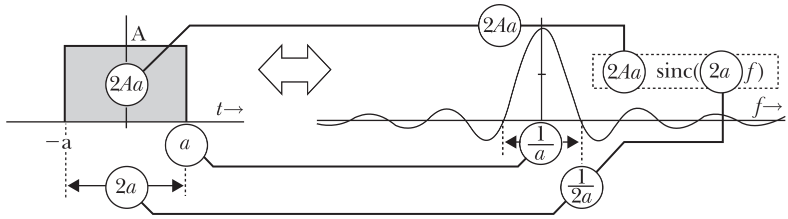



**FIGURE 5.10 A Fourier Transform Pair** A rectangle and the corresponding Fourier transform sinc function. This result can be written and sketched by inspection for any rectangle and sinc.

1. **Value at the Origin**
   - From Eqn. 5.6: Since $\text{sinc}(0) = 1$, then $2Aa\,\text{sinc}(0)$ gives the value $2Aa$ at the origin.
   - From the rectangle: The value of the sinc function at the origin equals the area under the transform rectangle.
2. **First Zero Crossing** (Other zero crossings lie at integer multiples of the first zero crossing.)
   - From Eqn. 5.6: The first zero crossing lies at $\text{sinc}(1)$, where $f = 1/2a$.
   - From the rectangle: The first zero crossing occurs at 1/(total width of the transform rectangle).

In summary, the sinc function for a given rectangle can be readily sketched without computation by observing the rectangle width and area. The corresponding sinc function has a first zero crossing at $1/W$ Hz, and its value at the origin equals the area of the rectangle.

Working backwards from a sketch of sinc to the defining equation, any sinc function with a peak amplitude B and a first zero crossing at $f = b$, has the equation: $S(f) = B\,\text{sinc}(f/b)$. This Fourier transform corresponds to a rectangle with Area $= B$ and Width $= 1/b$.

**EXAMPLE 5.3**

*Find the Fourier transform of the "stepped" function* s(t) *shown in Figure 5.11.*

**Solution:**

Without explicitly computing $S(f)$, some predictions about $S(f)$ can be made immediately. $S(f)$ should be a real and even signal, since $s(t)$ is real and even. The value $S(0)$ should equal the area under $s(t)$. The area of $S(f)$ should equal the value $s(0) = 2$.

The Fourier transform equation could be used to compute $S(f)$ directly, but this would require integration by parts, and the resulting formula for $S(f)$ would carry little meaning unless it was cleverly rearranged to resemble some familiar form. It can be more instructive to use the properties of the Fourier transform and the known behavior of some common component signals.

Another strategy is often useful when examining a new signal such as this. Consider this stepped function as the sum of two rectangles, then apply the linearity of the transform to find an exact result for $S(f)$. By inspection, the Fourier transform of the sum of two rectangles should take the form of the sum of two sinc functions. This result can be determined graphically, and then the exact numerical values can be readily filled in. This process is illustrated in Figure 5.12 where the result is:

$$S(f) = 4\,\text{sinc}(4f) + 2\,\text{sinc}(2f)$$

**Check:** $S(0) = 4 + 2 = 6 =$ area under $s(t)$, as expected.

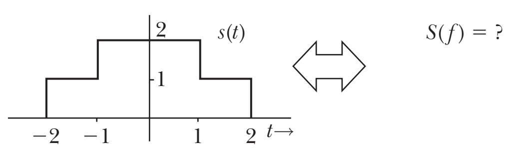



**FIGURE 5.11** What is the Fourier transform of this signal $s(t)$? By inspection, $S(f)$ should be real and even with area $= 2$ and $S(0) = 6$.

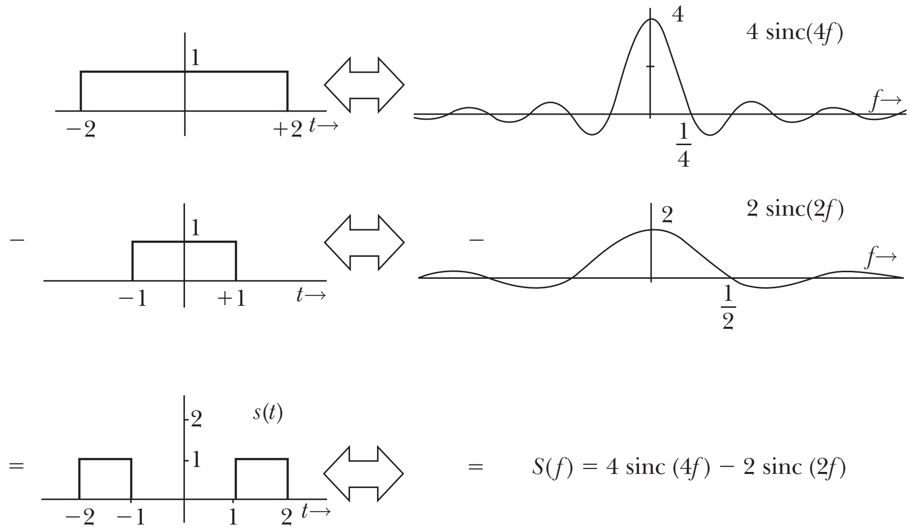



**FIGURE 5.12** The Fourier transform of $s(t)$ can be found graphically once it is recognized that $s(t)$ is the sum of two rectangles. In this figure, the signal components are manipulated in both the time and frequency domains. The result is obtained by inspection.

**EXAMPLE 5.4**

*Find the Fourier transform of the even pair of unit rectangles shown in Figure 5.13.*

**Solution:**

Observe that the Fourier transform $S(f)$ should be even and real, with Area = 0, and $S(0) = 2$. Note that the function $s(t)$ may be created from the difference between the same two rectangles as in Example 5.3. By inspection in Figure 5.14,

$$S(f) = 4\,\mathrm{sinc}(4f) - 2\,\mathrm{sinc}(2f)$$

**Check:** $S(0) = 4 - 2 = 2 =$ area under $s(t)$, as expected.

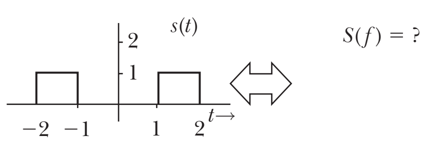



**FIGURE 5.13** What is the Fourier transform of $s(t)$? By inspection, $S(f)$ should be real and even with area = 0 and $S(0) = 2$.

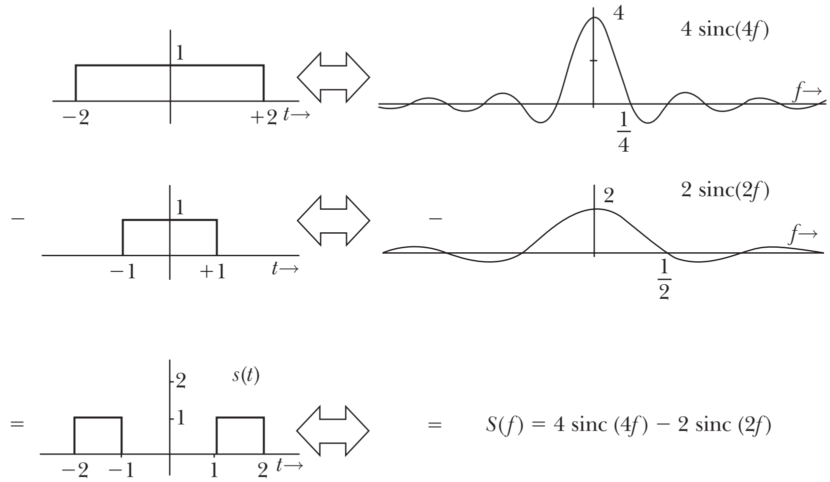



**FIGURE 5.14** The Fourier transform of $s(t)$ can be found graphically once it is recognized that $s(t)$ is the difference of two rectangles. In this figure, the signal components are manipulated in both the time and frequency domains. The result is obtained by inspection.

## 5.4.3 Magnitude of the Fourier Transform

The magnitude of any complex value is both real-valued and nonnegative. The magnitude of the sinc function will have the same value at the origin, and the zero crossings will be found in the same places. The outcome is illustrated in the magnitude of the Fourier transform $S(f)$ of the signal $s(t) = \mathrm{rect}(t/20)$, as shown in Figure 5.15.

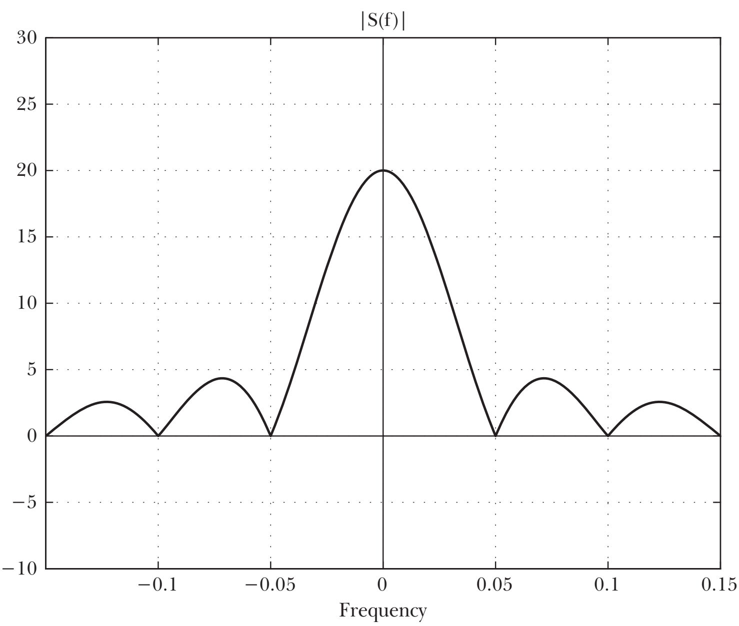



**FIGURE 5.15** The Fourier transform of a real function has a magnitude that is real, even, and nonnegative. The magnitude of the function $S(f) = 20\,\mathrm{sinc}(20f)$ shown here is even and real valued and nonnegative.

# 5.5 Signal Manipulations: Time and Frequency

The frequency- and time-domain forms are different means of representing *the same signal* expressed as either $s(t)$ or $S(f)$. Any manipulation of a signal in one domain has a direct effect on its representation in the other domain. The linkage between the two representations is the mathematics of the Fourier transform equation. By understanding this relationship, the Fourier transform can be applied to maximum benefit in both the time and frequency domains. In this section, the relationship between time-domain and frequency-domain signals is explored, and several useful properties of the Fourier transform emerge.

## 5.5.1 Amplitude Variations

It has already been established that varying the amplitude of the time-domain signal $s(t) \to ks(t)$ will simply change the frequency-domain amplitude by a corresponding fraction.

$$k\, s(t) \xleftrightarrow{\mathcal{F}} k\, S(f)$$

This relationship is assured by the linearity of the transform.

## 5.5.2 Stretch and Squeeze: The Sinc Function

Consider the time-domain rectangle in Figure 5.16, and suppose that the rectangle signal $s(t)$ having width $2a$ is expanded symmetrically along the horizontal axis by making $a$ larger. Observe that

- The width of the sinc function varies as $1/a$, so the sinc becomes narrower.
- Since making $a$ larger increases the area under the rectangle, then the value of the sinc function at the origin must become larger.
- Since the value at the origin $S(0)$ does not change, then the area under the sinc function remains constant.

All the while, of course, the characteristic sinc shape is maintained, since it is still the Fourier transform of a rectangle.

In summary, if a signal is stretched wider in one domain, then its Fourier transform will squeeze narrower and grow higher. The observation may be generalized to all signals.

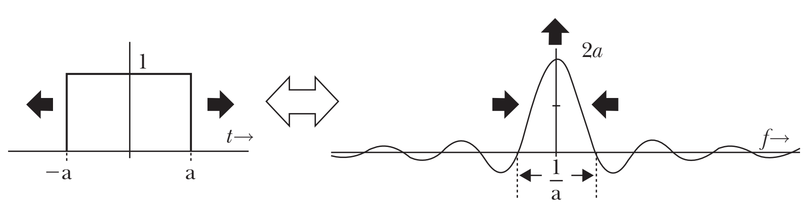



**FIGURE 5.16** As a rectangle widens, the corresponding sinc narrows and grows higher. The area under the sinc function, given by value of the rectangle at the origin, remains constant.

### 5.5.3 The Scaling Theorem

Consider a signal $s(t)$ having the Fourier transform $S(f)$. The signal $g(t) = s(kt)$ will be stretched along the time axis for $0 < k < 1$, and will be squeezed narrower for $k > 1$. The Fourier transform of the resulting signal $g(t)$ may be computed as:

$$S(f) = \int_{-\infty}^{+\infty} s(kt)e^{-j2\pi ft}\, dt$$

Let $x = kt$, then $t = x/k$, $dt = dx/k$, and the limits of integration remain unchanged.

$$S(f) = \frac{1}{k}\int_{-\infty}^{+\infty} s(x)e^{-j2\pi fx/k}\, dx$$

This integral may be recognized as the Fourier transform of $s(t)$ with $f$ replaced by $f/k$. In other words,

**THEOREM 5.2** *(Time Scaling)*

*If*

$$s(t) \stackrel{\mathcal{F}}{\longleftrightarrow} S(f)$$

*then for* k $\neq$ *0,*

$$s(kt) \stackrel{\mathcal{F}}{\longleftrightarrow} \frac{1}{|k|} S\left(\frac{f}{k}\right)$$

The absolute value $|k|$ in the denominator allows for negative $k$. This result is also known as the *similarity theorem* [6]. It is left as an exercise to prove the dual of this relationship for frequency-domain scaling:

**THEOREM 5.3** *(Frequency Scaling)*

*If*

$$s(t) \stackrel{\mathcal{F}}{\longleftrightarrow} S(f)$$

*then for* k $\neq$ *0,*

$$\frac{1}{|k|} s\left(\frac{t}{k}\right) \stackrel{\mathcal{F}}{\longleftrightarrow} S(kf)$$

After any computation such as this, it is useful to make some simple checks to ensure that the result is at least reasonable. This check will not prove the result, but it will detect obviously incorrect answers.

**Check:** Try $k = 1$ in the above theorems (no scaling). Then $s(t) = S(f)$ as expected.

### 5.5.4 Testing the Limits

Consider the unit height rectangle of Figure 5.16, which is made wider as in Section 5.5.2 where the corresponding sinc function narrows and becomes higher while retaining the same area. In the limit, as $a \to \infty$, the rectangle becomes very wide, and its appearance approaches a constant for all time. Meanwhile, the sinc function tends to an impulse with near-zero width, height approaching infinity, and unit area. By this reasoning, a new Fourier transform pair emerges that can be sketched in Figure 5.17.

$$s(t) = A \xleftrightarrow{\mathcal{F}} S(f) = A\delta(f)$$

The inverse Fourier transform of the impulse $S(f) = A\,\delta(f)$ will now be computed.

**Proof:** The sifting property of the impulse allows direct evaluation of the inverse Fourier transform:

$$s(t) = \int_{-\infty}^{+\infty} A\,\delta(f)e^{+j2\pi ft}\,df = Ae^0 = A$$

which gives the same result as predicted above.

The properties of this Fourier transform pair should be confirmed in light of the definition of the impulse. To appreciate these properties graphically, it is necessary to carefully interpret the meaning of the above sketch. Recall that the impulse is an arrow drawn with a height reflecting its area. The arrow shows that its amplitude tends to infinity.

The properties of the unit impulse $\delta(x)$ were introduced in Chapter 1:

1. $\delta(x) = 0$ for $x \neq 0$
2. $\int_{-\infty}^{+\infty} \delta(x)\,dx = 1$
3. $\delta(0) \to \infty$

1. The value at the origin of the time-domain constant equals A. This corresponds to the area of the impulse in the frequency domain. [Property 2 above.]
2. The area under the time-domain function tends to infinity. This corresponds to the definition of the value of the impulse at the origin. [Property 3 above.]

By the symmetry of a Fourier transform pair, it can be predicted that the Fourier transform of an impulse $s(t) = A\,\delta(t)$ is a constant $A$ in the frequency domain. Proof of this follows the same logic as the time-to-frequency case.

$$S(f) = \int_{-\infty}^{+\infty} A\,\delta(t)e^{-j2\pi ft}\,dt = Ae^0 = A$$

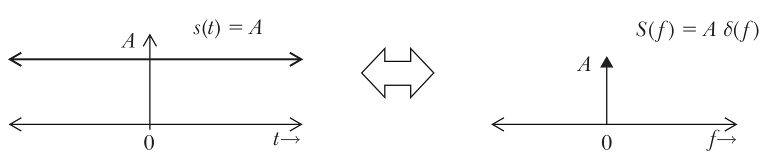



**FIGURE 5.17** The Fourier transform of a constant value $A$ is an impulse function having area $A$.

The time-domain impulse signal represents a very short duration signal. This might be the model for a spark or a lightning strike or a glitch in an electric circuit. The Fourier transform of this impulse can be interpreted as a signal with constant amplitude frequency components at all frequencies going to infinity. This result is intuitively correct since a lightning strike, or the sparks produced by the motor in an electric hair dryer, can produce interference in a radio signal across a wide range of frequencies.

Similarly, whenever a linear system response function is to be determined by applying an impulse input, the same operation can be described in the frequency domain as using an input test signal that simultaneously applies all possible input frequencies.

## 5.5.5 A Shift in Time

Consider the rectangle $s(t)$ with a small constant delay $t_0$ s, described by $g(t) = s(t - t_0)$, and sketched as shown in Figure 5.18. What is the Fourier transform $G(f)$?

Note that the area of the shifted rectangle $s(t - t_0)$, compared to $s(t)$ has not changed, so the value of the transform at $G(0)$ must remain constant for any time shift. It follows that $G(f)$ cannot shift along the frequency axis since its value at the origin $G(0)$ must remain equal to the area of the time-domain rectangle. Furthermore, the shifted rectangle is no longer an even function, so the new Fourier transform $G(f)$ must be complex, with real part even and imaginary part odd. Since the real and imaginary parts in $G(f)$ will each be symmetrical about the origin, the resulting $G(f)$ is definitely centered on the origin $f = 0$.

The effect of shifting a signal was observed in the Fourier series, when considering the case of a shifted sinusoid. Recall that only the phase of the frequency components was affected. In the complex Fourier series, use of the complex exponential describes phase directly. Also, when odd and even square wave signals were compared, the exact same frequency components were necessarily present in both waveforms, with each component undergoing a phase change proportional to its frequency.

The same logic applies to the Fourier transform of any shifted signal $s(t - t_0)$, as shown above. Because the overall shape or appearance of the rectangle is not changed by a simple time shift, the same frequency components are required to describe it. The Fourier transform $G(f)$ should still resemble the sinc function $S(f)$, although each component should undergo a phase change proportional to its frequency. An exact expression for $G(f)$ can now be found.

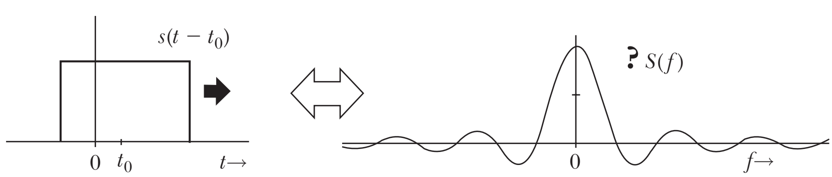



**FIGURE 5.18** Shifting the even time-domain rectangle $s(t)$ will yield the neither signal $g(t) = s(t - t_0)$. A time-domain shift does *not* cause a shift in the frequency domain. The effect on $S(f)$ is to be determined.

### 5.5.6 The Shifting Theorem

Consider a signal $s(t)$ having the Fourier transform $S(f)$. The signal $g(t) = s(t - t_0)$ will be shifted along the time axis. The corresponding Fourier transform of the time-shifted signal $g(t)$ may be computed as:

$$G(f) = \int_{-\infty}^{+\infty} g(t)e^{-j2\pi ft}\, dt = \int_{-\infty}^{+\infty} s(t - t_0)e^{-j2\pi ft}\, dt$$

Let $x = t - t_0$, then $t = x + t_0$, $dt = dx$, and the limits of integration remain unchanged.

$$G(f) = \int_{-\infty}^{+\infty} s(x)e^{-j2\pi f[x+t_0]}\, dx$$

$$G(f) = e^{-j2\pi ft_0} \int_{-\infty}^{+\infty} s(x)e^{-j2\pi fx}\, dx$$

This integral may be recognized as the Fourier transform of $s(t)$, multiplied by $e^{-j2\pi ft_0}$.

**THEOREM 5.4** *(Time Shifting)*

*If*

$$s(t) \stackrel{\mathcal{F}}{\longleftrightarrow} S(f)$$

*then*

$$s(t - t_0) \stackrel{\mathcal{F}}{\longleftrightarrow} e^{-j2\pi ft_0} S(f)$$

**Check:** Try $t_0 = 0$ (no time shift), then $s(t) = S(f)$, as expected.

For a shift in the opposite direction, the constant $t_0$ may be made negative. This result confirms the observations made above. It is left as an exercise to prove the inverse relationship for a constant frequency shift $f_0$.

**THEOREM 5.5** *(Frequency Shifting)*

*If*

$$s(t) \stackrel{\mathcal{F}}{\longleftrightarrow} S(f)$$

*then*

$$S(f - f_0) \stackrel{\mathcal{F}}{\longleftrightarrow} e^{+j2\pi f_0 t} s(t)$$

Note the difference in sign in the complex exponential between the above theorems. This reflects the difference between the Fourier transform and its inverse.

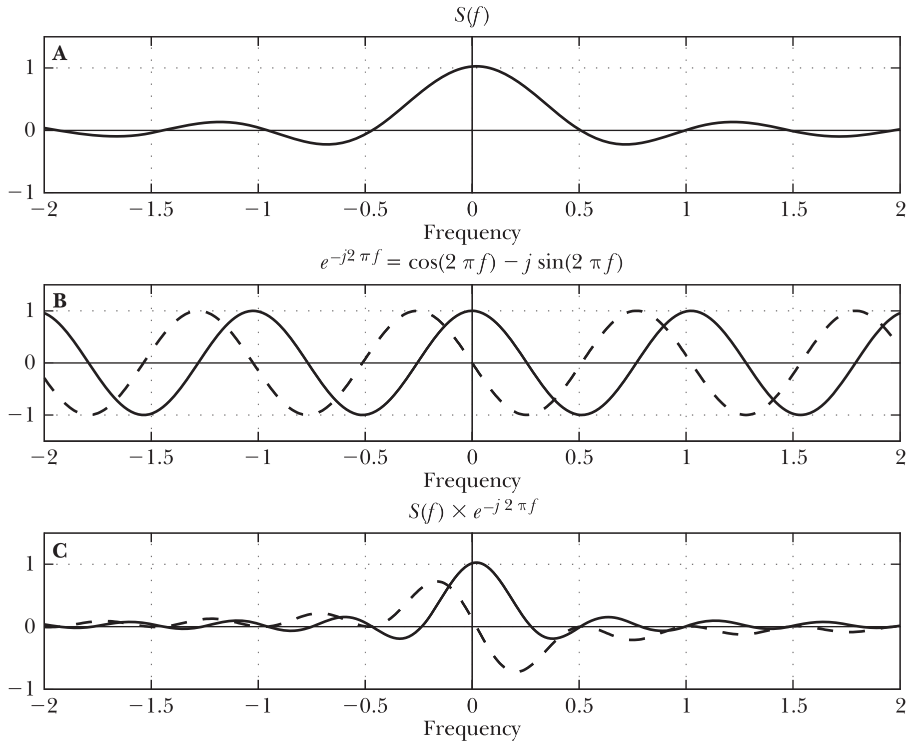



**FIGURE 5.19** The sinc function in A represents the Fourier transform of an even rectangle. The function in C is the Fourier transform of the same rectangle delayed by one second. When the real and even function (A) is multiplied by the complex exponential (B) the result has real part even and imaginary part odd (C). See also Figure 5.20.

It is worthwhile to explore what multiplication by a complex exponential really does to $S(f)$. In Figure 5.19A, the familiar sinc function Fourier transform of an even rectangle is shown. If the rectangle is shifted, this sinc function is multiplied by the complex exponential in Figure 5.19B, which gives the result in Figure 5.19C. Because the shifted rectangle has not changed in area, the value at the origin in both A and C is unchanged. Because the shifted rectangle is no longer an even function, the result in C has real part even and imaginary part odd.

## 5.5.7 The Fourier Transform of a Shifted Rectangle

From the above properties, the Fourier transform of the shifted rectangle $g(t) = s(t - k)$ is the original $S(f)$, multiplied by the complex exponential $G(f) = e^{-j2\pi fk}S(f)$. This real part even, imaginary part odd frequency-domain waveform has an unusual form, which, when sketched, is not especially revealing. The result is shown in Figure 5.20B for a small time shift $k$.

**Magnitude of $G(f)$** On the other hand, consider the magnitude of the complex exponential given by $|e^{-j2\pi fT}| = 1$. Consequently, the magnitude $|G(f)| = |S(f)|$ is

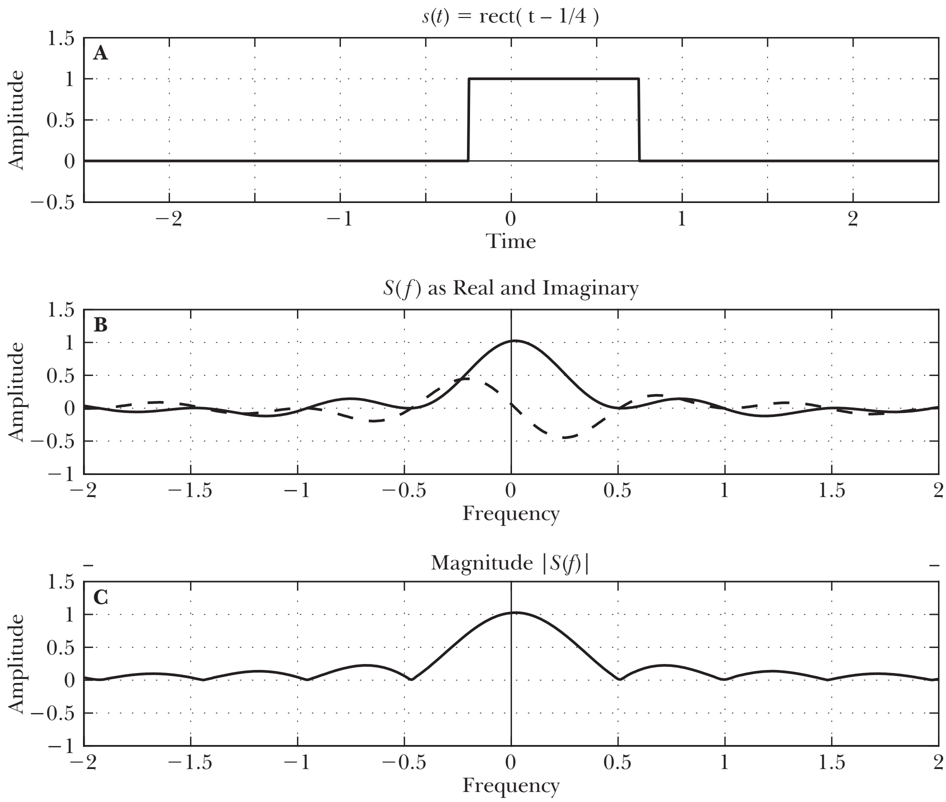



**FIGURE 5.20** The real signal $s(t)$, a shifted rectangle (A). Its Fourier transform $S(f)$ has real part even and imaginary part odd (B). The time-domain shift operation has no effect on the magnitude $|S(f)|$ of the Fourier transform (C).

unaffected by the shifting operation. The same frequency components are present in the same relative proportions in Figure 5.20B and C.

**Phase of $G(f)$** The (real, even) and (imaginary, odd) parts in the Fourier transform of the shifted $g(t)$ reflect a phase shift undergone by individual components. Recall the rotated phase vector described in the three-dimensional Fourier series graph. The phase rotation is function of frequency for a given constant time shift k, suggesting that components at higher frequency rotate faster than low frequency components as $k$ is varied. Furthermore, the exact phase change in each is given by:

$$\Phi(f) = \tan^{-1}\left[\frac{Im(e^{-j2\pi fT})}{Re(e^{-j2\pi fT})}\right] = \tan^{-1}\left[\frac{-\sin(2\pi fT)}{\cos(2\pi fT)}\right] = -2\pi fT$$

This result is the straight-line phase vs. frequency relationship that was shown in Figure 1.28 when a sinusoid was shifted by a constant time. In this case, the sinusoidal components upon which the Fourier transform is founded must each undergo the same time shift $T$, which requires a different phase shift in radians depending on the frequency of each component. This is the effect of multiplying $e^{-j2\pi fk}S(f)$ for the shifted signal $s(t-k)$. Fortunately, the magnitude of the components is often all that is important in signals analysis, and the magnitude of the

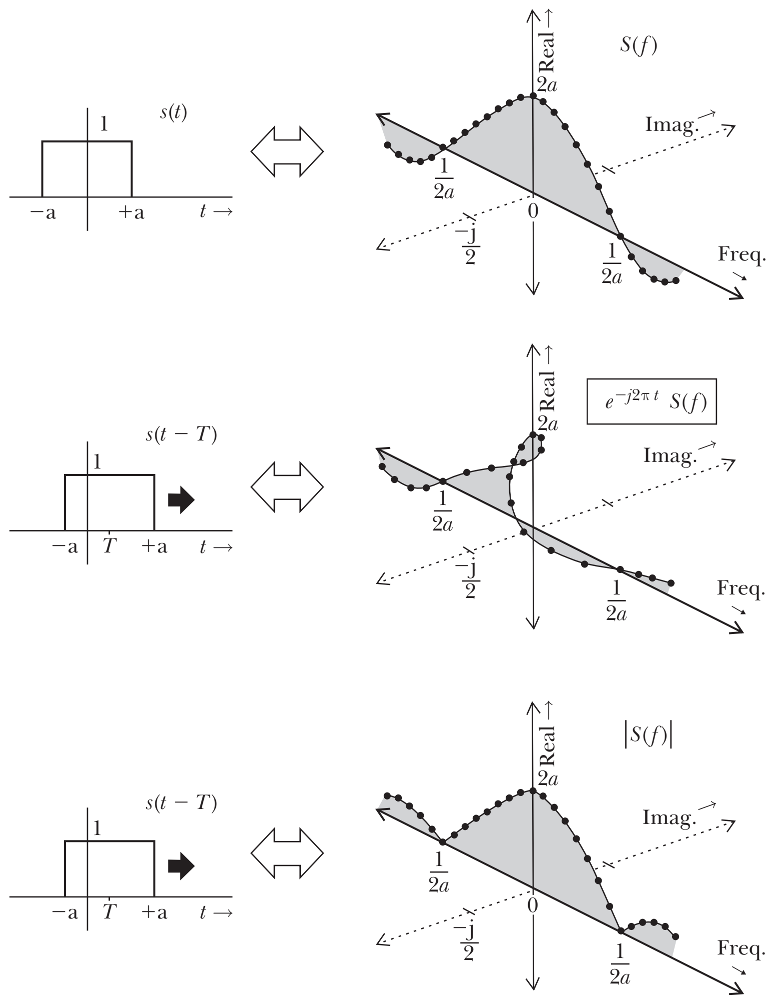



**FIGURE 5.21** A three-dimensional graph illustrates how real and imaginary components contribute to the Fourier transform of a shifted rectangle. The magnitude of the Fourier transform is unaffected by the time-domain shift.

Fourier transform $|S(f)|$ is unaffected by a time shift $s(t-k)$. Furthermore, in any practical application, delays are inevitable, so this result is significant.

The effect of shifting in time, multiplication by the complex exponential, and the meaning of the magnitude $|S(f)|$ can be seen more clearly if the above sketches are redrawn in three dimensions. The Figure 5.21A shows $s(t)$ and the corresponding real and even $S(f)$. In Figure 5.21B, $s(t-T)$ has the effect of twisting $S(f)$ through multiplication by the complex exponential. Each component undergoes a phase rotation depending on its frequency $\Phi(f) = -2\pi fT$. Finally, the even and real magnitude $|S(f)|$ is shown in Figure 5.21C where any delay $T$ is irrelevant.

## 5.5.8 Impulse Series—The Line Spectrum

The Fourier transform of a constant value $s(t) = A$ is an impulse $S(f) = A\delta(f)$ in the frequency domain. This corresponds to the DC offset, or "zero frequency" component of $s(t)$. Such a component was previously associated with the complex

Fourier series; an impulse now replaces the single value $C_0$. Shifting the single impulse reveals the meaning of each Fourier series component.

## 5.5.9 Shifted Impulse $\delta(f - f_0)$

If the impulse $A\delta(f - f_0)$ is shifted by a constant $f_0$, then by direct application of the frequency shift theorem:

$$A\delta(f - f_0) \xleftrightarrow{\mathcal{F}} Ae^{+j2\pi f_0 t}$$

## 5.5.10 Fourier Transform of a Periodic Signal

Any frequency-domain sketch of Figure 5.22 conveys the interpretation of impulses as components of $e^{-j2\pi ft}$. The above result precisely reflects the definition of a component $C_n$ at $nf_0$ in the complex Fourier series. Indeed, the study of the Fourier series can be directly applied to the Fourier transform by substituting, for each complex component $C_n$, an impulse response with area $C_n$, drawn with a height reflecting the value $C_n$. It follows that the Fourier transform of a cosine signal with period $T = 1/f_0$ will be a pair of half-area impulses at $f = +f_0$ and $f = -f_0$, as shown in Figure 5.23.

$$A\cos(2\pi f_0 t) \xleftrightarrow{\mathcal{F}} \frac{A}{2}\delta(f + f_0) + \frac{A}{2}\delta(f - f_0)$$

In general, the Fourier transform of a periodic signal will be characterized by impulses spaced at integer multiples $nf_0$ of a fundamental frequency $f_0$. Given the Fourier series of a periodic signal $s(t)$:

$$s(t) = \sum_{n=-\infty}^{+\infty} C_n e^{-j2\pi n f_0 t}$$

By inspection and from the above graph the corresponding Fourier transform is:

$$s(t) \xleftrightarrow{\mathcal{F}} \sum_{n=-\infty}^{+\infty} C_n \delta(f - nf_0)$$

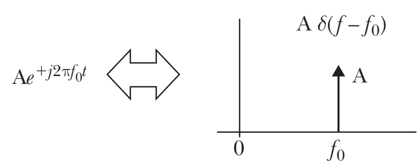



**FIGURE 5.22** A shifted impulse function corresponds to a single component of $e^{-j2\pi f}$, located at $f = f_0$.

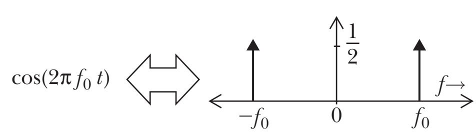



**FIGURE 5.23** The Fourier transform of a cosine signal with period $T = 1/f_0$ will be a pair of half-area impulses at $f = +f_0$ and $f = -f_0$.

**EXAMPLE 5.5**

*Sketch the Fourier transform of an even square wave with period* T $= 4a$ s, *and amplitude 1.*

**Solution:**

The Fourier series components of an even square wave have already been computed. In the Fourier transform, even components $C_n$ correspond to regularly spaced impulses, each with area $= C_n$.

Now, the Fourier transform of a constant $s(t) = A$ is an impulse with area $A$. On the other hand, the constant $s(t) = A$ is a power signal. Recall that the complex Fourier series representation of this same signal $s(t) = A$ is a single component $C_0$, or the zero-frequency component. It follows that the representation of power signals using the Fourier transform will involve impulses spaced at integer multiples $nf_0$, each with area corresponding directly to the complex Fourier series coefficients $\{C_n\}$. Consequently, the Fourier transform of a periodic signal takes on the appearance of a line spectrum, as seen in Figure 5.24.

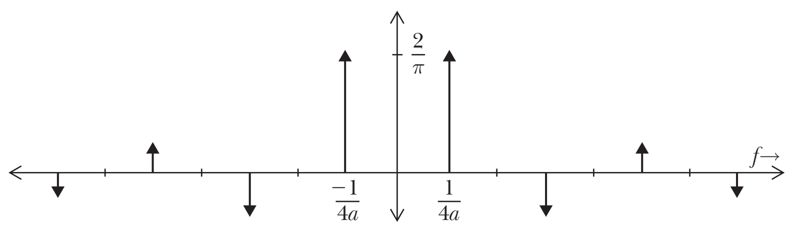



**FIGURE 5.24** The Fourier transform of an even square wave, period $T = 4a$, Amplitude $= 1$. The Fourier transform of any periodic signal is a *line spectrum*, consisting of impulses located at integer multiples of the period.

# 5.6 Fourier Transform Pairs

Figure 5.25 examines a series of *Fourier transform pairs* relating time- and frequency-domain signals. In each case, time-domain signals are shown on the left, with the corresponding Fourier transforms on the right. Note that the (frequency vs. time) columns could be switched with no loss of generality.

These four figures illustrate all the properties of the Fourier transform described in the previous discussion.

- Figure 5.25A shows a constant function $s(t) = 1$. Its transform is an impulse with unit area: $S(f) = \delta(f)$. Note the value of $s(0) = 1 =$ area of $\delta(f)$. Similarly, the value of $S(0) \rightarrow \infty$ equals the area under the function $s(t)$.
- Figure 5.25B shows the signal $s(t) = e^{+j2t}$, drawn in the time domain as real and imaginary parts. The corresponding Fourier transform shows the single component of $e^{+j2\pi ft}$ at frequency $f = 1/\pi$ Hz and unit area, as $S(f) = \delta(f - 1/\pi)$. The area of the unit impulse corresponds to the fact that $s(0) = 1$ in the time domain. Note the use of a broken line to indicate that $s(t) = j\sin(2t)$ is imaginary. This real-part-even, imaginary-part-odd signal has a real Fourier transform.

- Figure 5.25C shows an imaginary sinewave $s(t) = j\sin(2t)$ and its corresponding Fourier transform as two real half-unit components (one positive and one negative) equally spaced at $f = -1/\pi$ and $f = +1/\pi$ Hz. The total area in the frequency-domain equals zero, as given by the sum of the two impulses; also $s(0) = 0$. This odd and imaginary signal has an odd and real Fourier transform.
- Figure 5.25D shows a cosine waveform $s(t) = \cos(2t)$ and its corresponding Fourier transform as two half-unit components (both positive) equally spaced at $f = -1/\pi$ and $f = +1/\pi$ Hz. The total area in the frequency domain is 1, as given by the sum of the two impulses; also $s(0) = 1$. Similarly, the area under the cosine is zero (complete periods over all time) and $S(0) = 0$. This even and real signal has an even and real Fourier transform.
- The linearity of the Fourier transform can be confirmed by observing that signals $C + D = B$, in both the time and frequency domains.
- The frequency shifting property can be seen by observing that in the frequency domain, $B$ is a version of signal $A$, shifted by $-1/\pi$ Hz. The corresponding effect on signal $A$ in the time domain is multiplication by $e^{-j2t}$ to give signal $B$. Of course, the magnitude of the constant signal $A$ equals 1, for all time. This magnitude was not affected (in either domain) by the frequency shift, as the magnitude $|s(t)| = 1$ for $s(t) = e^{-j2t}$ independent of $t$.

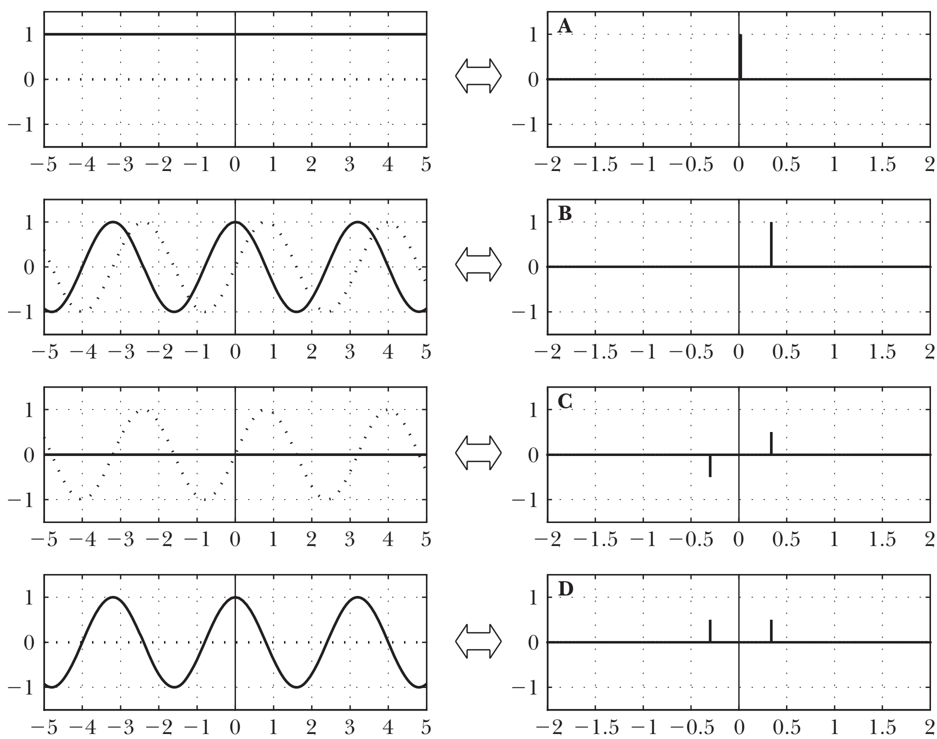



**FIGURE 5.25 Impulse Fourier transform pairs** Each pair shows a time-domain signal on the left and the corresponding frequency-domain signal on the right.

### 5.6.1 The Illustrated Fourier Transform

Appendix A presents a number of common signals and their Fourier transforms shown graphically along with the corresponding mathematical form of each. Where functions include real and imaginary parts, the imaginary components are drawn using a dotted line.

As the various properties of the Fourier transform are introduced, it is useful to refer to this appendix as confirmation of each formulation. For example, in every case the value of a signal at the origin in one domain must equal the area under the other. The Fourier transform properties relating odd and even functions to real and imaginary components may also be readily confirmed in the graphs. Different transforms may be combined graphically (as above) to extend the usefulness of this presentation, while noting that any manipulation in the time domain has a corresponding effect in the frequency domain and vice versa. Finally, the form of the inverse Fourier transform implies that in general the time and frequency columns for a given transform pair can be switched to give the inverse relationships.

Knowledge of a small number of basic Fourier transform pairs such as these, combined with a few simple manipulations (e.g., shift a signal, add two signals together, multiply by a constant), will allow rapid determination (or at least *estimation*) of the Fourier transform of new or unknown signals. Using the rules for odd and even functions, knowledge of the area under a function, and a few simple manipulations, a host of Fourier transform pairs will effectively be at hand.

## 5.7 Rapid Changes vs. High Frequencies

It is clear that the peak-to-peak transitions (maximum-to-minimum) in a cosine waveform take place more rapidly as frequency grows. Similarly, for any frequency spectrum, high-frequency components correspond to rapid changes in the waveform of the time-domain signal. It has been observed that the appearance of any signal that is squeezed in the time domain will stretch to higher frequencies when viewed in the frequency domain. Conversely, a slowly changing waveform will have a Fourier transform characterized by a concentration of low-frequency components, located near the origin or “zero-frequency” position in the frequency domain. In other words, the (energy or power) spectral density of the signal will be concentrated near $f = 0$.

One very rapid change is an infinite slope, a step function, or an on-off transition in a signal. Digital signals are characterized by such transitions, and, in operation, a digital circuit typically radiates frequencies across the spectrum. One need only try to listen to an AM radio while working near a unshielded digital circuit or computer to verify this fact. Modern computers require well-designed shielding to avoid radio interference. Significantly, such circuits are not primarily intended as or designed to be radio emitters!

Perhaps, the best example of a rapid change is a *glitch* or a spark. The near-instantaneous transitions produced by sparks on an electric motor, for example, will radiate across the frequency spectrum to be heard in nearby radios and televisions. Such a change can be modelled as an impulse. The Fourier transform of an impulse $\delta(t)$ is a constant value in the frequency domain, implying equal components of all frequencies, from DC to infinity.

Whenever an otherwise smoothly changing (low-frequency) signal is suddenly turned off, the rapid transition resulting from this step change will inevitably correspond to high-frequency spectral components. Some circuit designs purposely synchronize on-off activity to the zero-crossings in a periodic signal, where sharp voltage or current transitions at turn-on will be minimized.

Amateur radio operators who transmit Morse code signals are well aware of the requirement to smooth the harsh transitions when dots and dashes are transmitted. Any square edge, or a rapid on-off transition in a signal, will inevitably contain frequency components that extend beyond the bandwidth required for efficient communication and that may instead overlap frequencies occupied by other users of the radio spectrum.

## 5.7.1 Derivative Theorem

One means of determining how rapidly a signal changes is to study its derivative with respect to time. What happens in the frequency domain when the time derivative of a signal $s(t)$ is taken?

Before demonstrating the result, consider that the signal $s(t)$ can be expressed as a linear combination of orthogonal sine and cosine terms. Each of these terms can be differentiated separately, yielding:

$$\begin{aligned}&\frac{d}{dt}[A\cos(2\pi ft) + B\sin(2\pi ft)]\\ &\quad = 2\pi f[-A\sin(2\pi ft) + B\cos(2\pi ft)]\end{aligned}$$

The differentiation operator will effectively swap the even and odd components, and multiply each by a factor $2\pi f$. In the frequency domain, this will have the effect of swapping real and imaginary terms.

This question could be answered directly by starting with the definition of the inverse Fourier transform, and taking $d/dt$ of both sides of this equation:

$$\begin{aligned}\frac{d}{dt}s(t) &= \frac{d}{dt}\int_{-\infty}^{+\infty} S(f)e^{+j2\pi ft}\,df\\ &= \int_{-\infty}^{+\infty} [j2\pi f S(f)]e^{+j2\pi ft}\,df\end{aligned}$$

which may be recognized as the inverse Fourier transform of $[j2\pi f S(f)]$. In other words:

**THEOREM 5.6** *(Time Derivative)*

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*then*

$$\frac{d}{dt}s(t) \xleftrightarrow{\mathcal{F}} j2\pi f S(f)$$

**EXAMPLE 5.6** **(Time Derivative Theorem)**

*Use the time derivative theorem to find the Fourier transform of* $\sin(2\pi t)$ *by starting with the known Fourier transform of* $s(t) = \cos(2\pi t)$.

**Solution:**
Observe that $s(t)$ has a frequency of 1 Hz. Begin with the Fourier transform relationship:

$$\cos(2\pi t) \xleftrightarrow{\mathcal{F}} \frac{1}{2}\delta(f-1) + \frac{1}{2}\delta(f+1)$$

describing an even pair of real half-area impulses in frequency, and apply the time derivative theorem to obtain:

$$\frac{d}{dt}\cos(2\pi t) = -2\pi \sin(2\pi t) \xleftrightarrow{\mathcal{F}} j2\pi f\left[\frac{1}{2}\delta(f-1) + \frac{1}{2}\delta(f+1)\right]$$

which simplifies to give:

$$\sin(2\pi t) \xleftrightarrow{\mathcal{F}} \left[\frac{-j}{2}\delta(f-1) + \frac{+j}{2}\delta(f+1)\right]$$

which is non-zero only when $f = \pm 1$ and describes the expected odd pair of imaginary half-area impulses in frequency.

The corresponding derivative theorem for the inverse Fourier transform differs in a sign, as:

**THEOREM 5.7** *(Frequency Derivative)*

If

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

then

$$-j2\pi t\, s(t) \xleftrightarrow{\mathcal{F}} \frac{d}{df}S(f)$$

## 5.7.2 Integration Theorem

The inverse of the Fourier transform properties of differentiation are those related to integration. However, the integration operation may change the nature of a signal in important ways. For example, let the signal $a(t)$ be a unit rectangle and let $b(t)$ be its time integral as shown in Figure 5.26. The rectangle is bounded in time and has finite energy; however, its integral has infinite energy. The rectangle has no DC component (over all time); however, its integral has DC component equal to $1/2$ (over all time). The signal $b(t)$ is an odd signal except for this DC component, and, by inspection, its Fourier transform should be imaginary and odd, plus a delta function.

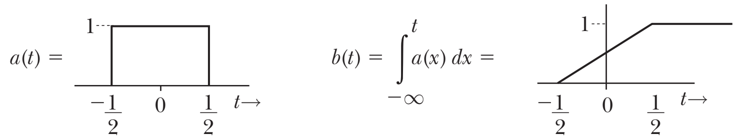



**FIGURE 5.26** The time integral of $a(t) = \text{rect}(t)$.

The appearance of the delta function is characteristic of the Fourier transform of an integral and leads to:

**THEOREM 5.8** *(Time Integration)*

*If*

$$s(t) \overset{\mathcal{F}}{\longleftrightarrow} S(f)$$

*then*

$$\int_{-\infty}^{t} s(x)\, dx \overset{\mathcal{F}}{\longleftrightarrow} \frac{1}{j2\pi f}\, S(f) + \frac{1}{2}\, S(0)\, \delta(f)$$

## 5.8 Conclusions

The Fourier transform has been developed as a generalization of the complex Fourier series, suitable for both periodic and nonperiodic signals. In general, nonperiodic signals will have a continuous Fourier transform, while the Fourier transform of a periodic signal is characterized by a line spectrum with components spaced at integer multiples of the period. Furthermore, the overall envelope of the spectral components follows a shape distinguished by the continuous Fourier transform of a signal period of the signal. Indeed, when computing Fourier series components of a periodic signal, only a single period is considered.

The properties of the continuous Fourier transform have been explored in detail, revealing its utility in the study of both periodic and nonperiodic signals. By understanding the properties of the Fourier transform, reflected in both the time domain and the frequency domain, and by knowing a few basic transform relationships, the properties of many new and interesting signals can be appreciated without explicitly computing the Fourier transform. These properties can now be put to use in the study of various Electrical Engineering applications. In Chapter 8, the class of signals resulting when input signals are sampled (perhaps in a digital computer) will be explored, as will computation of the Fourier transform and its inverse for these discrete signals. This will lay the groundwork for the study of digital signal processing (DSP), which is fast becoming the technique of choice in many signal-processing applications. In Chapter 6, the Fourier transform is put to use in the study of practical problems in the context of linear systems analysis.

# End-of-Chapter Exercises

**5.1** Use the definition of the inverse Fourier transform to show that for a signal $s(t)$, the area under the frequency domain equals the value at $s(0)$. In other words, show that:

$$s(0) = \int_{-\infty}^{+\infty} S(f)\, df$$

**5.2** The signal $s(t)$ has the Fourier transform $S(f)$. Describe the Fourier transform of $a(t) = 1 + s(t)$.

**5.3** Consider the signal $a(t)$, which is a 100 Hz cosine. A new signal $b(t)$ is the signal $a(t)$ shifted right by 1 rad.

**(a)** Sketch $b(t)$ in the time domain.
**(b)** Determine the phase shift in seconds.
**(c)** Write an expression for $B(f)$ in terms of $A(f)$.
**(d)** What is the area under $B(f)$?
**(e)** In $S(f)$, what is the area of the real part of the impulse found at 100 Hz?

**5.4** Consider the signal $s(t) = a(t) + b(t)$ where $a(t)$ and $b(t)$ are both even functions. Is the Fourier transform $S(f)$ odd or even or neither?

**5.5** Consider the signal $s(t) = \text{rect}(2(t-5))$ and its Fourier transform $S(f)$.

**(a)** Sketch $s(t)$ in the time domain.
**(b)** Sketch the magnitude $|S(f)|$.
**(c)** What should be the value of $S(f)$ for $f = 0$?
**(d)** What should be the area under $S(f)$?

**5.6** Consider the signal $s(t) = u(t-10)$ and its Fourier transform $S(f)$, where $u(t)$ is the unit step function.

**(a)** Sketch $s(t)$ in the time domain.
**(b)** What should be the value of $S(f)$ for $f = 0$?
**(c)** What should be the area under $S(f)$?
**(d)** Describe $S(f)$ as odd or even, real or imaginary.

**5.7** Consider the signal $s(t) = \text{rect}(2(t-5))$ and its Fourier transform $S(f)$.

**(a)** Determine $S(f)$ using the time-shifting theorem and $a(t) = \text{rect}(2t)$.
**(b)** Evaluate the Fourier transform integral to find $S(f)$ and $|S(f)|$.
**(c)** From $S(f)$, what is the area under $s(t)$?

**5.8** Consider the signal $s(t) = \cos(250\pi(t + 0.001))$ and its Fourier transform $S(f)$.

**(a)** What phase shift (rad) in $s(t)$ corresponds to the time shift by 1 msec?
**(b)** Sketch $s(t)$ in the time domain.
**(c)** What should be the value of $S(f)$ for $f = 0$?
**(d)** What should be the area under $S(f)$?
**(e)** Sketch $S(f)$.
**(f)** Sketch the magnitude $|S(f)|$.

**5.9** Consider the signal $s(t) = \cos(250\pi(t + 0.001))$ and its Fourier transform $S(f)$.

**(a)** Determine $S(f)$ using the time-shifting theorem and $a(t) = \cos(250\pi t)$.
**(b)** Evaluate the Fourier transform integral to find $S(f)$ and $|S(f)|$.
**(c)** From $S(f)$, what is the area under $s(t)$?

**5.10** Consider the signal $s(t) = e^{-a|t|}$ for $a > 0$ and its Fourier transform $S(f)$.

**(a)** Sketch $s(t)$ in the time domain.
**(b)** What should be the value of $S(f)$ for $f = 0$?
**(c)** What should be the area under $S(f)$?
**(d)** Describe $S(f)$ as odd or even, real or imaginary.
**(e)** Sketch the magnitude of the Fourier transform $|S(f)|$.

**5.11** Consider the signal $s(t) = e^{-a|t|}$ for $a > 0$ and its Fourier transform $S(f)$.

**(a)** Evaluate the Fourier transform integral to find $S(f)$ and $|S(f)|$.
**(b)** From $S(f)$, what is the area under $s(t)$?

**5.12** Starting with $a(t) = \cos(20\pi t)$ and its Fourier transform $A(f)$, use the time derivative theorem to find the Fourier transform of $b(t) = \sin(20\pi t)$.

**5.13** Consider the signal $s(t) = e^{-|t|}$ and its Fourier transform $S(f)$.

**(a)** Sketch $s(t)$ in the time domain.
**(b)** From $s(t)$, determine the value of $S(0)$.
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** From $s(t)$, determine the area under $S(f)$.
**(e)** Calculate the Fourier transform of $s(t)$.

**5.14** Consider the *sign function* defined as:

$$s(t) = \begin{cases} t < 0 & -1 \\ t = 0 & 0 \\ t > 0 & +1 \end{cases}$$

**(a)** Sketch $s(t)$ in the time domain.
**(b)** From $s(t)$, determine the value of $S(0)$.
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** From $s(t)$, determine the area under $S(f)$.
**(e)** Calculate the Fourier transform of $s(t)$.
**(f)** Sketch $S(f)$.

**5.15** Consider the signal $s(t) = u(t)$ where $u(t)$ is the unit step function.

**(a)** Express $s(t)$ in terms of the sign function in Question 5.14.
**(b)** From $s(t)$, determine the value of $S(0)$.
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** From $s(t)$, determine the area under $S(f)$.
**(e)** Calculate the Fourier transform of $s(t)$.
**(f)** Sketch $S(f)$.

**5.16** The signals $a(t) = \cos(2\pi t)$ and $b(t) = \cos(20\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

**(a)** What should be the value of $S(f)$ for $f = 0$?
**(b)** What should be the area under $S(f)$?
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** Sketch the Fourier transform $S(f)$.

**5.17** The signals $a(t) = \sin(2\pi t)$ and $b(t) = \cos(20\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

**(a)** What should be the value of $S(f)$ for $f = 0$?
**(b)** What should be the area under $S(f)$?
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** Sketch the Fourier transform $S(f)$.

**5.18** The signals $a(t) = \text{rect}(t)$ and $b(t) = \cos(200\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

**(a)** What should be the value of $S(f)$ for $f = 0$?
**(b)** What should be the area under $S(f)$?
**(c)** Describe $S(f)$ as odd or even, real or imaginary.
**(d)** Sketch the Fourier transform $S(f)$.

**5.19** Consider the signal $s(t)$ from Figure 5.27.

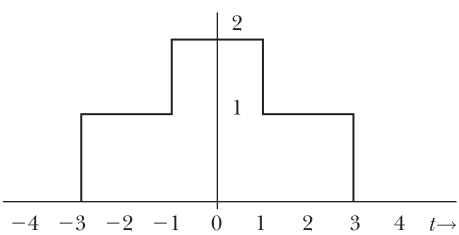



**FIGURE 5.27** Diagram for Questions 5.19 to 5.22.

**(a)** By inspection, what is the DC component in the Fourier transform?
**(b)** What is the area under the Fourier transform?
**(c)** Express $s(t)$ as the sum of two simple signals.
**(d)** By inspection, give an exact expression for the Fourier transform $S(f)$.

**5.20** Calculate the Fourier transform of the signal $s(t)$ from Figure 5.27.

**5.21** Consider the signal $s(t)$ from Figure 5.27 and the signal $c(t) = s(t - 2)$.

**(a)** By inspection, give an exact expression for the Fourier transform $S(f)$.
**(b)** Sketch $c(t)$.
**(c)** By inspection, sketch $C(f)$ as magnitude vs. frequency.
**(d)** Describe a sketch of $C(f)$ and as phase vs. frequency.
**(e)** What is the total energy contained in $C(f)$?

**5.22** Calculate the Fourier transform of the signal $c(t) = s(t - 2)$ where $s(t)$ is from Figure 5.27.

**5.23** Consider the unit rectangle $s(t)$ and its time derivative $b(t) = s'(t)$ in Figure 5.28.

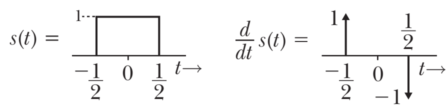



**FIGURE 5.28** Figure for Question 5.23.

**(a)** By inspection, determine an expression for the Fourier transform $B(f)$.
**(b)** Determine the Fourier transform of the derivative by using $S(f)$ and the time derivative theorem.
**(c)** Calculate $B(f)$ and check this result against the predictions from part (a).
**(d)** Compare the two answers from (a) and (b) above.

**5.24** Write a simplified expression for the inverse Fourier transform $s(t)$ of the signal $S(f)$ below.

$$S(f) = 2 - \frac{1}{5}\delta(f-5) + \frac{1}{3}\delta(f-3) - \frac{1}{5}\delta(f+5) + \frac{1}{3}\delta(f+3)$$

**5.25** Consider the signal $s(t)$, where $k$ is a constant.

$$s(t) = \frac{e^{+j2\pi(1-k)t} + e^{-j2\pi(1+k)t}}{2}$$

**(a)** Simplify $s(t)$ as the product of two functions of time.
**(b)** Describe the Fourier transform of $s(t)$.
**(c)** Sketch $S(f)$ for $k = 0.5$.
**(d)** Compute the Fourier transform $S(f)$.

# CHAPTER 6

# Practical Fourier Transforms

LEARNING OBJECTIVES

By the end of this chapter, the reader will be able to:

- Explain how convolution and multiplication relate through the Fourier transform
- Interpret the significance of impulse response vs. transfer function
- Describe causality and its meaning for linear systems analysis
- Compute the phase and frequency response of common systems
- Identify graphically various frequency-domain filtering operations
- Use the Fourier transform to accomplish frequency-domain filtering
- Derive a time-domain filter from a frequency description
- Explain the role of the modulation theorem in signals and systems
- Create a Bode plot from a time-domain impulse response
- Apply the Fourier transform to periodic signals graphically from first principles

## 6.1 Introduction

In this chapter, the techniques of signals analysis will be explored in the context of practical applications. Further properties of the Fourier transform will be developed, in particular the convolution property. The ideal linear system introduced in Chapter 1 will serve as the model for these applications. A thorough analysis of the behavior of signals and systems depends on an appreciation of both the time and frequency domains. The Fourier transform will be further treated in both domains, and will be found to be appropriate for both periodic and nonperiodic signals.

## 6.2 Convolution: Time and Frequency

Although the role of convolution is central to the analysis of linear systems, it has been shown that evaluating the convolution integral can be a computationally demanding exercise. Even when evaluated by computer, as many integrations are necessary as there are points in the result. Fortunately, this situation is greatly simplified when approached from the perspective of the frequency domain. In practice, convolution can often be avoided completely by working in the transform domain. Despite the fact that the Fourier transform itself is not always easy to compute, this is often the preferred method to avoid performing the tedious convolution operation. To this end, it will be shown that time-domain convolution corresponds to simple multiplication in the frequency domain, or:

$$s(t) * h(t) \Leftrightarrow S(f) \times H(f)$$

Consequently, in a linear system a simple multiplication relates the input and output signals in the frequency domain.

**The Logarithm Domain** Transformation to another domain is a commonly employed technique for simplifying computations. Seen from the perspective of a different domain, complicated (or impossible) operations can often be reduced to simple arithmetic. Consider the multiplication problem $Z = X \times Y$. By transforming each value to the *logarithm domain*, the problem reduces to an addition, as $\log(Z) = \log(X) + \log(Y)$. This simplification comes at the price of having to compute the logarithm of $X$ and $Y$, and then performing the inverse logarithm operation to find $Z$. Before the days of electronic pocket calculators, long multiplication was routinely simplified in this way using printed tables of logarithms and (inverse) anti-logarithms to supply the transformations. The slide rule, also made obsolete by pocket calculators, performed multiplication based on sliding logarithmic scales to add in the *logarithm domain*. Even today, engineering units such as decibels (dB) are employed to take advantage of the opportunity to add values expressed in the logarithm domain.

## 6.2.1 Simplifying the Convolution Integral

The defining relationship between time-domain convolution and frequency-domain multiplication is:

**THEOREM 6.1** *(Convolution)*

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*and*

$$h(t) \xleftrightarrow{\mathcal{F}} H(f)$$

*then*

$$s(t) * h(t) \Leftrightarrow S(f) \times H(f)$$

where $*$ is the convolution operator.

**Proof:**
Consider the convolution $g(t) = s(t) * h(t)$. Find $G(f)$ in terms of $S(f)$ and $H(f)$.
Begin with the definition of the Fourier transform of $g(t)$:

$$G(f) = \int_{-\infty}^{+\infty} [s(t) * h(t)] e^{-j2\pi ft}\, dt$$

where the integral defining the convolution $s(t) * h(t)$ may be substituted:

$$= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} s(t - x) h(x)\, dx \right] e^{-j2\pi ft}\, dt$$

This equation may be rearranged to give:

$$= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} s(t-x) e^{-j2\pi ft} dt \right] h(x) dx$$

After applying the time-shift property to the inner integral:

$$= \int_{-\infty}^{+\infty} [e^{-j2\pi fx} S(f)] h(x) dx$$

$$= S(f) \int_{-\infty}^{+\infty} h(x) e^{-j2\pi fx} dx$$

The remaining integral describes the Fourier transform $H(f)$.

$$G(f) = S(f)H(f)$$

Leaving the desired relationship:

$$s(t) * h(t) \Leftrightarrow S(f) \times H(f) \tag{6.1}$$

To find the convolution $g(t) = s(t) * h(t)$ using this result, apply the Fourier transform to $s(t)$ and $h(t)$ to obtain $S(f)$ and $H(f)$. Multiplication then gives $G(f) = S(f)H(f)$. Finally, compute the inverse Fourier transform of $G(f)$ to arrive at $g(t)$. The Fourier transform would be a valuable tool, even if this property of simplifying convolution was its only use.

The converse relationship may also be defined as:

**THEOREM 6.2** *(Multiplication)*

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*and*

$$h(t) \xleftrightarrow{\mathcal{F}} H(f)$$

*then*

$$s(t) \times h(t) \Leftrightarrow S(f) * H(f)$$

where $*$ is the convolution operator.

It is concluded that time-domain multiplication corresponds to convolution in the frequency domain and vice versa. This very useful property may be applied effectively in situations wherever computation might be easier in the transform domain. By recognizing such situations, a great deal of computational effort can often be saved. Example 6.1 illustrates this application of the convolution property.

## EXAMPLE 6.1 (Fourier Transform of a Triangle)

*Find the Fourier transform of the triangle shown in Figure 6.1.*

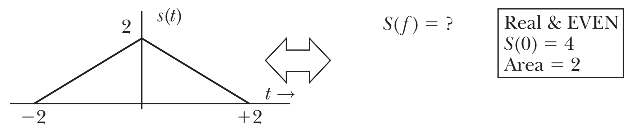



**FIGURE 6.1** The unknown Fourier transform of this triangle must be even and real with area = 2 and value at the origin = 4. Note that this triangle can be formed by the convolution of a rectangle with itself.

**Solution:**

By inspection, the Fourier transform $S(f)$ should be real and even, with Area = 2, and $S(0) = 4$. (The triangle has Area = 4 and $s(0) = 2$.) Another observation can now be made, which will lead directly to the desired solution. Observe that the triangle $s(t)$ can be obtained from the convolution of a suitable rectangle with itself. Knowing that the Fourier transform of a rectangle is a sinc function, and using the convolution property, the Fourier transform $S(f)$ should be given by the multiplication of the corresponding sinc function with itself. In other words, the Fourier transform of a triangle should resemble $\text{sinc}^2(f)$. This insightful observation leads directly to an answer that otherwise would not be self-evident, requiring explicit computation of the Fourier transform integral.

An exact answer for $S(f)$ may be derived directly by sketching the time- and frequency-domain results as in Figure 6.2. Recall that when two functions are convolved, the resulting width is the sum of the width of each function, while the resulting area is the product of the two areas. Consequently, since the triangle has Area = 4, Width = 4, the two rectangles must each have Area = $\sqrt{4} = 2$, and Width = $4/2 = 2$. Each rectangle is therefore Height = 1, and Width = 2. The Fourier transform for this rectangle is shown. Finally, since convolution in one domain is equivalent to multiplication in the other, the desired result is given by: $S(f) = 4\,\text{sinc}^2(2f)$.

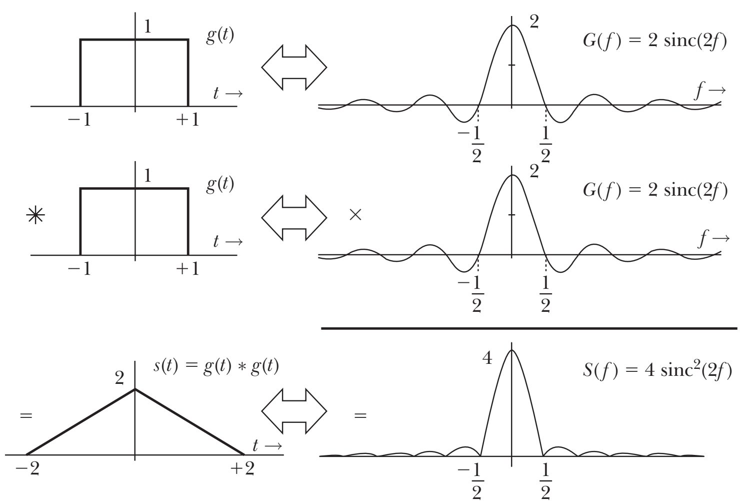



**FIGURE 6.2** The Fourier transform of the triangle can be found directly once the signal is recognized as being the convolution of a rectangle with itself. In the frequency domain, the corresponding operation is a sinc function multiplied by itself. See Figure 6.1.

**EXAMPLE 6.2 (Fourier Transform of Two Pulses)**

*Use the convolution property to find the Fourier transform of the even pair of unit rectangles shown in Figure 6.3.*

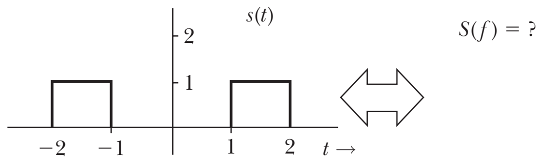



**FIGURE 6.3** The unknown Fourier transform of this pair of rectangles must be even and real with area = 0 and value at the origin = 2. Note that the rectangles can be formed by the convolution of a single rectangle with a pair of impulses.

**Solution:**

Observe that the Fourier transform $S(f)$ should be real and even, with Area = 0, and $S(0) = 2$.

The function $s(t)$ may created from the convolution of a single unit rectangle with a pair of unit impulses, one located at $t = +3/2$ and the other at $t = -3/2$. A copy of the rectangle will be obtained corresponding to each impulse. After sketching this operation in the time and frequency domains, the desired result is found as shown in Figure 6.4.

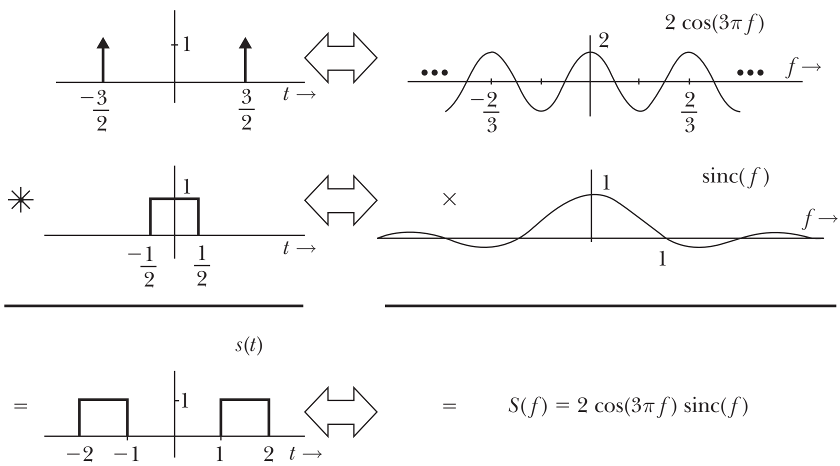



**FIGURE 6.4** The Fourier transform of the two rectangles can be found directly once the signal is recognized as being the convolution of a rectangle with a pair of impulses. In the frequency domain, the corresponding operation is a sinc function multiplied by cosine. Compare to Figure 5.14.

The inverse Fourier transform of an even pair of impulses yields a cosine as a function of frequency. Convolution in one domain is equivalent to multiplication in the other, leaving the desired result: $S(f) = 2\cos(3\pi f)\,\text{sinc}(f)$.

## 6.3 Transfer Function of a Linear System

It has been shown that if $h(t)$ is the response function of a linear system, any input signal is convolved with $h(t)$ to produce the output signal $g(t) = s(t) * h(t)$. It is now established that in the frequency domain, the same operation corresponds to a

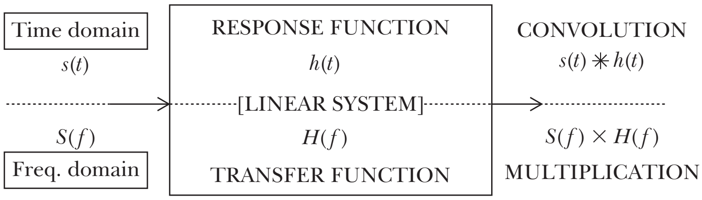



**FIGURE 6.5** A linear system can be categorized in either the time domain or the frequency domain. The time-domain impulse response $h(t)$ is convolved with an input signal. In the frequency domain, input signals are multiplied by $H(f)$.

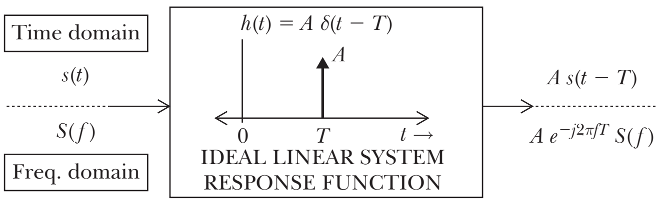



**FIGURE 6.6** An ideal linear system can be categorized in either the time domain or the frequency domain. Computations are greatly simplified in the frequency domain as convolution is avoided in favor of a simple multiplication.

multiplication: $G(f) = S(f)H(f)$. The Fourier transform $H(f)$ of the response function $h(t)$, is identified as the *transfer function* of the system. Either the (impulse) response function or the transfer function can be used to uniquely characterize a specific linear system as seen in Figure 6.5.

Many systems are commonly described by their frequency characteristics rather than their time-domain behavior, using the transfer function $H(f)$ rather than the impulse response. For example, an ideal *lowpass filter* is a circuit that blocks all frequencies above a certain frequency while lower frequencies pass unaffected. It would be no less correct, but much less descriptive, if this same system was described as one that *convolves input signals with a sinc function.* This specific example is to be studied in Section 6.7.

## 6.3.1 Impulse Response: The Frequency Domain

The function $h(t)$ is also known as the impulse response of the system. When a unit impulse $\delta(t)$ is input to a linear system, the response function $h(t)$ emerges. Now the Fourier transform of an impulse is a constant equal to one, since the area under $\delta(t)$ equals one. Using the time-shift property, the Fourier transform of a shifted impulse, multiplied by a constant should be:

$$A\,\delta(t-T) \Leftrightarrow A\,e^{-j2\pi fT}$$

If any signal $s(t)$ passes through this system, it is expected to emerge undistorted. In the time domain, the input signal $s(t)$ is convolved with the impulse response $h(t)$, to produce a shifted and amplified version of $s(t)$. The corresponding effect in the frequency domain is to multiply $S(f)$ by the Fourier transform of the response function as shown in Figure 6.6.

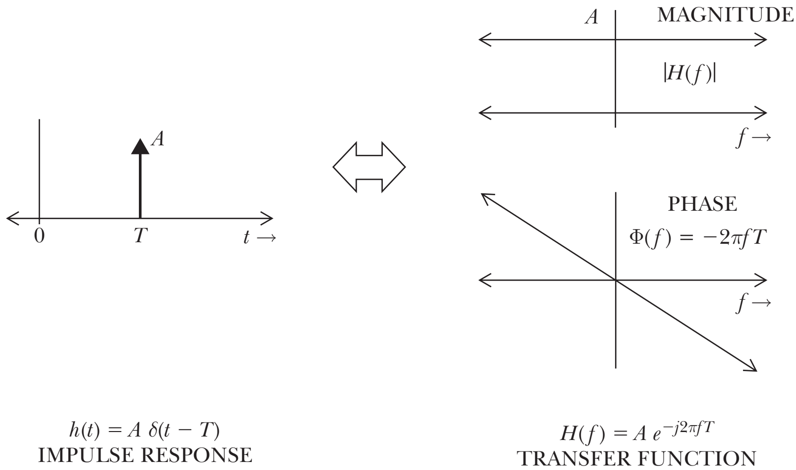



**FIGURE 6.7** A ideal linear system can be categorized in either the time domain or the frequency domain. The distinctive ideal impulse response $h(t)$ corresponds to the frequency-domain transfer function $H(f)$ shown here as magnitude and phase. Input signals are multiplied by $H(f)$ in the frequency domain.

In other words, the output from this system, seen in both time and frequency domains, is:

$$A\,\delta(t-T) * s(t) \Leftrightarrow A\,e^{-j2\pi fT}\,S(f)$$

In the frequency domain, the signal $s(t)$ is multiplied by the complex exponential. While it may not immediately seem as though $A\,e^{-j2\pi fT}S(f)$ is an undistorted version of $S(f)$, recall that the magnitude of the complex exponential is unity, or: $|e^{-j2\pi fT}| = 1$. Therefore, the magnitude of $S(f)$ is unaffected by passing through this linear system. The corresponding phase response is a straight line through the origin, as seen in Figure 6.7. In Chapter 1, it was hinted that this complex exponential had a very important role in signals analysis. Here it is seen again, appearing in the ideal linear system. Whenever a signal is encountered that has been multiplied by $e^{-j2\pi fT}$, it should be remembered that this is merely the frequency-domain equivalent of a simple time-shift operation. In the frequency domain, this impulse test signal corresponds to a constant value that can be described as *a signal containing all frequencies.* In the frequency domain, the operation $G(f) = 1 \times H(f)$ corresponding to the impulse input leads directly to the transfer function $H(f)$.

Another approach to determining the transfer function involves checking every possible input by using a sinusoidal input signal that is swept through a range of frequencies, each corresponding to a single point in $H(f)$. In the same manner, a continuous *white noise* signal (a signal containing all frequencies equally) can be introduced to the system while a spectrum analyzer displays the system transfer function $|H(f)|$ at the output.

## 6.3.2 Frequency Response Curve

The transfer function $H(f)$ is often encountered in the description of audio equipment such as amplifiers, speakers, MP3 players, or headphones. Here, the magnitude $|H(f)|$ of the transfer function is referred to as the *frequency response.*

Typical audio specifications may state that the system works over a *frequency span from 0–20 kHz*, or a graph is sometimes included on the packaging or in the box showing (generally on a logarithmic scale) amplitude variations with frequency over these frequencies. Similarly, humans can typically hear frequencies up to about 15 kHz, although the upper limit of the frequency response for human ears falls off with age, or after overexposure to loud sounds.[1] It can now be seen that, mathematically, any music signal $s(t)$ played through an entertainment system, or recorded on an MP3, is multiplied in the frequency domain by the characteristic frequency response curve described by $|H(f)|$. This observation leads to the question of which ideal linear system would produce the perfect audio amplifier; the ideal linear system of Figure 6.7 would be a good candidate.

## 6.4 Energy in Signals: Parseval's Theorem for the Fourier Transform

It has been shown (Parseval's theorem) that for the Fourier series approximation to periodic signals the power of $s(t)$ with period $T$ can be computed in the frequency domain by adding the squared magnitude of each of the complex Fourier series components $C_N$, or:

$$\int_{-\infty}^{+\infty} |s(t)|^2 dt = \sum_{N=-\infty}^{+\infty} |C_N|^2$$

The total power of a periodic (power) signal $s(t)$ may be computed in either domain.

The corresponding result for the Fourier transform is called *Parseval's theorem for the Fourier transform* or *Rayleigh's theorem.* [6] Without risk of confusion, the result may simply be called *Parseval's theorem* for both energy and power signals.

**THEOREM 6.3** *(Rayeligh's)*

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*then*

$$\int_{-\infty}^{+\infty} |s(t)|^2 \, dt = \int_{-\infty}^{+\infty} |S(f)|^2 \, df$$

This relationship permits the evaluation of the total energy in a signal $s(t)$ in whichever domain the equations are most easily integrable.

**Proof: Parseval's Theorem for the Fourier transform**

Consider the result:

[1] Ironically, such permanent ear damage can result from long-term exposure to loud music, or from audio equipment played loudly, especially through headphones.

$$LHS = \int_{-\infty}^{+\infty} |s(t)|^2 \, dt$$

$$= \int_{-\infty}^{+\infty} s(t) s^*(t) \, dt$$

Express (the complex conjugate of) the second $s(t)$ as the inverse Fourier transform of $S(f)$:

$$= \int_{-\infty}^{+\infty} s(t) \left[ \int_{-\infty}^{+\infty} S^*(f) e^{-j2\pi ft} df \right] dt$$

which can be rearranged as:

$$= \int_{-\infty}^{+\infty} \left[ \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft} dt \right] S^*(f) \, dt$$

where the inner term is the Fourier transform of $s(t)$, leaving:

$$= \int_{-\infty}^{+\infty} S(f) \, S^*(f) \, dt = RHS$$

since $LHS = RHS$, Parseval's theorem is confirmed.

### 6.4.1 Energy Spectral Density

Using Parseval's theorem, it is possible to determine the energy content of a nonperiodic signal with respect to frequency. As above, the total energy in a signal $s(t)$ is given by:

$$\text{Total Energy} = \int_{-\infty}^{+\infty} |S(f)|^2 \, df$$

Now, if the integration over all frequencies gives the total energy in a signal, it follows that the term $|S(f)|^2$ can be interpreted as the energy per frequency in $S(f)$ or, in other words, the *energy spectral density* of the signal $s(t)$, where:

$$\text{Energy in } [a, b] = \int_{a}^{b} |S(f)|^2 \, df$$

Just as the Fourier series components could be used to find the power in each of the component frequencies of a signal $s(t)$, the energy spectral density can be used to analyze the energy content of signals within any continuous range of frequencies. A signal can be measured before and after interacting with a specific linear system, and the effect of passing through that system can be precisely described as a function of frequency. Now, the output signal from a linear system is described in the frequency domain by the multiplication: $G(f) = S(f)H(f)$, where $S(f)$ is the input signal and $H(f)$ is the transfer function. Moreover, if the above equation is squared, it can be written as:

$$|G(f)|^2 = |S(f)|^2 \times |H(f)|^2$$

In other words, the output energy spectral density can be found directly from the product of the input energy spectral density and the squared magnitude of the transfer function $|H(f)|^2$.

## 6.5 Data Smoothing and the Frequency Domain

Consider an experiment in which a number of data points are to be plotted on a graph, as shown in Figure 6.8A. Inevitable experimental error results in data that appears to follow a straight line, but the data points lie scattered above or below the expected line. Such data is often *smoothed* to attenuate (random) experimental error, perhaps by replacing each point with the average of adjacent points. The result is a smoother curve, one that exhibits less variation between one point and the next. In effect, the rapid changes (high-frequency components) in the signal points have been reduced. Now, reducing high-frequency components is fundamentally a frequency-domain operation, yet this smoothing result came about by averaging points in the time domain. A more careful study of this procedure will reveal what is actually happening.

Smoothing data is a commonplace time-domain operation. To smooth data, each point might be replaced by the average of that point and the two neighboring points. In other words,

$$data[t] = \frac{1}{2}[data[t] + data[t+1]]$$

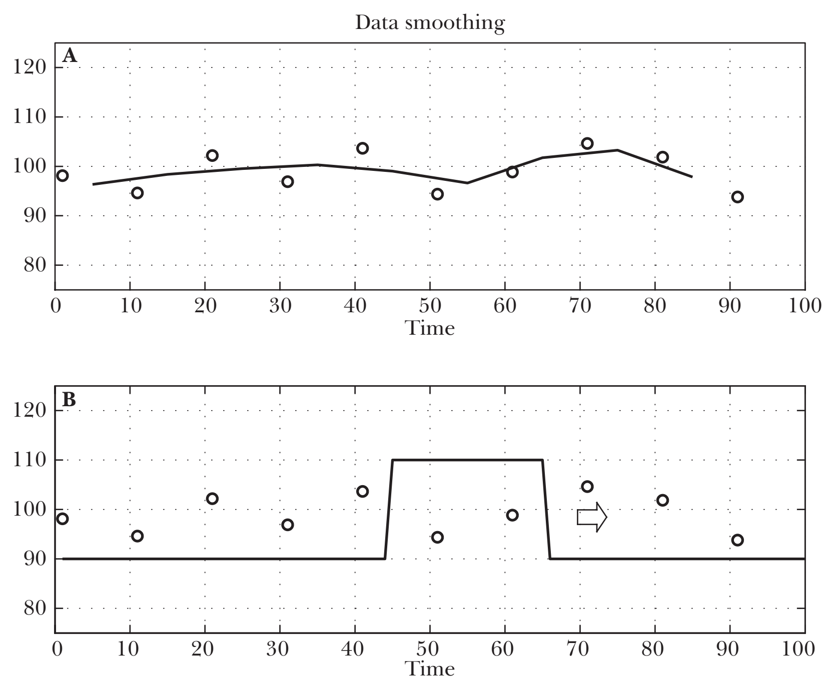



**FIGURE 6.8** The solid line in **A** shows experimental data that has been *smoothed* by averaging adjacent points. This common practice limits rapid variations in the data values. **B** shows that smoothing data by averaging adjacent values amounts to convolving the data with a rectangle spanning two points.

This operation would have to be repeated for all $t$; in other words, for every point in the data. This might be written as a summation or more properly, as a convolution in which two adjacent points are included as one signal moves across another. Consider the sketch in Figure 6.8B, where a rectangle is constructed to span two points. The smoothing operation is equivalent to convolution with this narrow rectangle. In MATLAB, the line in plot **A** was generated by convolving the data set with the rectangle modelled as a signal two points wide using `[1 1]` as:

```
output = conv(data, [1 1]) / 2;
```

using the convolution operator directly. Now, by the convolution theorem, the same results could be obtained in the frequency domain with a single multiplication by the sinc function Fourier transform of the rectangle. This operation would effectively attenuate high frequencies where the sinc amplitude is small in favor of low frequencies where the sinc reaches a maximum at the origin. This is a type of lowpass filter.

If more elaborate smoothing techniques are envisaged, the Fourier transform can be used to accurately describe the effect. For example, better smoothing using the average of three points instead of two is like using a wider rectangle. In the frequency domain, the sinc function would be narrower, cutting off more high frequencies. In the limit, if all the points are used to find one average value for the entire set, then the rectangle would become very wide, and the corresponding sinc would converge to a delta function that reflects the constant value result in the time domain. Other smoothing techniques such as taking three points and weighting the middle point can also be evaluated in this way.

Smoothing is also used in image processing where an image is to be processed to remove unwanted noise. A particular case of convolutional filtering is the so-called *zero-area filter* in which the filter shape has no DC component. Such filters are useful because they remove any DC component from the original data; in the frequency domain, any DC component is multiplied by zero.

Although data smoothing is often considered to be a time-domain operation, it can easily be analyzed and even performed in the frequency domain if so desired.

## 6.6 Ideal Filters

Signal filters are generally described in the frequency domain, as these elements selectively block or pass certain frequency components. Specific filters are described by name according to their frequency response or transfer function characteristics as, for example, *lowpass* filters, *highpass* filters, or *bandpass* filters. In the frequency domain, filtering involves multiplying an input signal by a transfer function $H(f)$, as shown in Figure 6.9.

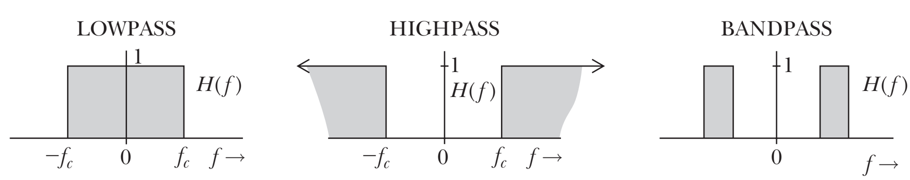



**FIGURE 6.9 Ideal Filters** Lowpass, highpass, and bandpass filters are defined in the frequency domain to reflect the range of input frequencies that can pass through the filter. In the frequency domain, a filter $H(f)$ is multiplied by an incoming signal to keep only frequencies in the defined intervals (shown shaded).

These filters are all defined in the frequency domain, and their names reflect their frequency-domain characteristics. The symmetrical appearance of each of these ideal filters is understandable when considering that the paired impulses of a sinusoid must both be included. The use and operation of these filters can now be explored with an example.

## EXAMPLE 6.3 (Filtering Frequency Components)

*The signal* $s(t) = \cos(4\pi t) + \cos(8\pi t) + \cos(16\pi t)$ *in Figure 6.10A. This signal has frequency components of 2, 4, and 8 Hz, respectively. Its Fourier transform consists of three pairs of impulses seen in Figure 6.10B, one for each of the cosine components.*

1. What filter could be used to eliminate components above 3 Hz?
2. What filter could be used to eliminate components below 5 Hz?
3. What filter could be used to pass only the 4 Hz component?

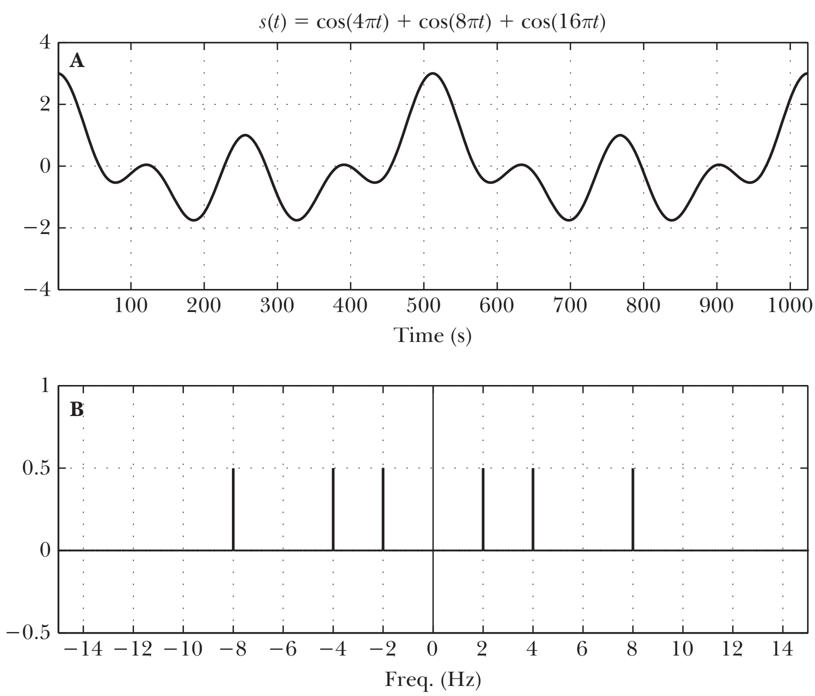



**FIGURE 6.10 Ideal Filters** The periodic signal $s(t)$ shown in **A** consists of three equal amplitude cosine components at (2 Hz, 4 Hz, and 8 Hz), as may be identified in the Fourier transform $S(f)$ shown in **B**. Filters may be used to extract the components, as seen in Figure 6.11.

**Solution:**

Figure 6.11 shows three different ideal filters acting on the signal $s(t)$ from Figure 6.10. In each case, the filters $H(f)$ are multiplied by $S(f)$ and only certain frequency components are not set to zero. For these ideal filters, the frequency threshold above which signals will or will not pass is called the *cutoff frequency*.

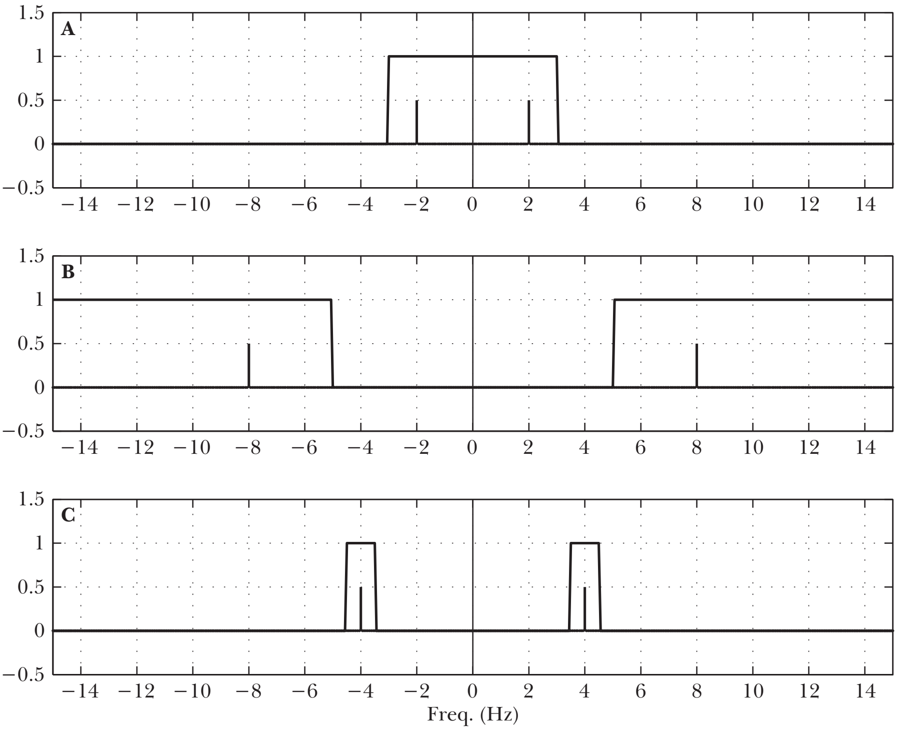



**FIGURE 6.11 Ideal Filters** Using the signal in Figure 6.10, the component cosines are revealed when an ideal filter $H(f)$ is multiplied in the frequency domain. The lowpass filter in **A** passes only frequencies below 3 Hz, leaving a 2 Hz cosine. The highpass filter in **B** passes only frequencies above 5 Hz, leaving a 8 Hz cosine. The bandpass filter in **C** passes only frequencies around 4 Hz, leaving a 4 Hz cosine.

1. Filter **A** is a lowpass filter with a cutoff frequency of 3 Hz. Only components below 3 Hz can pass through. Only the 2 Hz cosine from $s(t)$ remains after filtering.
2. Filter **B** is a highpass filter with a cutoff frequency of 5 Hz. Only components above 5 Hz can pass through. Only the 8 Hz cosine from $s(t)$ remains after filtering.
3. Filter **C** is a bandpass filter with a passband around 4 Hz. Only components between 3 Hz and 5 Hz can pass through. Only the 4 Hz cosine from $s(t)$ remains after filtering.

Of course, each of the above filtering operation could be accomplished with identical results in the time domain by convolving $s(t)$ with the inverse Fourier transform $h(t)$ of each filter transfer function $H(f)$. For example, a lowpass filter would be accomplished through convolution with a suitable sinc function. The Fourier transform provides a mechanism to design such *convolutional filters* in a straightforward way. It is left as an exercise to compute the three corresponding time-domain filters.

Ideal filters should exactly block or pass certain frequencies without affecting the signal otherwise. An ideal response function $h(t)$ has already been described. An ideal filter should have distortionless behavior outside of the filtered region of frequencies. This implies that the phase response can be linear, meaning that a simple delay is acceptable in the output signal. One practical difficulty of constructing ideal filters will now be investigated.

### 6.6.1 The Ideal Lowpass Filter Is Not Causal

While the ideal lowpass filter is readily visualized as an even rectangle in the frequency domain, inspection in the time domain reveals a significant limitation of this model. In Section 3.6, it was seen that the response function $h(t)$ of a *causal system* requires that $h(t)$ is zero for all $t < 0$. Because the ideal lowpass filter $H(f)$ has an impulse response $h(t)$ that is in the form of a sinc function, as in Figure 6.12, *the ideal lowpass filter is noncausal.*

The sinc function impulse response violates causality because it is not zero for all $t < 0$. This observation alone confirms that it is impossible to construct an ideal lowpass filter. While very good filters can nonetheless be designed, the goal of constructing the ideal filter is nonrealizable. Figure 6.13 shows one approach to resolving this dilemma.

The problem with a noncausal filter is that the circuit must respond to an input that has not yet arrived. A practical high-order lowpass filter will trade a delay in time for knowing what is coming next, and thereby seemingly achieve the impossible. Consider the impulse response based on $h(t - T)$ as shown in Figure 6.13. By allowing some delay in the system, a causal impulse response can be created that encompasses much of sinc function shape. The magnitude of the Fourier transform of $h(t - T)$ remains unaffected by this time shift. Using this approach, what is the best ideal lowpass filter that could be constructed? Regardless of how much shift is contemplated, the sinc function must be cut off for $t < 0$ if the lowpass filter is to obey causality. One way to accomplish this cutoff is to first multiply the sinc function by a wide rectangle to render $h(t)$ zero at its two extremes. Of course, this will have some effect on the frequency-domain ideal lowpass filter characteristics. Specifically, multiplying $H(f)$ by a wide rectangle in the frequency domain will convolve $h(t)$ with a narrow sinc function, thereby introducing small ripples, or ringing, on the edges of the ideal rectangle filter shape. This observation characterizes the frequency-domain appearance of any high-order lowpass filter.

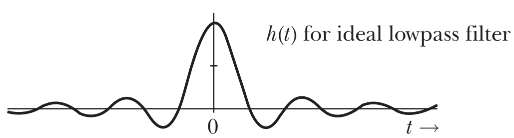



**FIGURE 6.12 A Noncausal Filter** For an ideal lowpass filter, $H(f)$ is a even rectangle. The corresponding impulse response $h(t)$ shown here is noncausal. Such a perfect filter is impossible to build.

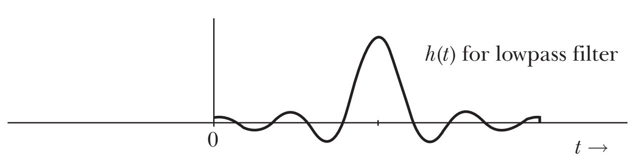



**FIGURE 6.13 A Causal Filter** The ideal lowpass filter impulse response of Figure 6.12 can be made causal by introducing a delay and truncating $h(t)$ in time to ensure that $h(t) = 0$ for $t < 0$. As the delay is increased, the filter $h(t)$ approaches the ideal sinc shape.

## 6.7 A Real Lowpass Filter

In Section 3.4.6, the RC-circuit of Figure 3.12 was presented to illustrate an impulse response convolution with an input signal in the time domain. In practice, this circuit is an example of a *first-order lowpass filter*. The specific characteristics of this filter may now be studied in the frequency domain.

Consider $h(t) = a\,e^{-at}$ for $t > 0$ and $a > 0$ as in Figure 6.14. For the capacitor and resistor as shown, $a = 1/RC$. For a real lowpass filter such as this, the cutoff frequency will not be abrupt and absolute as in the ideal lowpass filter; instead the output will exhibit gradually increasing attenuation at higher frequencies. In that case, the cutoff frequency $f_c$ is defined at the point where an input frequency component would be reduced to half its power; in this case, this *half power point* will be at $f_c = a/(2\pi)$.

The corresponding transfer function $H(f)$ should have real part even and imaginary part odd. The value $H(0)$ should equal the area under $h(t)$, or:

$$H(0) = \int_{-\infty}^{+\infty} h(t)\;dt = \int_{0}^{+\infty} a\,e^{-at}\;dt = 1$$

The Fourier transform of $h(t)$ is:

$$\begin{aligned} H(f) &= \int_{0}^{+\infty} a\,e^{-at}\,e^{-j2\pi ft}\;dt \\ &= \int_{0}^{+\infty} a\,e^{-(a+j2\pi f)t}\;dt \\ &= \frac{-a}{a+j2\pi f}\,e^{-(a+j2\pi f)t}\Big|_{t=0}^{\infty} \end{aligned}$$

therefore:

$$H(f) = \frac{a}{a+j2\pi f} \tag{6.2}$$

**Check:** $H(0) = 1$ as expected.

A deeper analysis of the circuit behavior is now possible, and phase of the (complex) transfer function $H(f)$. To this end, the real and imaginary parts are separated; the numerator and denominator are multiplied by the complex conjugate of the denominator as:

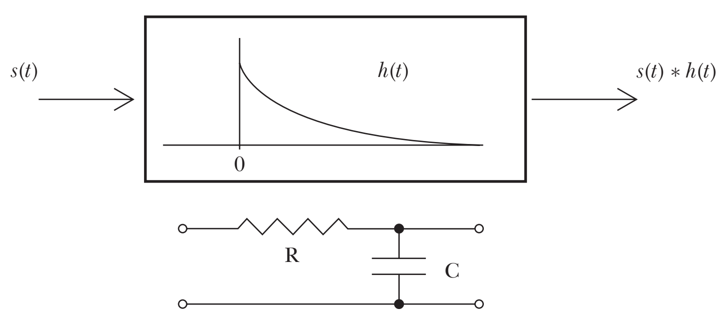



**FIGURE 6.14** A real-world causal impulse response (RC-circuit). The output cannot change until an input arrives: $h(t) = 0$ for $t < 0$.

$$H(f) = \frac{a}{a + j2\pi f} \times \frac{a - j2\pi f}{a - j2\pi f}$$

$$H(f) = \frac{a^2 - ja2\pi f}{a^2 + (2\pi f)^2}$$

Giving a **real part**:

$$Re(H(f)) = \frac{a^2}{a^2 + (2\pi f)^2}$$

And an **imaginary part**:

$$Im(H(f)) = \frac{-a2\pi f}{a^2 + (2\pi f)^2}$$

Which leads directly to the **phase** of $H(f)$ as:

$$\Phi(f) = tan^{-1}\left[\frac{Im}{Re}\right] = tan^{-1}\left[\frac{-2\pi f}{a}\right]$$

Similarly, the **magnitude** of $H(f)$ is given by:

$$|H(f)| = [H(f)H^*(f)]^{\frac{1}{2}} = \left[\frac{a}{a + j2\pi f} \times \frac{a}{a - j2\pi f}\right]^{\frac{1}{2}} = \frac{a}{\sqrt{a^2 + (2\pi f)^2}}$$

The real and imaginary components of $H(f)$ are shown in Figure 6.15A where it may be observed that the real part is even and the imaginary part odd, as expected. The magnitude $|H(f)|$ is shown in Figure 6.15B where it can be seen to be far from the rectangular shape of the ideal lowpass filter. An ideal lowpass filter would pass all frequencies below $f_c$ and block all frequencies above $f_c$. A simple filter such as $H(f)$ will slightly attenuate frequencies below the cutoff frequency $f_c$ and will not cleanly eliminate higher frequencies. For the closely spaced frequencies of Figure 6.10, this simple lowpass filter would have little useful effect in isolating specific frequency components.

As expected, the magnitude of $H(f)$ is real and even. Consequently it is unnecessary to show the negative side of the graph, and $|H(f)|$ is normally plotted on a one-sided log-log graph as magnitude and phase as in Figure 6.16. The cutoff frequency (set to 1 Hz) is shown highlighted. This is the *half power point* at which the power in an incoming cosine with frequency $f_c$ is reduced by half; its amplitude $A$ is reduced by a factor of $\sqrt{2}$ to be $0.707A$.

Finally, Figure 6.16B shows that there is a phase change associated with filter delay. This is the consequence of building a causal filter. In particular, the incoming signal component phase is changed by –45 degrees ($-\pi/4$ rad) at the cutoff frequency, and by up to 90 degrees at higher frequencies. The log-log graphs of magnitude Figure 6.16A and the log-linear graph of phase Figure 6.16B form a distinctive graphical presentation called a *Bode plot*[2] that is very useful in showing the frequency and phase response of systems and particularly filters.

[2] Pronounced *bow-dee.*

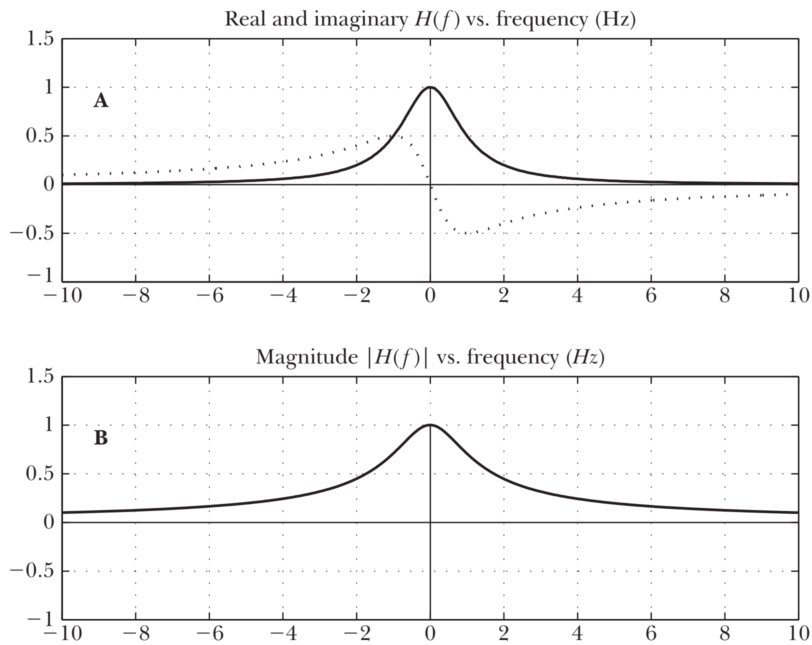



**FIGURE 6.15 Lowpass filter** The Fourier transform $H(f)$ of the impulse response $h(t)$ from Figure 6.14 is shown as real and imaginary components **A** and as magnitude vs. frequency **B**.

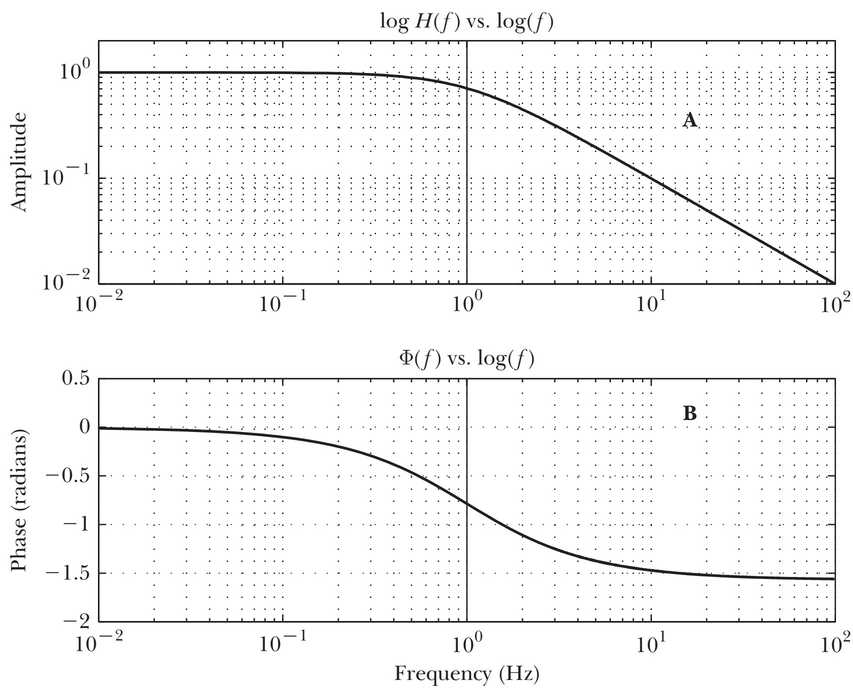



**FIGURE 6.16 Lowpass filter** The Fourier transform $H(f)$ of the impulse response $h(t)$ from Figure 6.14 is more commonly viewed as magnitude **A** and phase **B** on a one-sided graph with a logarithmic frequency scale. Here, the cutoff frequency (highlighted) is 1 Hz.

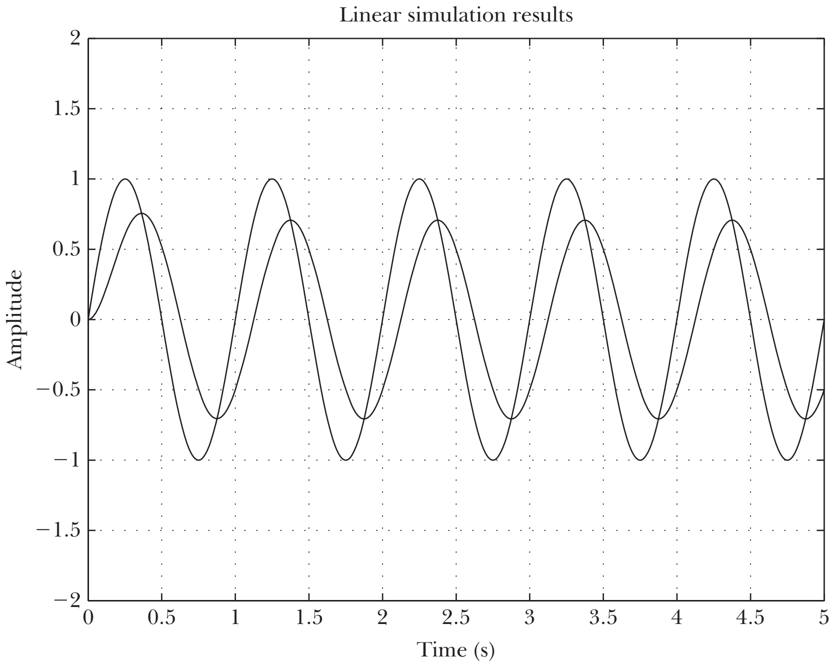



**FIGURE 6.17** Lowpass filter input and output signals. A sine wave input at the cutoff frequency is attenuated with amplitude $A = 0.707$ and phase change $\Phi = -45$ degrees.

From the Bode plot, the effect of this filter in the time domain can be illustrated in Figure 6.17, where an input sinusoid at the cutoff frequency is shown along with the output sinusoid, attenuated and shifted in phase as:

$$\sin(2\pi f_c t) \longrightarrow \frac{1}{\sqrt{2}} \sin\left(2\pi f_c t - \frac{\pi}{4}\right)$$

In every case, a sinusoidal input results in a sinusoidal output at the same frequency that is attenuated and that undergoes a frequency shift.

**MATLAB Example 1: First-Order Filter** The results obtained above can be readily duplicated in MATLAB, where the transfer function of the lowpass filter is used directly to obtain the Bode plot and to study the circuit behavior in more detail. The transfer function for this circuit was given in Eqn. 6.2 as:

$$H(f) = \frac{a}{a + j2\pi f}$$

Where $a = 1/RC$ is determined by the component values in the RC circuit, and $f_c = a/2\pi$ Hz is the lowpass filter cutoff frequency. The cutoff frequency can also be expressed in radians as $\omega_c = a$. To study this transfer function in MATLAB, let $s = j2\pi f$ or $s = j\omega$ to give the conventional form:

$$H(f) = \frac{a}{a + s}, \quad \text{with } s = j2\pi f \tag{6.3}$$

Now, this arithmetic expression in *s* will directly define a MATLAB system model named `sys`, as:

```
a = 2*pi;          % cutoff frequency = 1 Hz = 2 pi rad/s
s = tf('s');       % establish a transfer function on s
sys = a/(a+s)      % define this system
Transfer function:

           6.283
          ---------
          s + 6.283
```

Once `sys` is defined in this way,[3] this circuit can be studied in many ways, including:

- `bode(sys)`—create a Bode plot as in Figure 6.16.
- `impulse(sys)`—create a time-domain plot of impulse response $h(t)$ as in Figure 6.14.
- `step(sys)`—create a time-domain plot of response to a step input.
- `lsim(sys)`—create a time-domain plot of response to an arbitrary input, as shown below.

The function `lsim()` tests the system with a supplied input signal and then plots *both* the input signal and the output signal on the same graph. The sinusoidal test signal seen in Figure 6.17 is produced as:

```
t = 0:0.01:5;                 % define a time axis
input = sin(a*t);             % define a test input
lsim(sys,input,t);            % plot the test input response
axis([0 5 -2 2]);             % adjust the graph axes
grid on;                      % include grid lines
```

Finally, the above example can be repeated with a pulse input signal represented by a shifted unit rectangle, as seen in Figure 6.18:

```
t = 0:0.01:5;                   % define a time axis
input = rectpuls(t-1);          % define a test input
lsim(sys,input,t);              % plot the test input response
axis([0 5 -0.5 1.5);            % adjust the graph axes
grid on;                        % include grid lines
```

**Check:** The output signal in Figure 6.18 resembles the voltage across a charging-then-discharging capacitor.

## 6.8 The Modulation Theorem

Other properties of the Fourier transform are readily visualized or explained in light of the convolution theorem. Consider the short burst of cosine signal shown in Figure 6.19.

[3] Specifically, the use of `'s'` is *required* in the system definition.

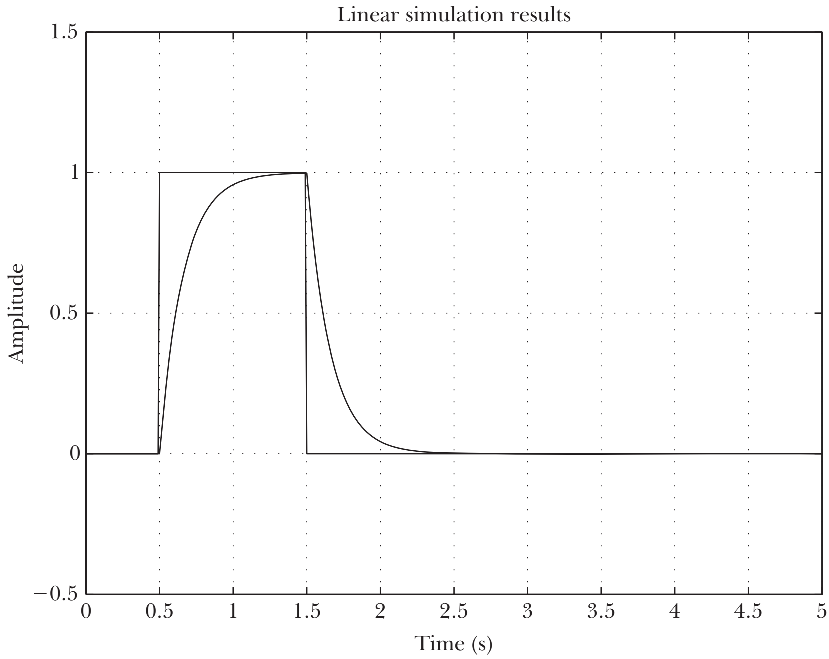



**FIGURE 6.18** A first-order lowpass filter *system* is defined in MATLAB by a transfer function, then the system behavior is examined here for a pulse input.

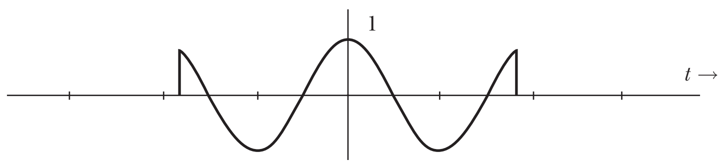



**FIGURE 6.19** These few cycles of a cosine may be modelled as a cosine multiplied by a rectangle.

It may be noted that this signal is simply the product of a cosine with a rectangle. The Fourier transform can be derived from the function $g(t) = \cos(2\pi f_0 t) \times s(t)$, where $s(t)$ is a rectangle spanning a fixed time interval. Truncation in time is commonly modelled using this approach. The Fourier transform of $\cos(2\pi f_0 t)$ consists of a pair of half-area impulses, located at $f = -f_0$ and $f = +f_0$, respectively. Multiplication in one domain is equivalent to convolution in the other, so the sinc function $S(f)$ is effectively copied, through the convolution operation, to lie centered at each of the above frequencies. The result is $G(f) = \frac{1}{2}S(f - f_0) + \frac{1}{2}S(f + f_0)$, as shown in Figure 6.20.

It was observed in Chapter 1 that when two sinusoids were multiplied together, the result was a sinusoid at both the sum and difference of the two sinusoidal frequencies. It has now been shown that this rule applies generally to the product of any signal $s(t) \times \cos(2\pi f_0 t)$. The interesting result of this example is in the observation that when the rectangle was multiplied by the cosine, the resulting Fourier transform consisted of two copies of the sinc function $S(f)$ positioned at $+f_0$ and $-f_0$, respectively.

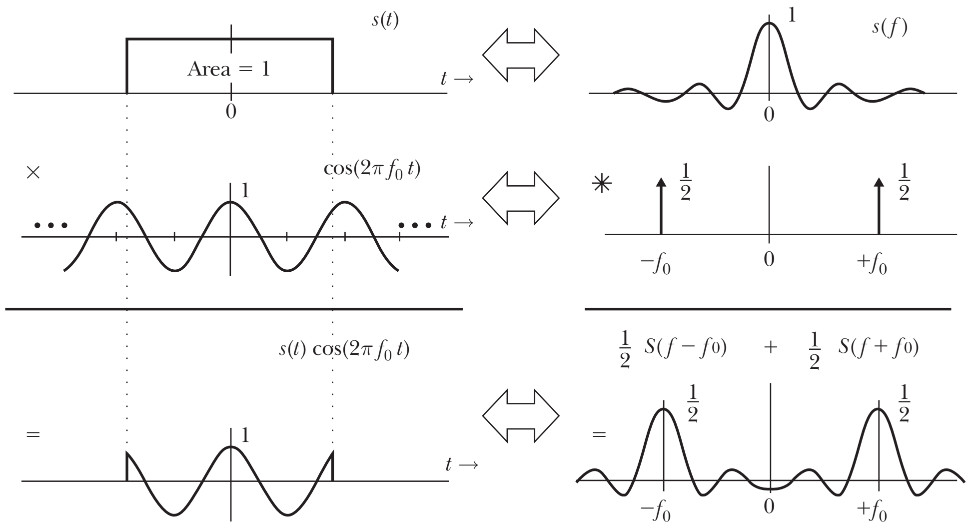



**FIGURE 6.20** In the time domain, a cosine multiplied by a rectangle. In the frequency domain, a pair of impulses convolved with a sinc function. The Fourier transform is obtained by inspection.

The modulation theorem of the Fourier transform can be generalized as:

**THEOREM 6.4** *(Modulation)*

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*then*

$$s(t)\cos(2\pi f_0 t) \xleftrightarrow{\mathcal{F}} \frac{1}{2}\, S(f + f_0) + \frac{1}{2}\, S(f - f_0) \tag{6.4}$$

Multiplication of $s(t)$ by a cosine in one domain leads to sum and difference copies of $S(f)$ in the other.

## 6.8.1 A Voice Privacy System

When sending voice messages from one place to another, it is possible that someone may be listening in to the conversation. This concern is most acute with radio devices such as cordless telephones. The modulation theorem leads to a simple method of *voice scrambling* that is commonly used where a degree of secrecy is desired in voice communications.

The problem of scrambling a voice is more difficult than it may appear at first glance. First, whatever method is used, the resulting scrambled signal must pass over the same telephone line or radio system that would normally carry the unscrambled voice. Therefore, the frequency components in the scrambled voice signal must still generally fall within the range of frequencies describing a normal voice signal. Second, the scrambled message must be readily reassembled into the

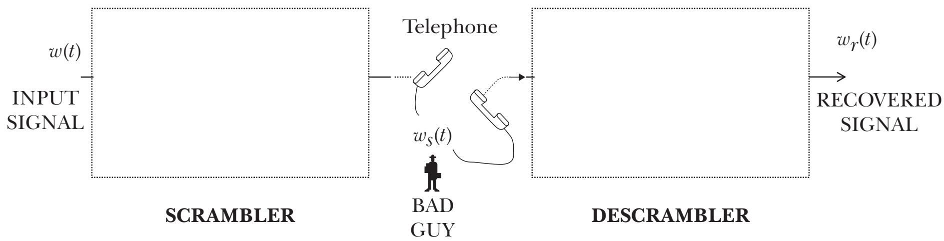



**FIGURE 6.21** Design of a voice-scrambling system requires that an intercepted message be unintelligible to a listener, but readily recovered by the intended recipient. Moreover, the signal on the phone line must match the general frequency characteristics of a voice.

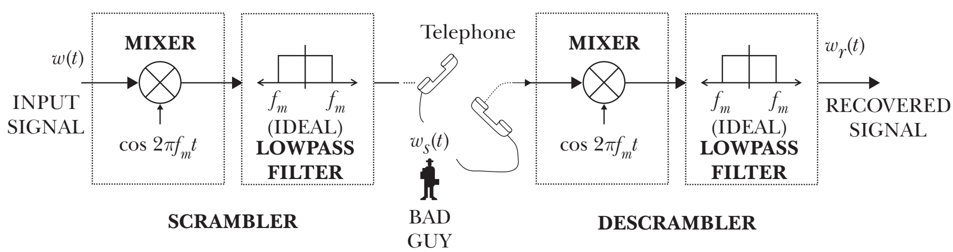



**FIGURE 6.22 Spectral Inversion** Multiplication by a suitable cosine followed by lowpass filtering effectively rearranges frequency components, making the signal unintelligible without destroying its spectral integrity. The original signal is restored when the process is repeated using an identical circuit.

original voice message for understanding by the intended recipient. Finally, of course, the scrambled signal must be unintelligible to anyone illicitly receiving the signal, as shown in Figure 6.21. Now, the human brain is a remarkable signal processing instrument in its own right, well adapted for making sense of all manner of distorted and noisy voice signals; more than one cunning approach to voice scrambling may well be decipherable by carefully listening to the message.

**Spectral Inversion** One simple method of voice scrambling uses the modulation theorem to accomplish *spectral inversion*, or *frequency inversion*. This approach involves swapping high- and low-frequency components in the frequency domain. The resulting signal obeys the three scrambling constraints outlined above. Because the same spectral components are present, but switched around, the signal can still pass through a voice transmission system. The signal can be readily descrambled by swapping the frequency components a second time; in fact, the same hardware can be used for both scrambling and descrambling. Finally, a frequency inverted signal is suitably unintelligible to the human ear; an eavesdropper can hear the signal, but cannot directly understand the message.

Consider a telephone voice signal $s(t)$, constrained or *bandlimited* to contain frequencies below 3000 Hz. No higher frequencies will pass over the channel. For simplicity, the signal $s(t)$ below will serve as a test signal, where $s(t)$ consists of the sum of two cosine signals having voice-like frequencies (800 Hz and 2000 Hz). The resulting signal is shown in time and frequency in Figure 6.23A, where there are two distinct cosine components visible in the frequency domain. The signal $s(t)$ would easily pass over a telephone line.

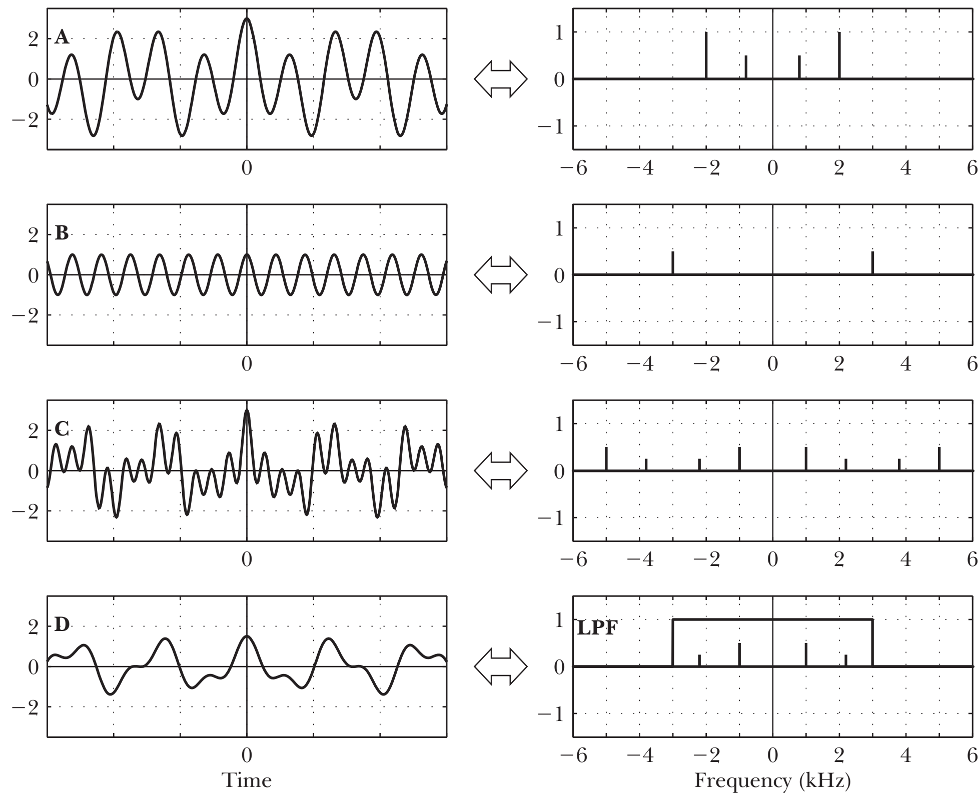



**FIGURE 6.23** **Spectral Inversion: Scrambling** Fourier transform pairs show the process of scrambling a voice signal. The original voice signal in **A** is multiplied by the cosine in **B** to give the signal **C**. After lowpass filtering, **D** is the scrambled signal, ready for transmission. Note that the spectral components in **A** and **D** are merely rearranged.

$$s(t) = \cos(1600\pi t) + 2\cos(4000\pi t)$$

Scrambling a voice message with this method consists of mixing the voice signal with a cosine before passing through the telephone lines. Let the mixing signal be a 3000 Hz cosine as in Figure 6.23B. In the frequency domain, the signal $S(f)$ is convolved with the paired impulses describing the cosine mixing signal. The resulting signal is shown in Figure 6.23C, where the modulation theorem describes a half amplitude copy of the original spectrum on either side of the origin, giving four components at the sum and difference frequencies.

$$s(t)\cos(6000\pi t) = \frac{1}{2}[\cos(7600\pi t) + \cos(4400\pi t)] + \frac{1}{2}[2\cos(10000\pi t) + 2\cos(2000\pi t)]$$

If a lowpass filter is applied to remove components above 3000 Hz, the signal in Figure 6.23D remains:

$$\text{scrambled } s(t) = \frac{1}{2}[\cos(4400\pi t) + 2\cos(2000\pi t)]$$

The result is a signal that is the sum of two cosines like the original $s(t)$ but at new frequencies (2200 Hz and 1000 Hz), which leads to the term *spectral inversion*. This is the signal that would be heard by a wiretapper, and that would emerge at

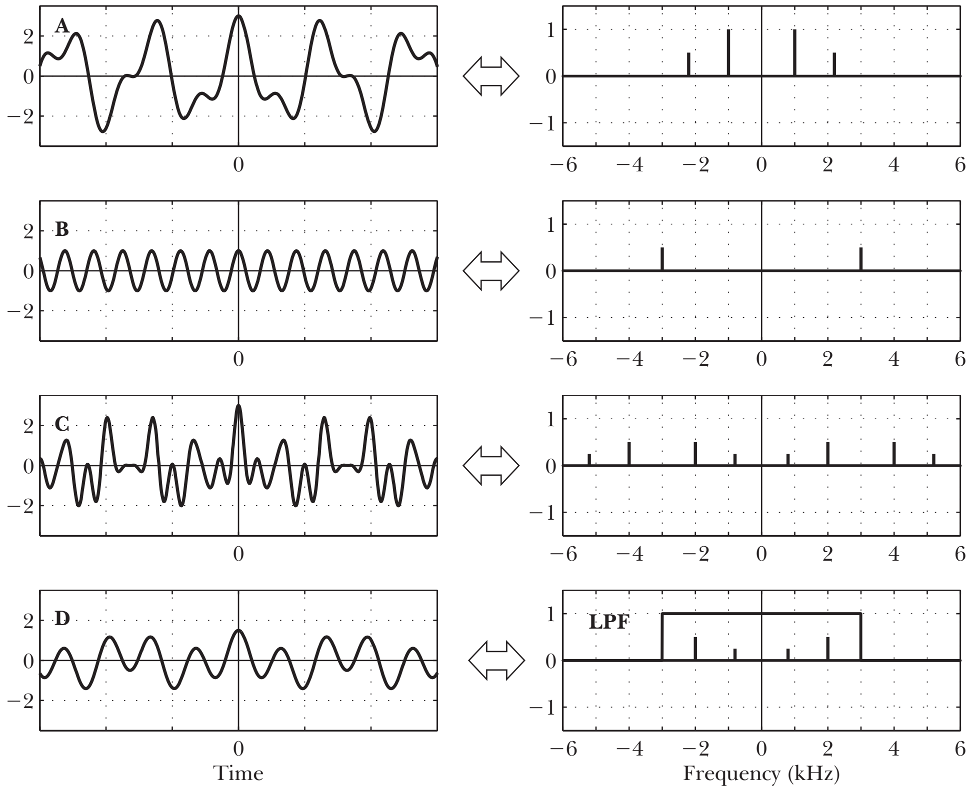



**FIGURE 6.24 Spectral Inversion: Descrambling** Fourier transform pairs show the process of descrambling the voice signal from Figure 6.23. The scrambled voice signal in **A** is multiplied by the cosine in **B** to give the signal **C**. After lowpass filtering, **D** is the descrambled signal, a copy of the original voice signal in Figure 6.23A.

the other end of the telephone line. Compared to the original $s(t)$ in **A**, the scrambled version in **D** is quite different. In the frequency domain, the high-frequency components in **A** are low in frequency in **D**, and vice versa, yet the overall bandwidth has not changed. The entire signal still fits within the available telephone bandwidth, although it would no longer sound intelligible.

Of course, inverting the frequencies again would yield the original signal. The voice message can be restored at the receiver by simply repeating the scrambling process with another 3000 Hz cosine. In Figure 6.24, the scrambled signal is passed through the same process to yield the signal in Figure 6.24D. Because half the signal is removed by lowpass filtering, the circuit output is only half amplitude. The input signal **A** has been multiplied by two, and the signal **D** is only ½ the original $s(t)$. The signal is easily amplified to restore the full signal.

While this is hardly a secure method of sending secret messages, it offers suitable protection from the casual listener. Anyone wise to this method would simply need his/her own 3000 Hz mixing signal to uncover the hidden voice. Still, this is a commonplace approach that is easy to implement, and this circuitry has been built in to some models of cordless telephones. For many years, it has been illegal to use frequency inversion descramblers intended for this purpose in the United States,[4]

[4] Public Law 99-508: The Electronic Communications Privacy Act of 1986 (ECPA).

although construction projects described as being *for educational use only* occasionally appear in electronics hobbyist magazines.

## 6.9 Periodic Signals and the Fourier Transform

Fourier series components were introduced to describe time-domain periodic signals. The Fourier transform was developed to handle nonperiodic signals, but its use extends to both periodic and nonperiodic signals. In particular, the Fourier transform of a periodic signal is distinguished by being a line spectrum, made up of equally spaced impulses, lying at integer multiples of the fundamental frequency. This line spectrum corresponds exactly to the Fourier series components where the value of each component can be found in the area of each impulse (moreover, an impulse is sketched with a height reflecting its area). The convolution property allows a unique perspective on the creation and analysis of periodic signals.

Since the convolution of any signal with an impulse gives a copy of that function, any periodic signal can be considered as the convolution of a single function with a series of impulses spaced by the period of the signal. For example, a 1 Hz square wave can be viewed as multiple copies of a rectangle of width 0.5 sec spaced every 1 sec. Mathematically, the square wave signal can be written as the convolution of the rectangle with a series of impulses spaced every 1 sec, as shown in Figure 6.25. This approach can be used to model any periodic signal. Once again, a complicated signal has been broken down into components that are more easily handled. Completion of this problem requires a study of the Fourier transform of an infinite series of impulses.

### 6.9.1 The Impulse Train

The *impulse train*, also known as a *comb function*, *Dirac comb*, *picket fence*, *sampling function*, or *Shah function*, is as useful as this variety of names suggests. Viewed as a periodic signal, the Fourier series is relatively easy to compute for this unit impulse $\delta(t)$ repeated with period $T$:

$$C_N = \frac{1}{T}\int_{-T/2}^{+T/2} \delta(t)e^{-j2\pi Nf_0t}\,dt = \frac{1}{T}e^{-j2\pi Nf_00} = \frac{1}{T}$$

This result shows that the complex Fourier series consists of components located at integer multiples of the fundamental frequency $f_0 = 1/T$, *each with the same amplitude* equal to $1/T$. The Fourier transform of an impulse train is equal to an impulse train in the transform domain, where the spacing and area of the transform domain impulses varies as $1/T$, as seen in Figure 6.26.

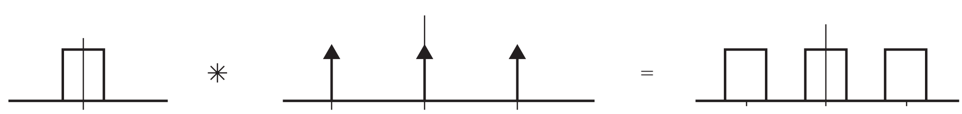



**FIGURE 6.25 Construction of a Periodic Signal** A periodic signal can be modelled as a single shape duplicated at regular intervals. A square wave can be formed by convolving a rectangle with an infinite series of impulses.

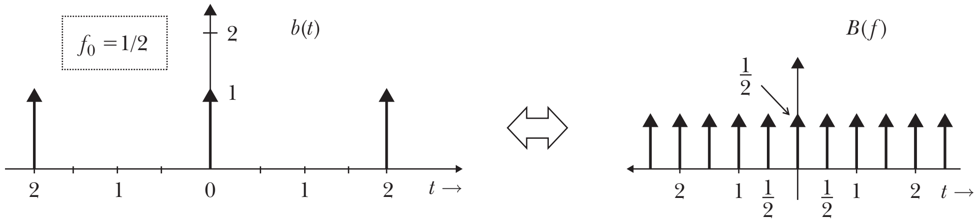



**FIGURE 6.26** The Fourier transform of an impulse train $b(t)$ is equal to an impulse train $B(f)$ in the frequency domain, where the spacing and area of the transform domain impulses varies as $1/T$.

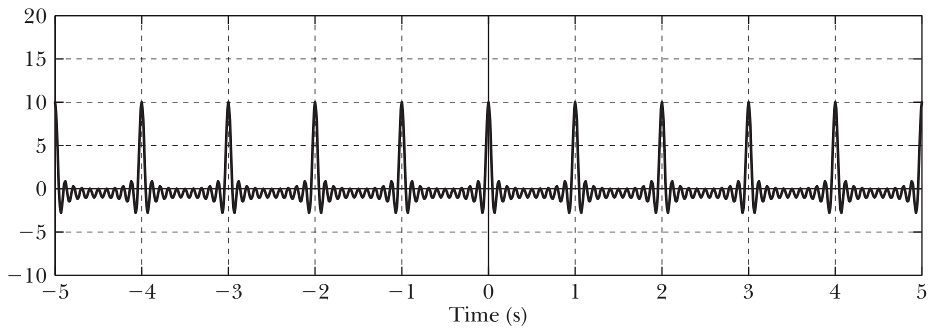



**FIGURE 6.27** As suggested by its Fourier transform, an impulse train can be approximated by the sum of an infinite series of cosine terms, each having the same amplitude. The sum of only 10 cosines is shown here.

From yet another perspective, the impulse train may be viewed as impulse pairs, centered on the origin. Each pair of impulses corresponds to a cosine in the transform domain. By this logic, the Fourier transform of an infinite impulse train is an infinite series of cosines at integral multiples of some base frequency, all with the same amplitude. This summation, computed below for 10 terms, should also describe an impulse train. Note there must be a component found at the origin in either domain, since there is an (infinite) area under the transform. The following MATLAB code may be used to confirm this observation, where the result is shown in Figure 6.27.

```
t=-5:0.01:5;
impulses = 0;
for n=1:10     % implement the summation as a loop
 impulses = impulses + cos(2*pi*n*t);
end            % NOTE: nothing happens until end
plot(t,impulses);
axis([-5, 5, -10, 20]);
grid on;
```

## 6.9.2 General Appearance of Periodic Signals

The convolution property may now be used to describe, in general terms, the Fourier transform of any periodic signal. Consider a signal with period $T$ s, modelled as the convolution of a nonperiodic element $x(t)$ with an impulse train of period $T$. With no further computation, the Fourier transform is given by the multiplication of $X(f)$ with an impulse train in frequency with components spaced $1/T$ Hz. In other words, the Fourier transform will be an infinite series of impulse spaces at $1/T$ Hz, and whose areas follow the shape of the Fourier transform $X(f)$. This observation will now be examined using the familiar square wave.

## 6.9.3 The Fourier Transform of a Square Wave

A square wave may be described as a unit rectangle convolved with a series of unit impulses spaced at the period $= T$ s. This corresponds to a frequency-domain signal in which a sinc function is multiplied by a series of unit impulses spaced at $1/T$ Hz, as in Figure 6.28. The relative weights of each spectral component follow a shape or *envelope* that is given by the Fourier transform of a single period of the square wave. In other words, a periodic square wave has a Fourier transform that is a line spectrum with a sinc envelope corresponding to the Fourier transform of a single rectangle.

It can be observed that the sinc function is precisely the width necessary so that zero crossings fall on even-numbered integral multiples of the fundamental frequency. This reflects the common observation that only odd-numbered harmonics are present in the distinctive square wave spectrum.

**Changing the Pulse Train Appearance** For any periodic signal, the line spacing in frequency is determined by the period of the signal in the time domain. The variations in the line areas are determined by the Fourier transform of the

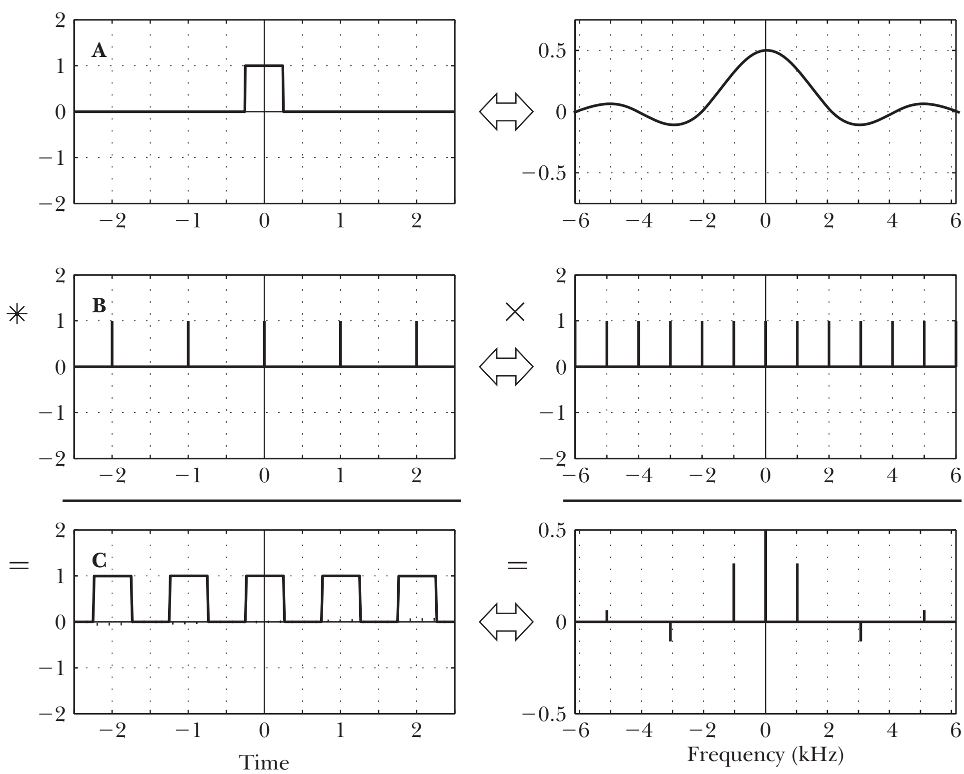



**FIGURE 6.28 Construction of a Periodic Signal** A periodic signal can be modelled as a single shape duplicated at regular intervals. Here, a square wave **C** is formed by convolving the rectangle in **A** with the series of impulses in **B**. In the frequency domain, a sinc shape is multiplied by a series of impulses, to give the characteristic Fourier transform of a square wave.

periodic shape in the time domain. The square wave is a rectangular pulse train with a 50 percent duty cycle, or having a period twice the width of each rectangle.

1. Moving the rectangles farther apart in time narrows the frequency-domain line spacing, while the sinc envelope remains unchanged.
2. Making the rectangles narrower makes the frequency-domain sinc envelope wider, while the line spacing remains unchanged (the period is unaffected).

If the rectangles are simultaneously made wider and farther apart together (or narrower and closer together) so as to maintain a square wave appearance, the line spacing changes in the frequency domain and proportionally, the width of the sinc envelope adjusts such that each line stays the same relative magnitude, while the spacing between lines varies. Just as its shape distinguishes a square wave without regard to period, its transform domain appearance is distinctive regardless of the exact line spacing.

**EXAMPLE 6.4 (Fourier Transform of a Pulse Train)**

*Describe the general appearance of the Fourier transform of a pulse train with a period of 2 s and pulse width of 0.2 s.*

**Solution:**
The Fourier transform of this periodic signal will be a line spectrum with lines spaced every ½ Hz along the frequency axis. The weight of each impulse is found by multiplying the impulse train by the Fourier transform of the rectangle, which is a sinc function in which the first zero-crossing occurs at $1/0.2 = 5$ Hz. Therefore, the central lobe of the sinc envelope will span 10 impulses. There will be a positive component at 0 Hz, because the pulse train has a DC component.

### 6.9.4 Other Periodic Waveforms

This analysis of periodic waveforms is quite useful. By knowing some simple transforms and the rules relating them, the Fourier transforms of many periodic signals can be derived by inspection. For example, the Fourier transform of a periodic triangle wave would have impulse components spaced at $1/T$ Hz, with an envelope described by $\text{sinc}^2(f)$, since the Fourier transform of a single triangle goes as $\text{sinc}^2(f)$.

Even the impulse train itself can be analyzed as a periodic repetition of unit impulses in the time domain, spaced every $T$ seconds. The corresponding line spectrum will have components spaced at $1/T$ Hz, and an envelope following the Fourier transform of a single impulse. Since the Fourier transform of a single impulse is a constant, this envelope is constant, and the Fourier transform of the impulse train is another impulse train, as expected. Specifically, a unit impulse train with a 1-second period consists of a line spectrum with components every 1 Hz, with each impulse having unit weight.

## 6.10 The Analog Spectrum Analyzer

The frequency components of a periodic waveform can be examined in the laboratory using a *spectrum analyzer*. Like an (analog) oscilloscope, an analog spectrum analyzer is best suited to the study of repetitive signals. A conventional (analog) spectrum analyzer works very much like a radio receiver to extract the frequency components present in an input waveform.

Consider an experiment in which an inexpensive AM radio receiver is used to tune across the broadcast band from 540 to 1600 kHz. While tuning the radio, the level of the signal from the radio speaker may be plotted as a function of frequency. The result shown in Figure 6.29 would be a measurement of the radio frequency spectrum in the area of the receiver. Peaks in the graph correspond to the presence of frequency components lying at the transmitting frequencies of local broadcast radio stations. This is the frequency spectrum of components from 540 to 1600 kHz.

In operation, the circuitry inside an analog spectrum analyzer essentially tunes across the frequency range of interest to produce a display that resembles Figure 6.29. In a like manner, the trace can generally be seen sweeping across the display as the spectrum is extracted. As in this example, the display on such a spectrum analyzer actually shows the magnitude of frequency components; furthermore, since the magnitude is always even, only a one-sided display is required. Like an oscilloscope, a spectrum analyzer cannot distinguish between a sine and a cosine waveform. Still, the magnitude shows the relative contribution of each frequency component. Modern digital spectrum analyzers actually calculate and display the Fourier series components of an input signal using an onboard computer, much as they would be computed in MATLAB.

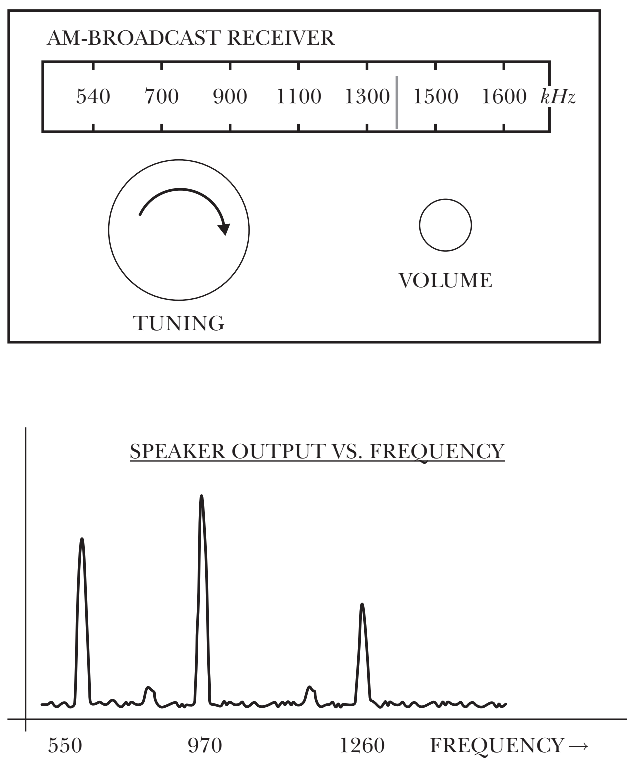



**FIGURE 6.29** Comparing the operation of an analog spectrum analyzer to that of a simple radio receiver.

## 6.11 Conclusions

The Fourier transform can be approached as the continuous form of the Fourier series. While the Fourier series applies specifically to periodic signals, the Fourier transform can be used for all manner of signals, whether periodic or nonperiodic, odd or even, real or imaginary. One of the most important properties of the Fourier transform is found in the observation that the (inverse) Fourier transform of two signals convolved in one domain becomes the product of their respective (inverse) Fourier transforms in the other domain. This property alone makes the use of the Fourier transform an attractive alternative to computing the convolution integral, as seen when implementing various filters defined in terms of their frequency characteristics. By studying the Fourier transform of a few common signal types, and appreciating the properties relating the time- and frequency-domain representations of signals, the concepts behind the Fourier transform can readily be applied to many diverse and interesting situations.

## End-of-Chapter Exercises

**6.1** A voltage signal $a(t)$ is defined as:

$$a(t) = \frac{100 \sin(100\pi t)}{1000\pi t} V$$

The signal passes through a linear system with response function $b(t)$, where:

$$b(t) = \frac{500 \sin(500\pi t)}{500\pi t}$$

Let the resulting output signal be $c(t)$.

**(a)** Describe the output signal $c(t)$ in terms of $a(t)$ and $b(t)$.

**(b)** By inspection, sketch the Fourier transform $A(f)$.

**(c)** By inspection, sketch the transfer function $B(f)$.

**(d)** Describe the output signal $C(f)$ in terms of $A(f)$ and $B(f)$.

**(e)** Sketch the output signal $C(f)$ in the frequency domain.

**(f)** From $C(f)$, sketch the output signal $c(t)$.

**6.2** Consider the signals $\{a(t), b(t), c(t)\}$ defined in Question 6.1.

**(a)** Find the energy contained in the input signal $a(t)$.

**(b)** Find the energy contained in the output signal $c(t)$.

**6.3** Consider the system described by the impulse response $b(t)$ in Figure 6.30.

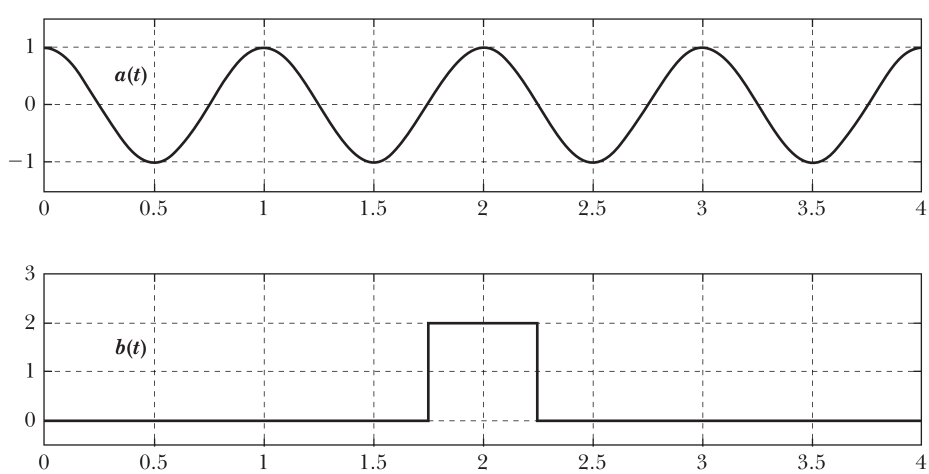



**FIGURE 6.30** Figure for Questions 6.3 to 6.5.

(a) Is this a causal system?
(b) Describe $b(t)$ in terms of an impulse and a unit rectangle.
(c) By inspection, give an exact expression for $B(f)$.
(d) Sketch the magnitude of the transfer function $|B(f)|$.
(e) What frequency input signal(s) would be exactly blocked by this system?

**6.4** Consider the system described by the impulse response $b(t)$ in Figure 6.30. Let the input to this system be the signal $a(t)$ as shown.

(a) Sketch the Fourier transform $A(f)$ as magnitude and phase.
(b) Sketch the magnitude of the transfer function $|B(f)|$ on the same graph as $A(f)$.
(c) Sketch the phase response of the transfer function.
(d) From the graph, determine the product $C(f) = A(f)\,B(f)$.
(e) Sketch $a(t)$ and $c(t)$ on the same graph.
(f) Give an exact expression for $c(t)$.

**6.5** Consider the system described by the impulse response $b(t)$ in Figure 6.30. Let the input to this system be the signal $a(t)$ as shown.

(a) Find the smallest value of $k$ where $(k > 0)$ such that $b(kt)$ would completely block the signal $a(t)$.
(b) Sketch the new impulse response $b(kt)$.

**6.6** Refer to Figure 6.31.

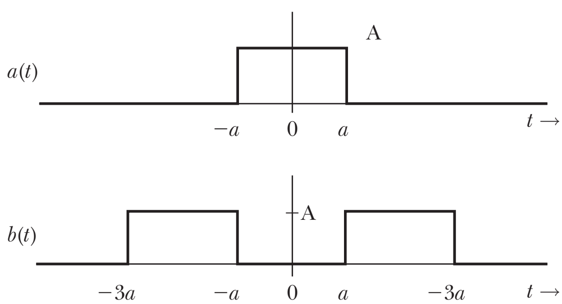



**FIGURE 6.31** Diagram for Question 6.6.

(a) Consider the Fourier transform $B(f)$ of signal $b(t)$.
  i. What is the value of $B(0)$?
  ii. Describe $B(f)$ as odd or even, real or imaginary.
(b) Express $b(t)$ as the convolution of $a(t)$ with a pair of impulses.
(c) From the answer for part (b), give an exact simplified expression for $B(f)$.
(d) Compute the Fourier transform $B(f)$ and compare to the answer for part (c).

**6.7** Refer to Figure 6.32 and determine the Fourier transform of this signal by inspection. *Hint: Express this signal in terms of the unit rectangle.*

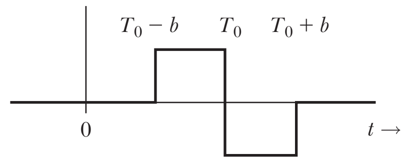



**FIGURE 6.32** Diagram for Question 6.7.

**6.8** (a) Accurately sketch $s(t) = 20\,\text{sinc}(2t)$ in the time domain.
(b) *By inspection,* accurately sketch the corresponding Fourier transform $S(f)$.
(c) From the sketch, give an expression for $S(f)$.

**6.9** A signal $s(t)$ is formed by the *convolution* of the two signals $a(t) = 100\,\text{sinc}(2t)$ and $b(t) = \text{rect}(2t)$. It is not necessary to sketch $s(t)$.

(a) *By inspection,* accurately sketch the corresponding Fourier transform $S(f)$.
(b) From the sketch of $S(f)$, give an exact expression for the Fourier transform $S(f)$.
(c) From the sketch of $S(f)$, determine:

$$\int_{-\infty}^{+\infty} s(t)\, dt$$

**6.10** The signals $a(t) = \cos(2\pi t)$ and $b(t) = \cos(20\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

(a) Sketch the Fourier transforms $A(f)$ and $B(f)$.
(b) Express $S(f)$ in terms of $A(f)$ and $B(f)$.
(c) What should be the area under $S(f)$?
(d) Describe $S(f)$ as odd or even, real or imaginary.
(e) Sketch the Fourier transform $S(f)$.

**6.11** The signals $a(t) = \sin(2\pi t)$ and $b(t) = \cos(20\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

(a) Sketch the Fourier transforms $A(f)$ and $B(f)$.
(b) Express $S(f)$ in terms of $A(f)$ and $B(f)$.
(c) What should be the area under $S(f)$?
(d) Describe $S(f)$ as odd or even, real or imaginary.
(e) Sketch the Fourier transform $S(f)$.

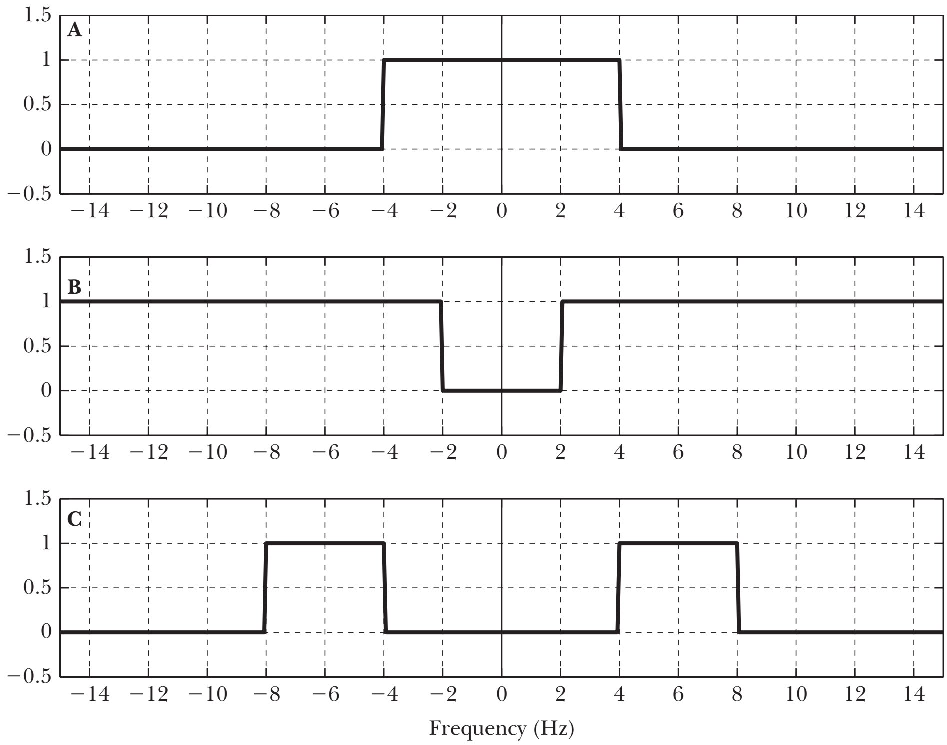



**FIGURE 6.33** Figure for Questions 6.13 to 6.18.

**6.12** The signals $a(t) = \text{rect}(t)$ and $b(t) = \cos(200\pi t)$ are multiplied to give the signal $s(t) = a(t)b(t)$.

**(a)** Sketch the Fourier transforms $A(f)$ and $B(f)$.
**(b)** Express $S(f)$ in terms of $A(f)$ and $B(f)$.
**(c)** What should be the area under $S(f)$?
**(d)** Describe $S(f)$ as odd or even, real or imaginary.
**(e)** Sketch the Fourier transform $S(f)$.

**6.13** Describe each of the ideal frequency-domain filters in Figure 6.33 as highpass, lowpass, or bandpass.

**6.14** Derive the equivalent time-domain (convolutional) filter for the ideal filter shown in Figure 6.33A.

**6.15** Derive the equivalent time-domain (convolutional) filter for the ideal filter shown in Figure 6.33B.

**6.16** Derive the equivalent time-domain (convolutional) filter for the ideal filter shown in Figure 6.33C.

**6.17** A ±2 V even square wave with a period of 1 s passes through the filter of Figure 6.33A. Sketch the output signal.

**6.18** A ±2 V even square wave with a period of 500 ms passes through the filter of Figure 6.33C. Sketch the output signal.

**6.19** Consider the voltage signal $s(t)$ from Figure 6.34.

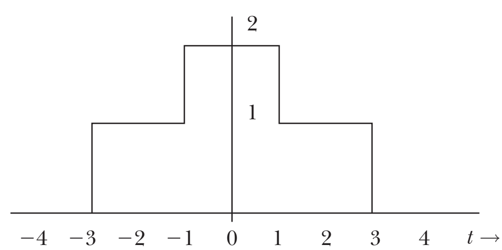



**FIGURE 6.34** Diagram for Question 6.19.

A periodic signal $c(t)$ is created by repeating $s(t)$ every 8 s.

**(a)** Sketch $c(t)$.
**(b)** Determine the $V_{rms}$ value of $c(t)$.
**(c)** Determine the $V_{DC}$ value of $c(t)$.
**(d)** Describe in words the overall appearance of $C(f)$.
**(e)** Derive the value of the Fourier series component at 8 Hz.
**(f)** Use the results for part (e) to determine the percentage of total power found in components beyond 8 Hz.

**6.20** Consider the signal $w(t) = \begin{cases} e^{-at} & t \geq 0 \\ 0 & otherwise \end{cases}$

Let $a = 2\pi$ and find what percentage of the total energy in $w(t)$ is found in frequencies less than 1 Hz.

**6.21** Refer to the signal $a(t)$ in Figure 6.35.

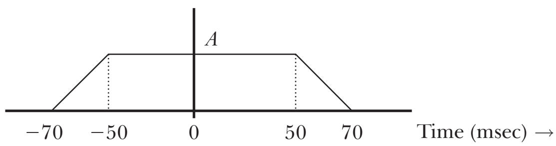



**FIGURE 6.35** Diagram for Question 6.21.

**(a)** Without computing $A(f)$, complete the following table concerning $a(t)$ and the corresponding Fourier transform $A(f)$.

**(b)** Compute $A(f)$ and confirm that this result matches the predictions from part (a).

| | |
|---|---|
| **Time Domain** $a(t)$ | ———————— |
| Power (W) | |
| Energy | |
| Describe as odd/even, real/imaginary | |
| **Frequency Domain** $A(f)$ | ———————— |
| Value at the origin $A(0)$ | |
| Area under $A(f)$ | |
| Energy | |
| Describe as odd/even, real/imaginary | |

**6.22** Consider the unit rectangle $a(t)$ convolved with a step function $u(t)$ as in Figure 6.36.

**(a)** By inspection, determine an expression for the Fourier transform of $b(t)$.

**(b)** Calculate $B(f)$ and check this result against the predictions from part (a).

**(c)** Observe that $b(t)$ can be viewed as the time integral of $a(t)$ and find the Fourier transform of the integral by using $A(f)$ and the time integral theorem.

**(d)** Compare the two answers from parts (a) and (b).

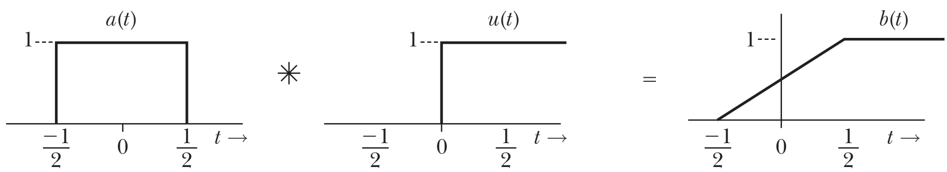



**FIGURE 6.36** Figure for Question 6.22.

**6.23** A radar system is to be built that sends a short electromagnetic pulse (a 2.0 GHz cosine sent for 1 *u*sec) that would normally return after reflecting from some distant object. The time $T$ between the outgoing pulse and its reflected copy is a function of distance to the object and the speed of light. The receiver sees the original pulsed signal and the reflected copy $T$ seconds later (the copies do not overlap). Sketch the signal at the receiver and estimate its Fourier transform $S(f)$ as a function of $T$.

**6.24** The sinc function of Figure 6.37 was generated in MATLAB from a unit height rectangle $N$ points wide defined on 1024 elements.

**(a)** Find the value of $N$.

**(b)** Use MATLAB to re-create the signal in the figure.

**(c)** Use MATLAB to plot the corresponding frequency domain rectangle.

**6.25** The sinc function of Figure 6.38 has been truncated to a width of 512 points through multiplication by a wide rectangle. The signal is defined on 1024 points.

**(a)** Describe the frequency-domain signal corresponding to this truncated sinc.

**(b)** Use MATLAB to re-create the signal in the figure.

**(c)** Use MATLAB to plot the corresponding frequency-domain rectangle.

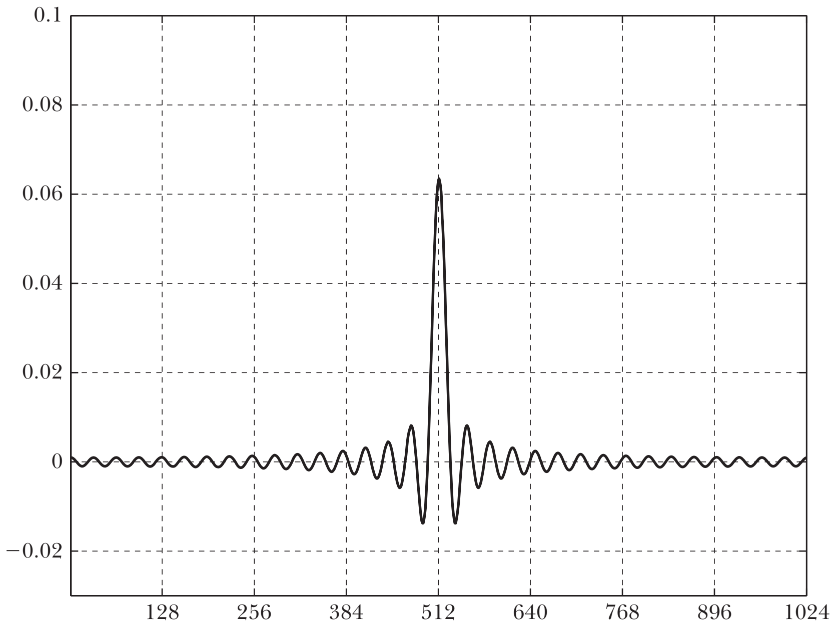



**FIGURE 6.37** Figure for Question 6.24.

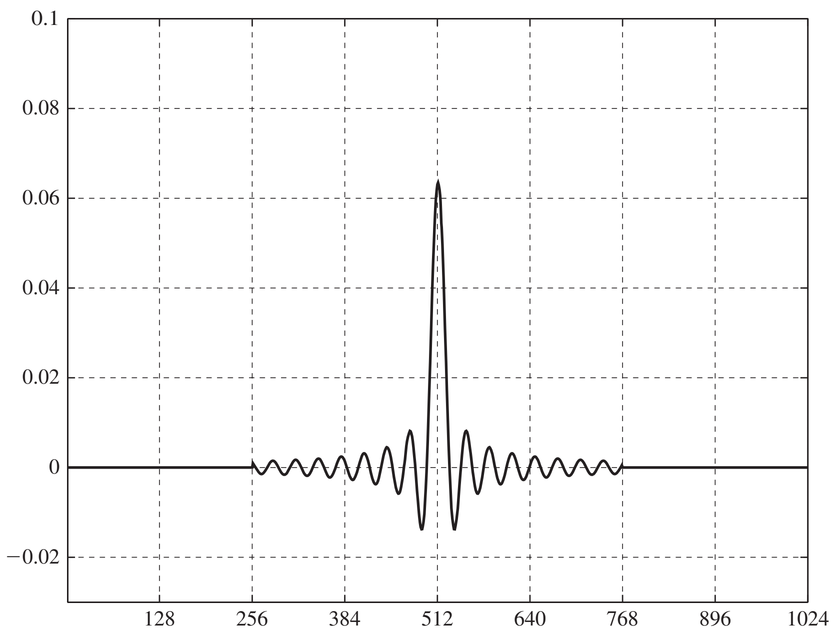



**FIGURE 6.38** Figure for Question 6.25.

CHAPTER 7

# The Laplace Transform

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Explain the relationship between the Laplace transform and the Fourier transform
- List the properties of the Laplace transform
- Compute the Laplace transform of common functions
- Identify the region of convergence
- Use the Laplace transform to analyze simple network circuits
- Derive a pole-zero diagram from a transfer function
- Interpret the significance of a pole-zero diagram
- Create a Bode plot by hand to estimate frequency and phase response curves
- Apply the Laplace transform to the design of filters with predictable properties

The Fourier transform has been studied as a way to represent a signal $s(t)$ as a linear decomposition of orthogonal sinusoidal components. By transforming a time-domain signal $s(t)$, calculations are often simpler, and the new perspective of a different domain can give a better understanding of signals behavior. In many applications, a different transform technique is required, one that goes beyond sinusoidal components and incorporates the ability to manipulate the derivative and integral response functions that are typically found in real systems. The *Laplace transform* is sometimes described as a *general-purpose Fourier transform,* although the two transforms usually serve in different applications.

While the Fourier transform relates the time and frequency domains, the Laplace transform adds an exponential dimension to the frequency domain and calls this the *s-domain*. Consequently, the Laplace transform of a time-domain signal $x(t)$ is the complex function of $X(s)$ having both frequency and exponential components. The $s$-domain function $X(s)$ describes a surface in two dimensions in which the Fourier transform can be identified as the frequency component taken alone. The Laplace transform offers an important new perspective on signals behavior and in many applications leads to a more concise and more descriptive approach to analyzing the interaction of signals and systems.

In this chapter, the *Laplace transform* is introduced, which incorporates the Fourier transform and shares many of its properties.

# 7.1 Introduction

The behavior of a linear system is characterized by its time-domain impulse response function $h(t)$. Consequently, the design and analysis of systems can focus on a study of the response function, either directly as $h(t)$ in the time domain or, as is often the case, in the alternate perspective of a *transform domain*. The Fourier transform $H(f)$ is used to examine a system from a frequency perspective. In many applications (such as filtering), the system characteristics are defined in terms of frequency behavior, and it is logical to study those systems from the frequency domain. In other systems, different characteristics such as transient and dynamic behavior may be important, and the use of another transform domain may be more appropriate. In particular, the Laplace transform is well suited to modelling linear systems that include differential or integral components in their response function. In practice, this includes a wide range of applications, and the Laplace transform is widely used in many different areas of engineering.

For electrical engineers, the Laplace transform is often the transform of choice in circuit analysis, since capacitors and inductors can be treated as simple impedances in the transform domain. In this way, complex circuits are essentially reduced to impedance networks without the need to manipulate the differential equations that govern the voltage and current relationships describing individual components. The nature of the Laplace transform facilitates the design and study of analog filters including frequency and phase response, and the overall behavior of a system can be defined by and determined by inspection of the transform-domain representation of a response function. The dynamics of many mechanical systems are also well served by the Laplace transform where system models include masses, springs, and dampers rather than resistors, inductors, and capacitors.

# 7.2 The Laplace Transform

Discussion of the Laplace transform begins with the familiar Fourier transform, which relates a time-domain function $x(t)$ to a frequency-domain function $X(f)$ as:

$$X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}\, dt$$

The Fourier transform is often written as a function of omega ($\omega$) where $\omega = 2\pi f$ to give:[1]

$$X(\omega) = \int_{-\infty}^{+\infty} x(t)e^{-j\omega t}\, dt$$

As usual, depending on the nature of $x(t)$ this integral may not exist. For example, the function $x(t) = u(t) \cdot t$ does not have a Fourier transform because this $x(t)$ grows without limit as $t$ increases. By introducing a convergence term, this difficulty can be overcome in many practical cases, and the resulting Laplace transform becomes applicable to a wider range of functions. In turn, the choice of

[1] This change of variables brings an extra factor of $2\pi$ into many Fourier transform operations; however, the use of $\omega$ is preferred for the Laplace transform.

an additional term adds a second dimension to the frequency domain. Let this new real valued term be sigma ($\sigma$) such that:

$$X(\sigma, \omega) = \int_{-\infty}^{+\infty} [x(t)e^{-\sigma t}]e^{-j\omega t}\, dt$$

which can be identified as the Fourier transform of $x(t)e^{-\sigma t}$. When $\sigma = 0$, the extra term vanishes to leave the Fourier transform of $x(t)$. It can be seen that for $t > 0$ and $\sigma > 0$ the term $e^{-\sigma t} \to 0$ as $t \to \infty$ and this integral will now converge for signals such as $x(t) = u(t)t$. Next, rearrange the equation as:

$$X(\sigma, \omega) = \int_{-\infty}^{+\infty} x(t)e^{-(\sigma + j\omega)t}\, dt$$

Finally, let $s = \sigma + j\omega$, and the Laplace transform is revealed. The real valued pair $(\sigma, \omega)$ now defines the *s*-domain. The Laplace transform relates a time-domain function $x(t)$ to a transform domain function $X(s)$ as:

$$X(s) = \int_{-\infty}^{+\infty} x(t)e^{-st}\, dt \tag{7.1}$$

The value of $\sigma$ plays a determining role in the existence of the above integrals. Provided that a given Laplace integral exists when $\sigma = 0$, the result matches the Fourier transform at $\sigma = 0$.

The Laplace transform is commonly used to describe the behavior of physical systems for which the causal response function $h(t)$ is necessarily zero for all $t < 0$. In that case, the transform is equivalent to that of a function multiplied by a unit step function $u(t)$.

$$H(s) = \int_{-\infty}^{+\infty} u(t)\, h(t)e^{-st}\, dt$$

Consequently, the Laplace transform is often defined as a *one-sided* or *unilateral* form as:

$$H(s) = \int_{0}^{+\infty} h(t)e^{-st}\, dt \tag{7.2}$$

The Laplace transform relationship may be abbreviated as:

$$x(t) \overset{\mathcal{L}}{\longleftrightarrow} X(s)$$

where the script $\mathcal{L}$ distinguishes the Laplace transform.[2] Where there is no chance of confusion, the simple notation $x(t) \longleftrightarrow X(s)$ can be used. Note that this is not an equality; the double arrow indicates the unique relationship of a Laplace transform pair. The form

$$\mathcal{L}[x(t)] = X(s)$$

is also seen. In this chapter, the unilateral form of the Laplace transform will be assumed unless otherwise stated. Furthermore, the lower limit will be assumed to include the value zero; this limit is sometimes written as $0^-$ to reinforce this condition.

[2] The letters $\mathcal{F}$, $\mathcal{L}$, and $\mathcal{Z}$ will specify the Fourier, Laplace, and z-transforms, respectively.

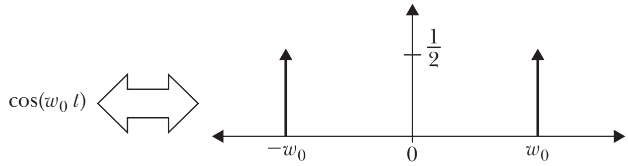



**FIGURE 7.1** **Sinusoidal** Fourier transform components of a cosine $s(t) = \cos(\omega_0\, t)$ where $\omega = 2\pi f$.

### 7.2.1 The Frequency Term $e^{j\omega t}$

From $s = \sigma + j\omega$, the $\omega$ term is a real value representing an imaginary exponential $e^{j\omega t}$. The presence of the frequency term $e^{j\omega t}$ in the Laplace transform incorporates the sinusoidal components associated with the Fourier transform as:

$$\cos(\omega t) = \frac{e^{+j\omega t} + e^{-j\omega t}}{2}$$

Sinusoidal components will always include a pair of terms in $\pm j\omega$ corresponding to the Fourier transform relationship, as seen in Figure 7.1. While the units of frequency are Hz, the units of $\omega = 2\pi f$ are rad/s.

### 7.2.2 The Exponential Term $e^{\sigma t}$

From $s = \sigma + j\omega$, the $\sigma$ term is a real value representing a real exponential $e^{\sigma t}$. The introduction of this exponential term has important consequences for the Laplace transform, both in where it is useful and how it can be calculated.

As seen in Figure 7.2 the exponential $e^{at}$ grows without limit if $a > 0$. Consequently, the specific form of $x(t)e^{-st}$ in the transform integral determines whether or not the Laplace transform integral exists. This issue of *convergence* is of foremost importance when the Laplace transform is computed and employed as there are generally regions of $s$ for which $X(s)$ does not exist.

### 7.2.3 The *s*-Domain

From $s = \sigma + j\omega$, the exponential $e^{st}$ incorporates both sinusoidal and exponential components in one unified complex valued representation called the *s*-domain. Because the variable $s$ is generally used without reference to the underlying $(\sigma, j\omega)$ real and imaginary components, it is especially important to understand the significance of the *s*-domain representation. In the following section the *s*-domain is explored in the context of some basic Laplace transforms and practical exercises.

## 7.3 Exploring the *s*-Domain

The Laplace transform of a signal $x(t)$ is the *s*-domain function $X(s)$. Each point $X(\sigma, \omega)$ represents a complex value on a two-dimensional surface in $e^{st}$, where $s = \sigma + j\omega$. In contrast, the Fourier transform $S(f)$ of the function $s(t)$ maps to a line called the frequency domain where each point $S(f)$ has complex components in $e^{j2\pi ft}$. The Fourier transform frequency domain can be identified in and

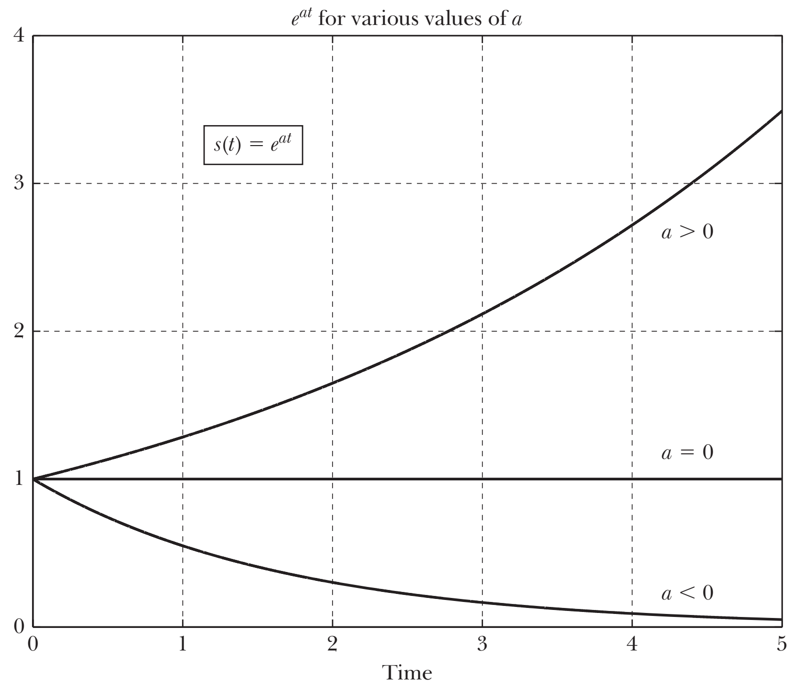



**FIGURE 7.2 Exponential** The exponential function $e^{at}$ will not converge for values of $a$ greater than 0.

calculated from the Laplace transform by setting $\sigma = 0$, whereby the Fourier transform $S(f)$ corresponds to one line on the Laplace transform surface $X(0, j\omega)$. To appreciate the utility of the *s*-domain representation, several examples are presented in the this section.

## 7.3.1 A Pole at the Origin

One particular Laplace transform serves to introduce the *s*-domain and to relate it to the corresponding time-domain functions. This is the Laplace transform of a constant value $h(t) = 1$, either as a step function $u(t)$ valid for all $t > 0$ or as the one-sided Laplace transform:

$$\mathcal{L}[u(t)] = H(s) = \int_{-\infty}^{+\infty} u(t)e^{-st}\,dt$$

$$H(s) = \int_{0}^{+\infty} 1e^{-st}\,dt \tag{7.3}$$

$$= \frac{-1}{s}e^{-st}\Big|_{t=0}^{\infty} \tag{7.4}$$

$$= \frac{-1}{s}e^{-\sigma t}e^{-j\omega t}\Big|_{t=0}^{\infty} \tag{7.5}$$

For the upper limit $(t \to \infty)$ the exponential term $e^{-\sigma t}$ converges to zero only for values of $\sigma > 0$. This observation defines the *region of convergence* (ROC) for which

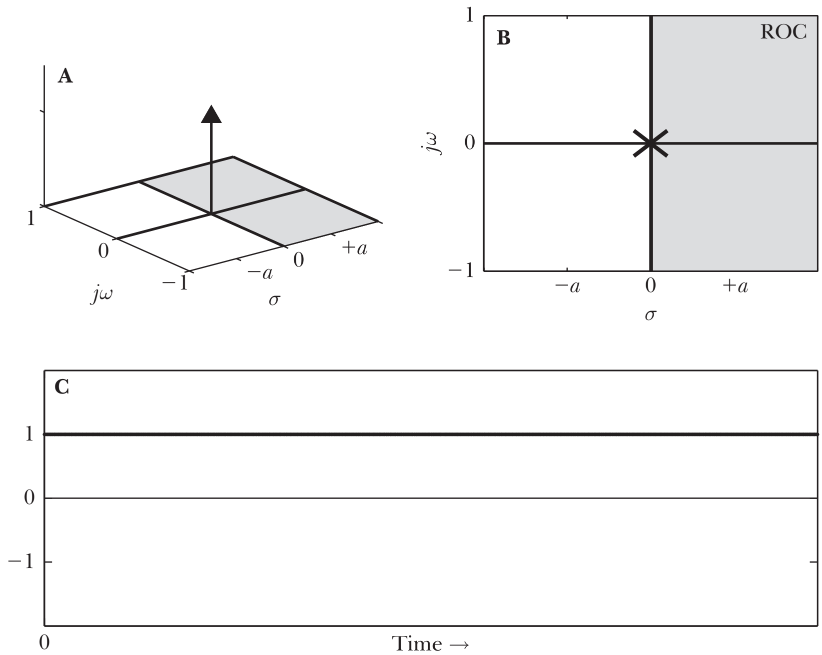



**FIGURE 7.3 The *s*-Plane: ($\sigma = 0, \omega = 0$)** The function $H(s) = 1/s$ characterizes the Laplace transform of a time-domain step function $u(t)$, which is a constant for all $t > 0$, and defines a *pole* having an infinite value when $s = 0$. The sketch in **A** shows the location of this pole at the origin in the *s*-domain. **B** shows the pole position as an $X$ on a pole-zero diagram. **C** shows the corresponding time-domain waveform equal to 1 for all $t > 0$.

this Laplace transform exists. On the $(\sigma, \omega)$ plane, the region includes all the values to the right of the vertical line $\sigma = 0$ as shown shaded in Figure 7.3. With this constraint,

$$H(s) = \frac{1}{s} \tag{7.6}$$

Significantly, the function $H(s) = 1/s$ goes to infinity when the denominator is zero. Since $s = \sigma + j\omega$, the denominator is zero only when both $\sigma = 0$ and $\omega = 0$, at the origin $(0, 0)$ of the *s*-domain. Such a point of infinite value is called a *pole*, while a point of zero value would be called a *zero*. There are no zeros in this example as the function $H(s) = 1/s$ has a constant non-zero numerator. In Figure 7.3A, a pole is sketched at position $(\sigma, \omega) = (0, 0)$ on the *s*-domain. The same information is shown in Figure 7.3B as a standard *pole-zero diagram*, where the pole position is marked with a cross (X). Any zeros would similarly be indicated with a circle (O). Establishing the pole-zero diagram is a key step in system analysis, as the positions of the poles and zeros define the Laplace transform surface for all $(\sigma, \omega)$. In practice, knowing the positions of the poles and zeros of the Laplace transform $H(s)$ is sufficient to describe the overall behavior of a linear system with response function $h(t)$.

At points other than the origin, the function $1/s$ falls off rapidly and symmetrically around the pole position and would never reach zero anywhere on the entire *s*-domain. With $s$ expressed in complex form:

$$H(\sigma, \omega) = \frac{1}{\sigma + j\omega}$$

it can be seen that as $\sigma \to 0$ the result matches the Fourier transform of the unit step $u(t)$ given by:

$$u(t) \stackrel{\mathcal{F}}{\longleftrightarrow} \frac{1}{j2\pi f}$$

Moreover, the pole represented by $H(s)$ can be shifted a distance $a$ to the left (for $a > 0$) along the horizontal axis as:

$$H(\sigma, \omega) = \frac{1}{(\sigma + a) + j\omega}$$

or, more conventionally as

$$H(s) = \frac{1}{s + a}$$

which is the subject of the following section. The region of convergence of this new pole position also shifts left to become $\sigma > -a$. In general, the ROC can never include a pole. This example demonstrates that for the unilateral Laplace transform the region of convergence will always be found as shown to the right of the right-most pole.

Finally it may be observed that, like the Fourier transform, the Laplace transform is linear such that $\mathcal{L}[ku(t)] = kH(s)$ for a constant $k$ and for $\mathcal{L}[ku(t)]$ gives:

$$H(s) = k\left[\frac{1}{s + a}\right]$$

which has the same pole position with or without the constant $k$. Consequently, the presence of an $s$-domain pole reflects a specific time-domain function that may differ by a constant $k$.

**Graphing the Function $H(s) = 1/s$** Generally, graphs that represent $f(x) = x^N$ appear as straight lines when shown on a logarithmic scale. Figure 7.4 shows the result of plotting both $f(x) = 1/x$ and $f(x) = 1/x^2$ on the same log-log graph. Both functions appear as straight lines having a slope that varies with the power ($N$) of $x^N$. The log-log format characterizes the *Bode plot* used to sketch transfer functions that typically vary as $s^N$ or as the corresponding frequency terms in $\omega^N$, such as the $H(s) = 1/s$ above.

MATLAB includes the function `loglog` to create a log-log plot.

```
x = 1 : 0.1 : 100;  %create x
y = 1./x;           %create y = f(x)
loglog(x, y);       %produce log-log plot of f(x) = 1/x
grid on
```

## 7.3.2 Decaying Exponential

Consider the Laplace transform $H(s)$ from above expressed as:

$$H(s) = \frac{1}{s + a} \tag{7.7}$$

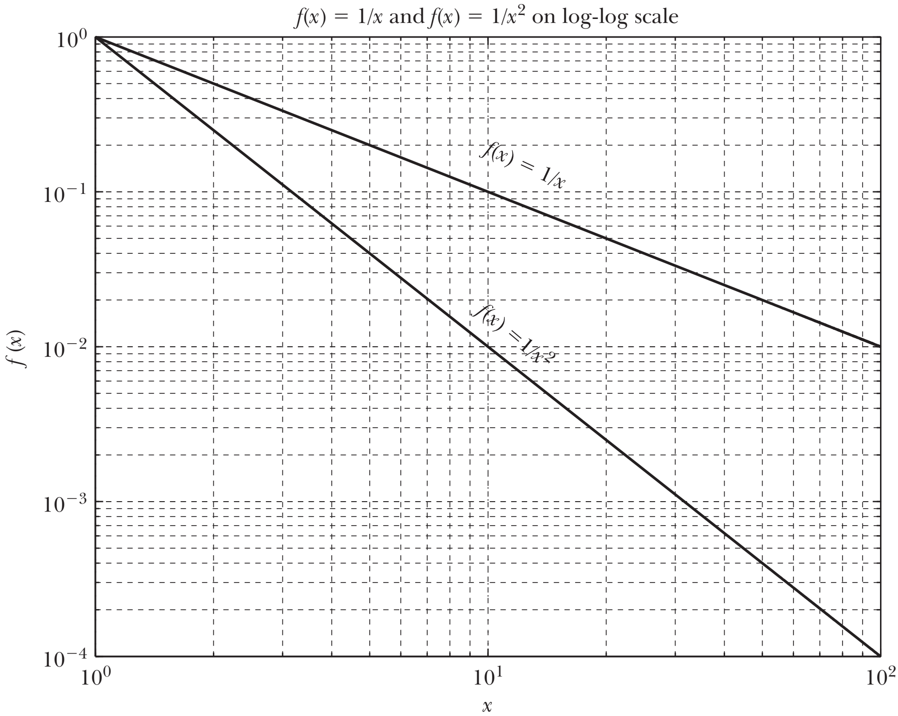



**FIGURE 7.4 A log-log Graph: f(x) = 1/x** When shown on a log-log scale, the function $f(x) = 1/x$ appears as a straight line. Similarly, the function $f(x) = 1/x^2$ appears as a straight line with twice the apparent downward slope.

By inspection, this $H(s)$ goes to infinity when the denominator goes to zero, which defines a pole located at $(\sigma, j\omega) = (-a, 0)$. From the definition of the Laplace transform, with $\sigma < 0$ and $\omega = 0$, this pole represents the time-domain exponential function $u(t)e^{-at}$ that is decaying with no frequency terms (since $\omega = 0$). This situation is shown in Figure 7.5B, where the pole-zero diagram shows a single pole to the left of the origin and directly on the $(\omega = 0)$ axis. The corresponding time-domain function is shown in Figure 7.5C.

Varying the pole position along the horizontal axis by varying $\sigma$ serves to change the rate of descent of the time-domain exponential function $h(t) = e^{\sigma t}$. As the pole is moved closer to the origin, the downward trend flattens out to become the unit step $u(t)$ as $\sigma \to 0$. However, if this pole is moved across to the right-hand side of the origin where $\sigma > 0$, the corresponding exponential grows without limit, and the response function $h(t)$ is no longer useable. In that case, a simple impulse input to $h(t)$ would lead to an *unstable* system. In general, poles will be expected only on the left-hand side of the pole-zero diagram.

In a more general sense, the emergence of the time-domain exponential term $e^{at}$ relates directly to the horizontal shift in the *s*-domain. Conversely, a shift in time would multiply the corresponding Laplace transform by an exponential term. As a Laplace transform property, if $x(t) \xleftrightarrow{\mathcal{L}} X(s)$, then:

$$e^{at}x(t) \xleftrightarrow{\mathcal{L}} X(s-a)$$

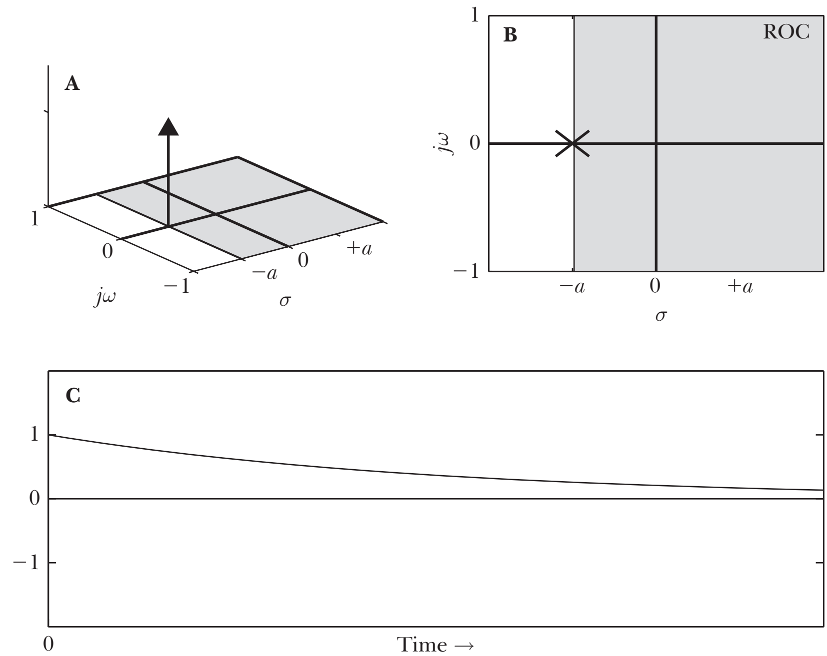



**FIGURE 7.5** **The *s*-Plane: ($\omega = 0$)** Values of $\sigma$ from the exponential function $e^{\sigma t}$ are represented along the $\sigma$-axis. In this figure, a *pole* is located at the position $(\sigma, j\omega) = (-a, 0)$. **A** is a three-dimensional view of the impulse showing its location on the *s*-plane. **B** shows the same component on a pole-zero diagram, where X marks the pole position. **C** shows the corresponding time-domain waveform, where only an exponential term is present whenever $\omega = 0$.

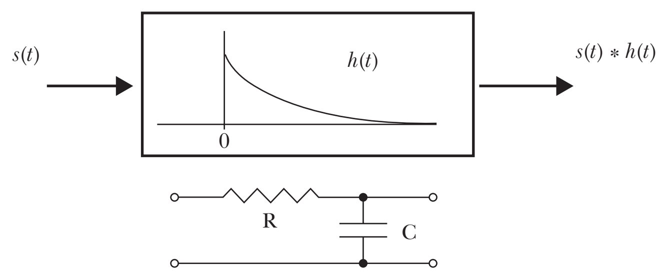



**FIGURE 7.6** An RC lowpass filter. The exponential response function $h(t)$ can be identified with the single pole of Figure 7.5.

Consider the causal impulse response $h(t)$ of the lowpass filter RC circuit in Figure 7.6. Because the form $h(t)$ is an exponentially decaying curve, this response function can be identified directly in the pole-zero diagram of Figure 7.5B where the exponential term $e^{\sigma t}$ is associated with the pole at $(\sigma, 0)$. It will be shown in Section 7.8 that a full analysis of the circuit using the Laplace transform reveals this exact response function for which $\sigma = -1/RC$. Otherwise, direct calculation in the time domain depends on the fact that current $i(t)$ flowing through a capacitor is related to the voltage $v(t)$ as the time derivative $i(t) = C\frac{dv(t)}{dt}$. Use of the Laplace transform in circuit analysis will avoid the need to solve differential equations.

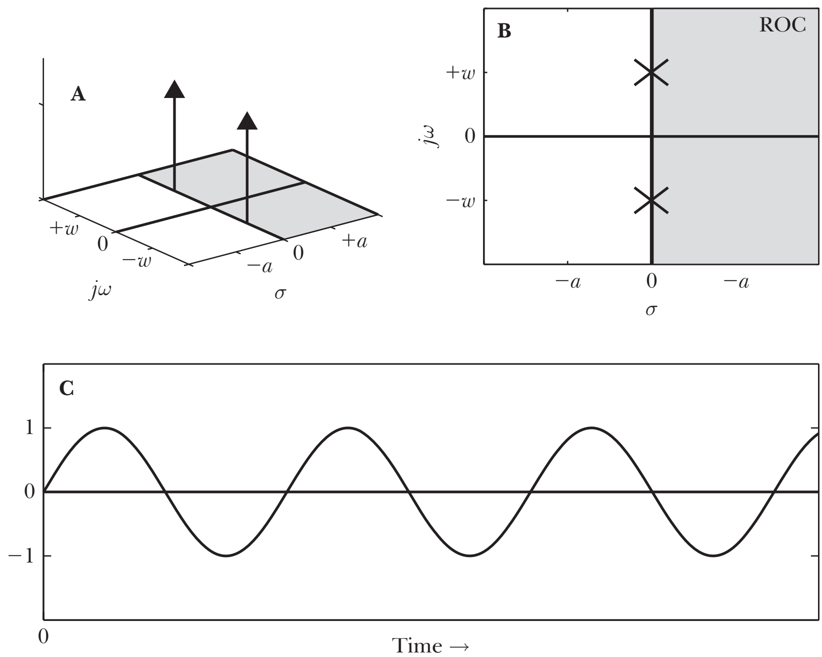



**FIGURE 7.7 The s-Plane: (σ = 0)** The exponential function $e^{\sigma t}$ represented along the $\sigma$-axis is multiplied by the complex exponential components shown on the $j\omega$-axis. **A** is a three-dimensional view of the familiar Fourier transform components of a sinusoid displayed with $\sigma = 0$. **B** shows the same components on a pole-zero diagram. **C** shows the corresponding time-domain waveform. The Laplace transform is identical to the Fourier transform when $\sigma = 0$.

## 7.3.3 A Sinusoid

In Figure 7.7A, a Laplace transform is sketched as a pair of poles at position $(\sigma, \omega) = (0, -\omega_0)$ and $(0, +\omega_0)$ on the s-domain. This sketch brings to mind the Fourier transform of a sinusoid, and there is no exponential term since $\sigma = 0$ for both poles. The pole-zero diagram in Figure 7.7B shows a pair of poles directly on the $(\sigma = 0)$ axis corresponding to a sinusoid with frequency $\omega_0$. Figure 7.7C shows the corresponding time-domain waveform. Note that if this example represents the response function of a linear circuit, then a simple impulse input would result in a sinusoidal output oscillation that goes on forever; in general, poles will not be found exactly on the line $(\sigma = 0)$ but may lie anywhere to the left of this line.

Shifting a pole vertically along the $j\omega$ axis by a distance $\omega_0$ moves $H(s) \to H(s + j\omega_0)$. The presence of a pair of poles at $(s + j\omega_0)$ and $(s - j\omega_0)$ leads to the form $s^2 + \omega_0^2$ that neatly incorporates both poles. The poles in Figure 7.7 are of the form below, where the denominator goes to zero for two different values of $s$.

$$H(s) = \frac{1}{(s + j\omega_0)(s - j\omega_0)} = \frac{1}{s^2 + \omega_0^2} \tag{7.8}$$

This denominator is found in both the sine and cosine Laplace transforms.

**The Generalized Cosine:** $A = \cos(\omega t + \Phi)$ The generalized cosine $A = \cos(\omega t + \Phi)$ includes a phase component $\Phi$ that emerges as a *zero* on the horizontal axis in the pole-zero diagram. The specific form of this Laplace transform is:

$$\mathcal{L}\left[\cos(\omega_0 t + \Phi)\, u(t)\right] = \frac{s\cos(\Phi) - \omega_0 \sin(\Phi)}{s^2 + \omega_0^2} \tag{7.9}$$

where the distinctive denominator reflects the sinusoid and the numerator holds the phase component. Specifically, the numerator is zero when:

$$s\cos(\Phi) - \omega_0 \sin(\Phi) = 0$$

and the zero appears on the horizontal axis in the pole-zero diagram for values of $s$ that satisfy this equation. For $cos(\omega_0 t)$ when $\Phi = 0$, the zero is at the origin $s = 0$. The term in $s$ disappears, and there is no zero for $\sin(\pm\omega_0 t)$ when $\Phi = \pm\pi/2$. For all other values, the $\sigma$-axis position of the zero varies as $tan(\Phi)$:

$$s = \frac{\omega_0 \sin(\Phi)}{\cos \Phi} = \omega_0 \tan(\Phi)$$

## 7.3.4 A Decaying Sinusoid

In Figure 7.8A, a Laplace transform is sketched as a pair of poles at position $(\sigma, \omega) = (-a, -\omega_0)$ and $(-a, +\omega_0)$ on the $s$-domain. This sketch represents the Fourier transform of a sinusoid with a decreasing exponential term (since $\sigma < 0$). The pole-zero diagram in Figure 7.8B shows the pair of poles directly on the $(\sigma = -a)$ line corresponding to a cosine with frequency $\omega_0$ rad/s multiplied by an exponential

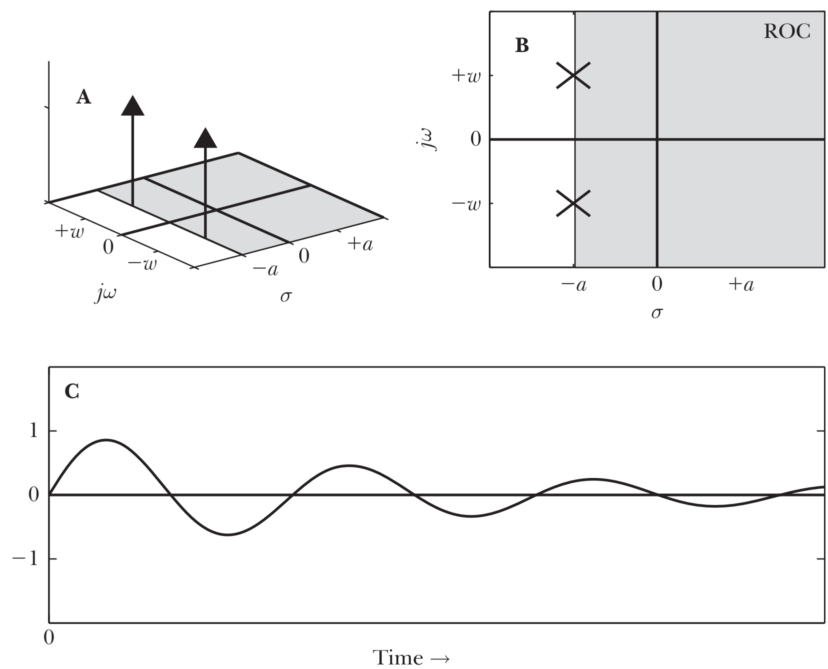



**FIGURE 7.8 The $s$-Plane: ($\sigma < 0$)** The exponential function $e^{\sigma t}$ represented along the $\sigma$-axis is multiplied by the complex exponential components shown on the $j\omega$-axis. **A** is a three-dimensional view of the familiar Fourier transform components of a sinusoid displayed with $\sigma < 0$. **B** shows the same components on a pole-zero diagram. **C** shows the corresponding time-domain waveform where the amplitude falls exponentially over time. See Figure 7.7.

decay. Figure 7.8C shows the corresponding time-domain waveform. Note that if this example represents the response function of a linear circuit, then a simple impulse input would result in a sinusoidal output oscillation that diminishes to zero after a time; this is a reasonable behavior that is consistent with the observation that poles are expected to be found in the left half plane of the pole-zero diagram.

Shifting a pole horizontally along the $\sigma$ axis by a distance $a$ moves $H(s) \to H(s+a)$. It follows that the shifted-paired poles in Figure 7.8 are of the form:

$$H(s) = \frac{1}{(s+a)^2 + \omega_0^2} \tag{7.10}$$

and the sinusoid is expected to decrease exponentially in amplitude over time for $a > 0$. Expanding the denominator gives:

$$H(s) = \frac{1}{s^2 + 2as + (a^2 + \omega_0^2)} \tag{7.11}$$

where in this second-degree polynomial, the term $2as$ totally depends on the presence of $a$ and reflects the rate of attenuation over time. The right-hand term $(a^2 + \omega_0^2)$ is recognized as the squared distance from each pole to the origin.

As in the case of a single pole shifted to the left or to the right, a paired pole located on the right-hand half plane corresponds to a sinusoid amplitude that is growing over time. This situation is examined in the next section.

### 7.3.5 An Unstable System

In Figure 7.9A, a Laplace transform is sketched as a pair of poles at position $(\sigma, \omega) = (+a, -\omega_0)$ and $(+a, +\omega_0)$ on the $s$-domain. This sketch evokes the Fourier transform of a sinusoid with an increasing exponential term (since $\sigma > 0$). The pole-zero diagram in Figure 7.9B shows a pair of poles directly on the $(\sigma = +a)$ line corresponding to a cosine with frequency $\omega_0$ multiplied by a growing exponential. Figure 7.9C shows the corresponding time-domain waveform. Note that if this example represents the response function of a linear circuit, then a simple impulse input would result in a sinusoidal output oscillation that unceasingly grows in amplitude; the presence of poles on the right half plane of the pole-zero diagram indicates instability. Significantly, the region of convergence (shaded) no longer includes the vertical axis $(\sigma = 0)$, and it is no longer valid to determine the Fourier transform from values found along this axis; in short, *the Fourier transform does not exist for this function.*

## 7.4 Visualizing the Laplace Transform

In applying the Laplace transform to the study of signals and systems, one of the principal goals is to find a system frequency response as defined by the magnitude of the Fourier transform. Since the Fourier transform is identified as the points on the vertical axis for $\sigma = 0$, it is important to appreciate how the pole and zeros of the Laplace transform surface serve to affect the form of the function of $\omega$ found along along that vertical axis. In this section, the effect of poles and zeros on the vertical axis is explored.

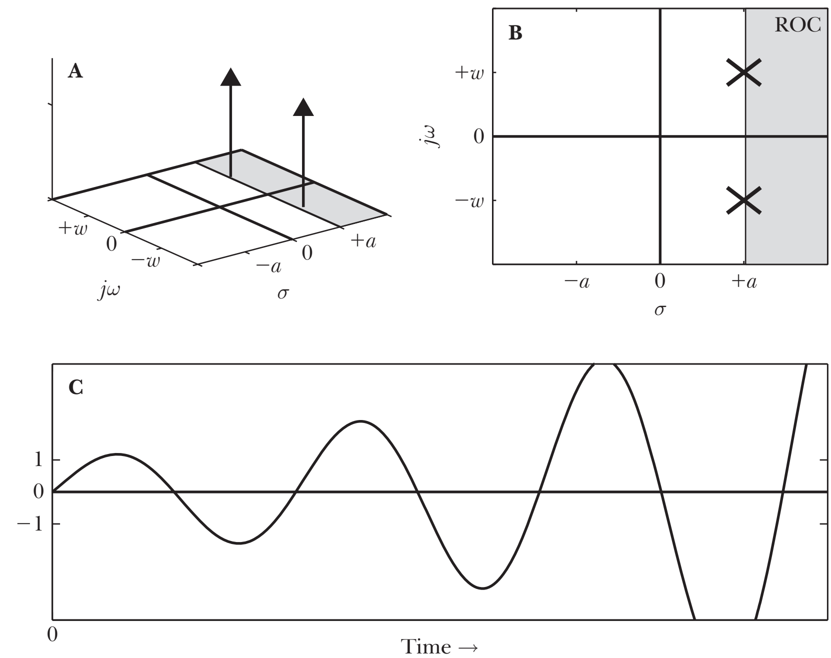



**FIGURE 7.9 The $s$-Plane: ($\sigma > 0$)** The exponential function $e^{\sigma t}$ represented along the $\sigma$-axis is multiplied by the complex exponential components shown on the $j\omega$-axis. **A** is a three-dimensional view of the familiar Fourier transform components of a sinusoid displayed with $\sigma > 0$. **B** shows the same components a pole-zero diagram. **C** shows the corresponding time-domain waveform where the amplitude grows exponentially over time. See Figure 7.7.

## 7.4.1 First-Order Lowpass Filter

Returning to the impulse response $h(t)$ of the lowpass filter RC circuit in Figure 7.6, the single pole is described by:

$$H(s) = k\left[\frac{1}{s+a}\right] \tag{7.12}$$

where $k = a$ for this particular circuit. The actual profile of the Laplace transform surface $H(s)$ depends directly upon the placement of the poles and zeros of the function $H(s)$. Figure 7.10 shows the Laplace transform $H(s)$ as a function of $(\sigma, j\omega)$ where the vertical axis is drawn on a log scale. Figure 7.10A shows the pole-zero diagram with a single pole in the left half plane; the region of convergence $(\sigma > -a)$ defined above is shaded. The pole position corresponds to an infinite value in the $H(s)$ function. Note that the line $\sigma = 0$ lies within the region of convergence, consequently the Fourier transform $H(j\omega)$ can be identified as Figure 7.10C.

Observe the overall appearance of the Fourier transform $H(j\omega)$ in Figure 7.10C. It is symmetrical around the origin $(\omega = 0)$, where it is at its highest value; otherwise, its amplitude diminishes with increasing frequency. This is consistent with the frequency response of the lowpass filter represented by this circuit. Next, observe how the pole position might affect the shape of $H(j\omega)$. If the pole were moved very far away along the $\sigma$-axis, then the surface near $H(j\omega)$ would flatten to leave little or no lowpass effect. Finally, if the pole were moved closer to the origin $\sigma = 0$, the surface

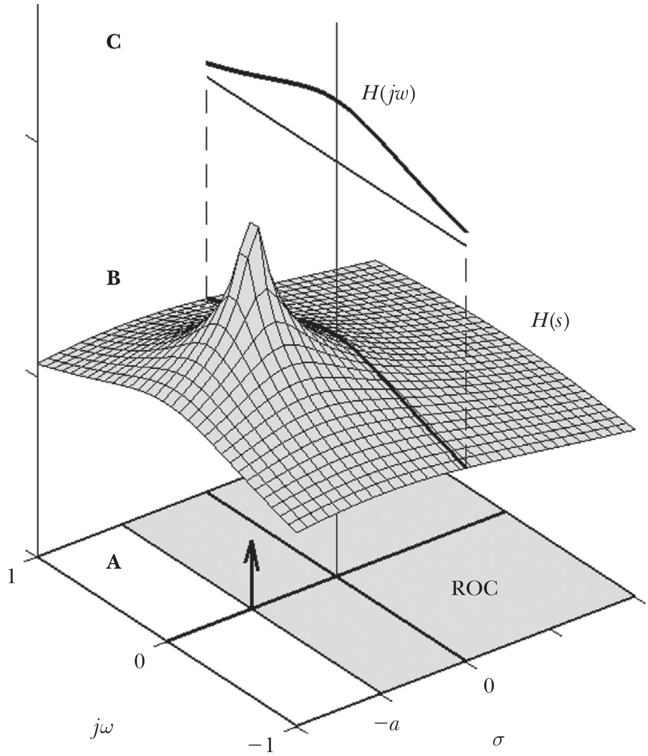



**FIGURE 7.10 The *s*-Plane—Laplace transform surface** The Laplace transform $H(s)$ of the lowpass filter from Figures 7.5 and 7.6 maps to a surface on the *s*-domain. Graph **A** shows a single pole located at $(\sigma, j\omega) = (-\text{a}, 0)$ and the region of convergence extending to the right of $\sigma = -a$. In graph **B**, the Laplace transform $H(s)$ rises to an infinite peak directly above the pole. Likewise, the Fourier transform $H(jw)$ is lifted from this surface in graph **C** along the line $\sigma = 0$ and serves to describe the *frequency response* of this lowpass filter.

would rise up, and $H(j\omega)$ would be much higher near $f = 0$ such that very low frequencies would be heavily favored over higher frequencies. In summary, the position of the pole affects the lowpass characteristics of this system by varying the cutoff frequency. It is this observation that makes the Laplace transform method such a valuable tool in system analysis; *knowing the position of this single pole is all that is necessary to describe completely the system response it represents.*

The corresponding Fourier transform $H(j\omega)$ is found by setting $\sigma = 0$ in Eqn. 7.12 (with $k = a$) to give:

$$H(j\omega) = \left[\frac{a}{j\omega + a}\right] \tag{7.13}$$

Or, with $\omega = 2\pi f$ this becomes:

$$H(f) = \left[\frac{a}{j2\pi f + a}\right] \tag{7.14}$$

In Section 6.7, this same result was found as Eqn. 6.2, working directly from the response function using the Fourier transform. In the present case, the frequency response has been found directly from the Laplace transform of that same impulse response.

## 7.4.2 Pole Position Determines Frequency Response

It has been shown that the single pole defined at $(\sigma = -a)$ defines a simple lowpass filter and that the pole position along the $\sigma$ axis affects the profile of the frequency response. This relationship will now be explored in detail.

Consider the single pole in Figure 7.11, which lies a distance $a$ from the origin. This pole corresponds to the transfer function:

$$H(s) = k\left[\frac{1}{s+a}\right]$$

As the value of points $H(s)$ around the pole fall off uniformly in all directions, every point at distance $d$ from the pole will have the same value as shown by the circle of radius $d$ on the figure. The distance from the pole to any point $\omega$ on the vertical axis $(\sigma = 0)$ has $d^2 = a^2 + \omega^2$. Points along the vertical axis define the Fourier transform $H(\omega)$, and the magnitude $|H(\omega)|$ is the *frequency response* of the time-domain system defined by $h(t)$. Any point on the $j\omega$ axis has a value given by:

$$H(j\omega) = \frac{k}{j\omega + a}$$

By rationalizing the denominator, the real and imaginary parts of $H(s)$ can be isolated.

$$H(\omega) = \left[\frac{k}{j\omega + a}\right]\left[\frac{-j\omega + a}{-j\omega + a}\right]$$

$$H(\omega) = \left[\frac{ka}{\omega^2 + a^2}\right] + j\left[\frac{-k\omega}{\omega^2 + a^2}\right]$$

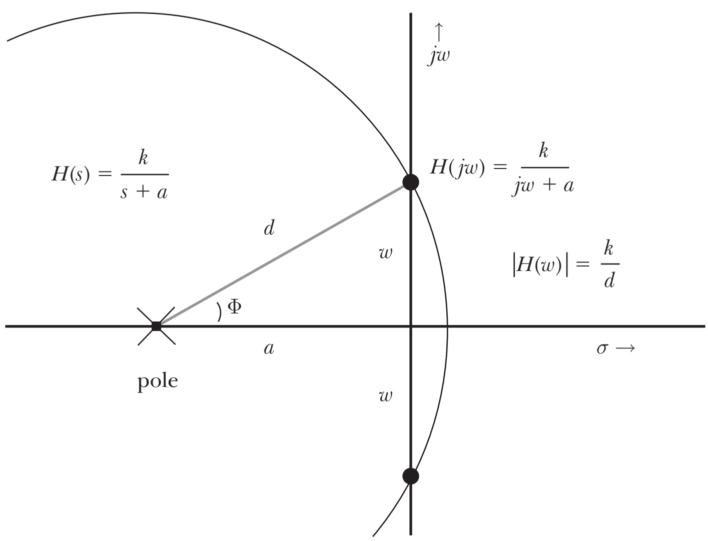



**FIGURE 7.11 The Role of Poles** The Laplace transform $H(s)$ of the lowpass filter from Figures 7.5 and 7.6 maps to a surface on the $s$-domain. The Fourier transform $H(jw)$ is identified along the vertical axis $(\sigma = 0)$ and serves to describe the *frequency response* of this lowpass filter. The single pole defines the form of the magnitude $|H(\omega)|$ simply as $k/d$ where $d$ is the distance from the pole to each point $\omega$.

The phase of $H(\omega)$ is the arc tangent of the $Im[H(\omega)]/Re[H(\omega)]$ or:

$$\Phi(\omega) = tan^{-1}\left(\frac{-\omega}{a}\right) \tag{7.15}$$

which is identified on Figure 7.11 as the angle between the horizontal and the vector leading from the pole to $H(j\omega)$.

The magnitude $|H(\omega)|$ is the square root of $H(\omega)H^*(\omega)$ or:

$$\begin{aligned}|H(\omega)| &= \left(\left[\frac{k}{j\omega + a}\right]\left[\frac{k}{-j\omega + a}\right]\right)^{\frac{1}{2}} \\ &= \left(\frac{k^2}{\omega^2 + a^2}\right)^{\frac{1}{2}} \\ &= \frac{k}{\sqrt{\omega^2 + a^2}}\end{aligned}$$

however, since $d^2 = \omega^2 + a^2$ by definition, then:

$$|H(\omega)| = \frac{k}{d}$$

Consequently, the magnitude $|H(\omega)|$ is a function of $d$ for a given pole position, where $d$ depends on $\omega$ and $a$. The corresponding phase angle is shown $\Phi$ that varies from 0 to $-90$ degrees as $\omega$ increases. It can also be seen that for every $H(\omega)$ there is a corresponding value $H(-\omega)$ where the vertical axis is the same distance $d$ from the pole.

1. For $\omega = 0$, at the origin, $d = a$ and $|H(0)| = k/a$. This fraction represents the *DC gain* of $H(s)$, and for $k = a$ the system has *unit gain.* The phase angle is $\Phi = 0$ *degrees*;
2. For $\omega = a$, then $d = \sqrt{2}a$ and $|H(a)| = \frac{k}{\sqrt{2}a} = 0.707$ times the DC value. The squared magnitude is $|H(a)|^2 = 0.5$ times the DC level, making this the *half-power point.* From Eqn. 7.15, the phase angle is $\Phi = -45$ *degrees*;
3. For $\omega \rightarrow \infty$, then $d \rightarrow \infty$ and $|H(\infty)| \rightarrow 0$ as $1/d$, $\Phi \rightarrow -90$ *degrees.*

The overall frequency response given by $|H(\omega)|$ varies with $\omega$ as a function of $d$ from a maximum of $k/a$ at the origin to approaching zero for very high frequencies. The shape of this curve is shown in Figure 7.15 for $k = a$ where the three reference points are labelled. This is a lowpass filter that favors lower frequencies and attenuates higher frequency components. The lower graph shows these same components on a log-log scale where the function $1/s$ approximates straight lines at the extremes. This form of the frequency-response graph is called a *Bode plot*[3] where the point ($\omega = a$) marks a turning point in both graphs and is called a *breakpoint.*

In summary, for

$$H(s) = k\left[\frac{1}{s + a}\right]$$

1. There is a single pole located at $s = -a$, where the denominator is zero.
2. The region of convergence lies to the right of the line defined by $\sigma = a$.

[3] Pronounced *bo ·dee.*

3. The corresponding transfer function $|H(w)|$ describes a lowpass filter with DC gain $= k/a$ and a breakpoint corresponding to the *half power point* of the circuit response.
4. The *breakpoint* is identified as the distance from the pole to the origin; here, $\omega_B = a$.
5. The transfer function includes a phase component $\Phi(\omega)$ that varies uniformly with frequency from 0 to $-90$ degrees and is $-45$ degrees at the breakpoint.
6. The pole corresponds to the impulse response $h(t) = u(t)ke^{-at}$; this is equivalent to taking the *inverse Laplace transform* of H(s).

## 7.4.3 Second-Order Lowpass Filter

Consider the paired poles in Figure 7.12, which lie at a distance $a$ from the vertical axis and a distance $\omega_0$ on either side of the horizontal axis. The presence of two poles defines this as a *second-order* system. These poles correspond to the transfer function:

$$H(s) = k\left[\frac{1}{(s+a)^2 + \omega_0^2}\right]$$

which by inspection corresponds to the Laplace transform of $h(t) = e^{-at}\sin(\omega_0 t)$ or a sinusoid of frequency $\omega_0$ rad/s with decreasing amplitude depending on the constant $a$.

Figure 7.12A shows the two poles and their mutual effect on a point $H(w)$ on the vertical ($\sigma = 0$) axis. The form of $H(\omega)$ is to be studied as $\omega$ is varied from 0 to $+\infty$. Let the poles be a distance $d_1$ and $d_2$ from $H(\omega)$. Just as the effect of a single pole at a distance $d$ varies as $1/d$, the effect of two poles at distances $d_1$ and $d_2$ varies as $1/d_1 d_2$. For large $\omega$, the value of $1/d_1 d_2$ falls off as $d^2$ as compared to the single pole case, so the Bode plot would have a line with double the downward slope as $\omega \rightarrow \infty$. In particular:

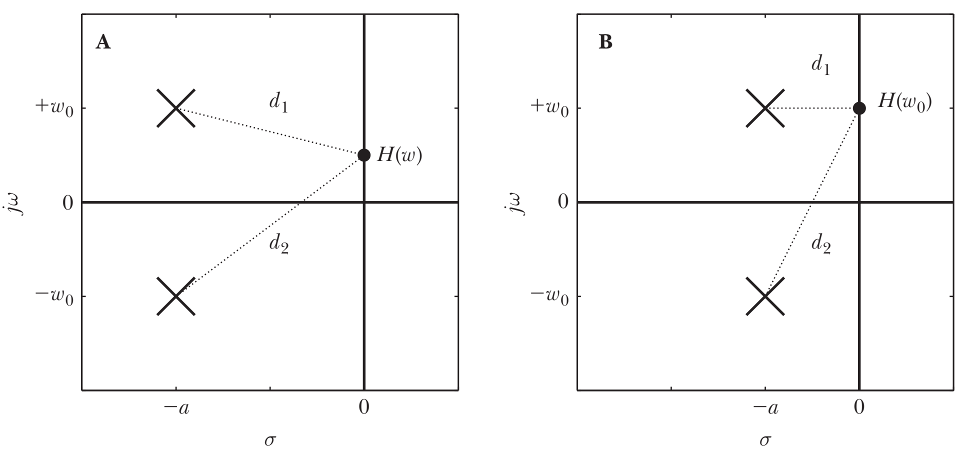



**FIGURE 7.12 Multiple Poles** The effect of multiple poles is determined by the distances from each to a specific point $H(s)$. In **A**, a pair of poles lie at distances $d_1$ and $d_2$ from the point $H(w)$ on the $\sigma = 0$ axis. **B** shows how the proximity of a pole to the vertical axis can dramatically affect the value of $H(\omega)$ as $d_1$ as $\omega$ approaches $\omega_0$. See Figure 7.13.

**1.** For $\omega = 0$, at the origin, $d_1$ and $d_2$ are equal, and $|H(0)| = k/d_1 d_2$. This fraction represents the *DC gain* of $H(s)$, and for $k = a^2 + \omega_0^2$ the system has *unit gain.*

As with the first-order filters discussed above, the breakpoint frequency corresponds to the distance from a pole to the origin. Each of these two poles lies the same distance from the origin, leading to a double breakpoint at:

$$\omega_B = \sqrt{a^2 + \omega_0^2}$$

corresponding to the radius of a circle centered at the origin and passing through both poles. Any poles lying on this circle would share the same breakpoint. Note that $\omega_B$ is larger than $\omega_0$ by definition and $\omega_B$ approaches $\omega_0$ as $a$ becomes small. The specific form of $H(\omega)$ near $\omega_0$ depends on both $\omega$ and $a$.

**a.** If $a$ is relatively large $(a \gg \omega_0)$, then $\omega_B \to a$. The two poles are far from the origin and appear as a double pole at the position $\sigma = a$, and the resulting $H(s)$ resembles the previous example with a value $-6$ *dB* at the breakpoint.

**b.** If $a$ is relatively small $(a \ll \omega_0)$, then $\omega_B \to \omega_0$. As seen in Figure 7.12B, near $\omega = \omega_0$ the value $d_1$ and the denominator $d_1 d_2$ may approach zero if the pole is very close to the vertical axis; this would create a potentially large excursion in the value of $H(\omega_0)$, as shown in Figure 7.13. This peak value is identified as a *resonance*; let this peak position be at frequency $\omega_P$.

Figure 7.12B shows that while $d_1$ is a minimum for $\omega = \omega_0$, the value of $d_2$ is growing as $\omega$ increases past $\omega_0$. Therefore, the peak in $|H(\omega)|$ is expected at a frequency $\omega_P$ that is somewhat *less than* $\omega_0$. It is left as an exercise to show that the maximum value of $1/d_1 d_2$ occurs for:

$$\omega_P = \sqrt{\omega_0 - a^2}$$

which is less than $\omega_0$ and approaches $\omega_0$ for small $a$. This expression is valid only for $a < \omega_0$; there is no peak when the poles lie at a distance greater than $\omega_0$ from the vertical axis. Similarly, the magnitude of the peak is most pronounced when $a$ is small and $\omega_P$ lies close to $\omega_0$, as in Figure 7.13.

**2.** In all cases, as $\omega \to \infty$, $d_1 d_2 \to \infty$ and $|H(\infty)| \to 0$.

In summary, the paired poles define a second-order filter that behaves as a lowpass filter that has a double breakpoint frequency $\omega_B$ where $|H(\omega)|$ falls off at

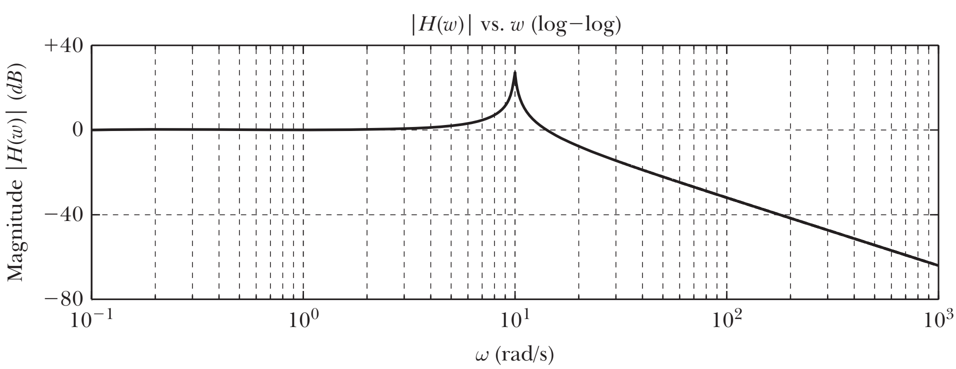



**FIGURE 7.13** **Resonance Frequency** Figure 7.12B shows the special case of the frequency response of a system in which a pole pair is located close to $\omega_R$ on the vertical axis. The resulting *resonance* peak is shown here for ($\omega_R = 10$ rad/s). At the extremes, the system behaves as a second-order lowpass filter.

twice the rate as compared to the previous single-pole example. For values of $a < \omega_0$ there may be a significant resonance at the peak frequency $\omega_P$, where $\omega_P < \omega_0 < \omega_B$. Both $\omega_P$ and $\omega_B$ converge to $\omega_0$ as $a \to 0$. In contrast to the resonance, if a zero is found on the $\sigma = 0$ axis, the frequency response goes to zero at that point.

**Resonance Frequency** The phenomenon of resonance as seen in Figure 7.13 can be desirable or undesirable depending on the expected system properties. Having a resonance point suggests that the system response is very different for specific input frequency components. Establishing a strong and very sharp resonance peak often becomes the focus of system design. For example, in a radio receiver where the goal is to listen to a particular station by tuning to its frequency (e.g., tuning an FM radio to 103.5 MHz), a *tuned circuit* having a very sharp resonance would be especially desirable as the signal arriving at the receiver antenna can be greatly amplified while other nearby stations would not. When the pendulum on a clock is adjusted to maintain accurate timekeeping, its length is varied so that the mechanical system resonance is close to 1 Hz. Another desirable and nearly undamped resonance is identified when a tuning fork is struck (impulse input) and it vibrates at a fixed audio frequency while the resulting sound only gradually fades away. An undesirable resonance may occur when an automobile wheel is out of balance and there is a particular speed at which the wheel begins to shake, sometimes violently; the solution is to slow down (or speed up) and move the system input away from the resonance frequency. In this case, the mechanical system response includes the suspension system, the coil spring, and dampers (shock absorbers) driven by a periodic source being the unbalanced wheel. The long-term solution is to modify the system by having the wheels balanced so that any resonance is at least outside of normal driving speeds.

**Multiple Poles and Zeros** Since all real poles are expected to be single poles on the horizontal axis or paired poles equally spaced above and below the horizontal axis, then the above examples embody all combinations of multiple poles. The presence of any other poles would have a similar influence on the magnitude $|H(\omega)|$ with the values of $(a, \omega_0)$ for each determining their contribution to the form of the frequency response. The combined effect of several poles determines the denominator of an overall output as:

$$|H(w)| = K\left[\frac{1}{d_1 \times d_2 \times d_3 \ldots \times d_n}\right]$$

Zeros on the $s$-domain with distances $D$ have a corresponding influence on the numerator as:

$$|H(w)| = K\left[\frac{D_1 \times D_2 \times D_3 \ldots \times D_n}{d_1 \times d_2 \times d_3 \ldots \times d_n}\right] \tag{7.16}$$

### 7.4.4 Two-Sided Laplace Transform

The Laplace transform is very often defined and used exclusively in its one-sided form (integrating from zero to infinity) although this is only a specific case of the more general two-sided or *bilateral* Laplace transform. The most significant difference with the two-sided transform is found in defining the region of convergence and recognizing situations where the ROC and consequently the Laplace transform does not exist.

Consider the function $x(t) = e^{-a|t|}$ that exists for all $t$ and its two-sided Laplace transform. The function $x(t)$ may be viewed as the sum $x(t) = a(t) + b(t)$, or the unilateral exponential $a(t) = e^{-at}u(t)$ plus its reflection $b(t) = e^{+at}u(-t)$, each having a region of convergence and together having a region of convergence that must satisfy both conditions.

$$\mathcal{L}\left[x(t)\right] = X(s) = \int_{-\infty}^{+\infty} e^{-a|t|}e^{-st}\, dt \tag{7.17}$$

The two-sided transform may be computed in two parts as shown below. For this particular $x(t)$, which includes an absolute value, the integration would follow this same approach in any case.

$$X(s) = \int_{-\infty}^{0} e^{-a|t|}e^{-st}\, dt + \int_{0}^{+\infty} e^{-a|t|}e^{-st}\, dt \tag{7.18}$$

Now, in the first integral $t < 0$ where $(|t| = -t)$, while in the second integral, $t > 0$ where $(|t| = +t)$, giving:

$$X(s) = \int_{-\infty}^{0} e^{(-s+a)t}\, dt + \int_{0}^{+\infty} e^{(-s-a)t}\, dt \tag{7.19}$$

$$X(s) = \frac{-1}{s-a}e^{-(s-a)t}\bigg|_{t=-\infty}^{0} + \frac{-1}{s+a}e^{-(s+a)t}\bigg|_{t=0}^{+\infty} \tag{7.20}$$

The regions of convergence are given by observing that:

- The left-hand integral converges as $t \to -\infty$ only when $-(s-a) > 0$ or $s < +a$.
- The right-hand integral converges as $t \to \infty$ only when $-(s+a) < 0$ or $s > -a$. This term alone corresponds to the one-sided Laplace transform.
- The overall result converges only for values of $-a < s < +a$. This interval is the overlap or intersection of the above two regions.

With the above constraint:

$$X(s) = \frac{-1}{s-a} + \frac{+1}{s+a} = \frac{-2a}{(s-a)(s+a)} = \frac{-2a}{s^2 - a^2} \tag{7.21}$$

As shown in Figure 7.14, there are two poles, located one on either side of the origin, defining a region of convergence that is a vertical strip. Significantly, this ROC includes the $\sigma = 0$ axis such that the Fourier transform of $s(t)$ exists and may be found directly from the above result by setting $s = j\omega$:

$$S(f) = \frac{-2a}{(j\omega)^2 - a^2} = \frac{2a}{a^2 + (2\pi f)^2} \tag{7.22}$$

This result may be confirmed in Figure A.20 in Appendix A, with $a = 1$.

The bilateral Laplace transform will always have an ROC describing a vertical strip, provided the ROC exists. In this example, for values of $a < 0$, there is no region of convergence, and the Laplace transform does not exist.

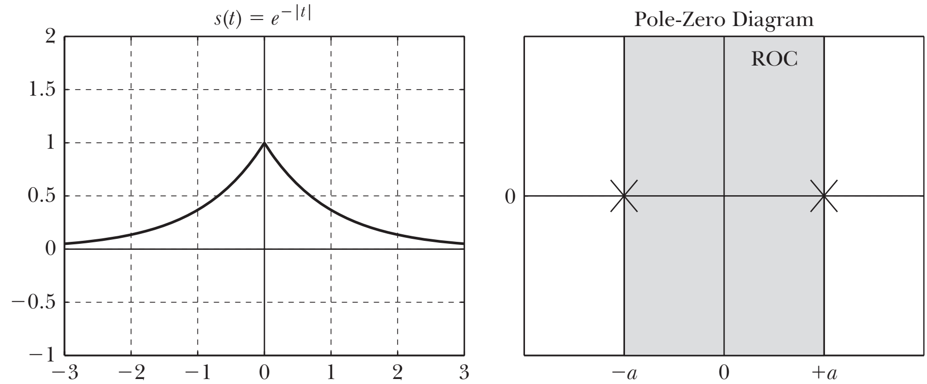



**FIGURE 7.14** **Two-Sided Laplace Transform ROC** The bilateral Laplace transform has a region of convergence (ROC) that defines a vertical strip in the *s*-plane (provided the ROC exists). In this example, the bilateral Laplace transform of the function $s(t) = e^{-a|t|}$ has a strip-like ROC that exists for all $a > 0$.

## 7.4.5 The Bode Plot

The three reference points identified above can be seen in Figure 7.15 where the overall frequency response of this filter is shown on a Bode plot, for positive values of $\omega$; this is a profile of the magnitude $|H(\omega)|$ vs. $\omega$ drawn on a log-log scale. The limiting points $(\omega \to 0)$ and $(\omega \to \infty)$ computed above match this graph, while the pole at the breakpoint position aligns with the point where the amplitude is $1/\sqrt{2} = 0.707$. This Bode plot has a distinctive form that resembles two straight lines on either side of the breakpoint. The nearly flat line to the left of the breakpoint represents low frequencies that pass relatively unaffected by the filter. The line to the right of the breakpoint has a distinct slope that can be observed to drop ten times in amplitude for every ten times increase in frequency. This decreasing amplitude with increasing frequency is expressed as *minus twenty dB per decade* or $-20$ dB/dec, or a *minus ten dB per decade* drop in power where:

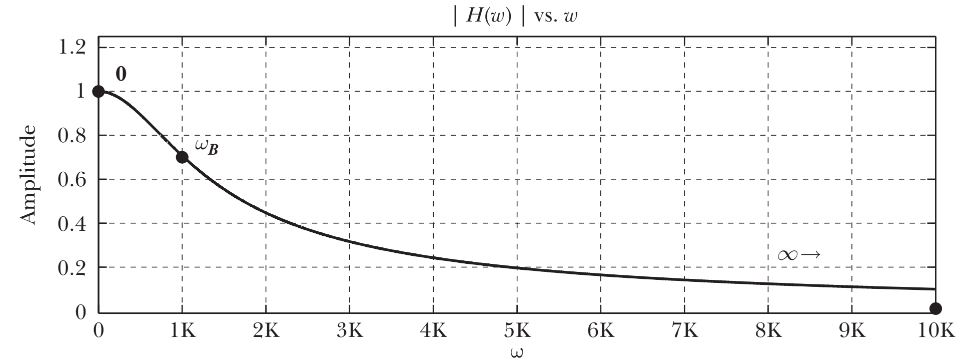
$$\textit{change in signal level } (dB) = 20log_{10}\left[\frac{V_{out}}{V_{in}}\right] = 10log_{10}\left[\frac{V_{out}^2}{V_{in}^2}\right]$$

A ten times drop in voltage (0.1) is $20log_{10}[0.1] = -20$ dB. Similarly, the value of the curve at the breakpoint is $20log_{10}(0.707) \approx -3$ dB down from the initial value of the curve. This same rate of change is often described as a drop by half in amplitude for every doubling in frequency, expressed as *minus six dB per octave* or $-6$ dB/oct.

One great advantage of the Bode plot is that it can be sketched by hand with a pencil and ruler by simply noting the reference frequency values shown above. In this case, two straight-line segments define the graph, one flat and the other with a slope of $-20$ dB/dec. Lines are drawn by identifying the limiting cases and the breakpoint for:

$$H(s) = k\left[\frac{1}{s+a}\right], \qquad \textit{with } k = a$$

1. $\omega \to 0$: value $H(0) = k/a = 1$, starting slope is zero; phase = 0 deg;
2. $\omega = a$: POLE : line $\downarrow$ $-20$ dB/dec; phase $\downarrow$ $-90$ deg;
3. $\omega \to \infty$: limiting slope = $-20$ dB/dec; limiting phase = $-90$ deg.



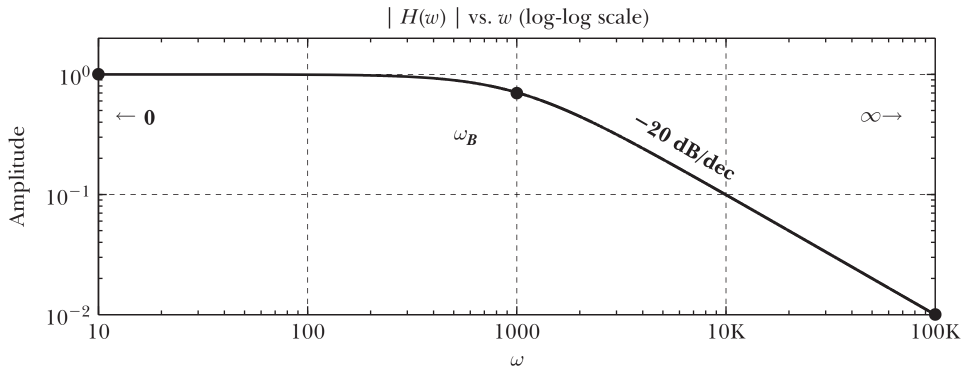




**FIGURE 7.15 Sketching the Lowpass Filter** The reference points shown are the limit as $(\omega \to 0)$, $(\omega \to \infty)$, and the breakpoint $\omega = a$ shown here for $(a = 10)$. The upper plot shows the magnitude $|H(\omega)|$ as a function of $\omega$ while the lower graph shows the same information plotted in log-log format as a *Bode plot* where notably the curve approximates two straight-line segments.

The resulting approximation is shown in Figure 7.16 along with the actual $|H(f)|$ curve. It can be seen that this straight-line approximation is very good with the largest difference being at the breakpoint where the curve lies $-3$ dB below the approximation. Similarly, there is a $-90$ deg phase change associated with the pole.

Moving along the frequency axis, multiple breakpoints corresponding to poles at various frequencies each add another $-20$ dB/dec to the slope, and another $-90$ deg to the overall phase change. In contrast, multiple breakpoints corresponding to zeros at various frequencies each add another $+20$ dB/dec to the slope, and another $+90$ deg to the overall phase change. In summary:

1. **POLE** at $\omega_B$: straight-line changes slope by $-20$ dB/dec, phase changes by $-90$ deg;
2. **ZERO** at $\omega_B$: straight-line changes slope by $+20$ dB/dec, phase changes by $+90$ deg.

**Bode Plot—Multiple Poles and Zeros** Consider the transfer function below with one zero and two poles. The Bode plot approximation to $|H(\omega)|$ can be sketched by hand following the above rules.

$$H(s) = 10\left[\frac{s+5}{(s+5)^2+100}\right]$$

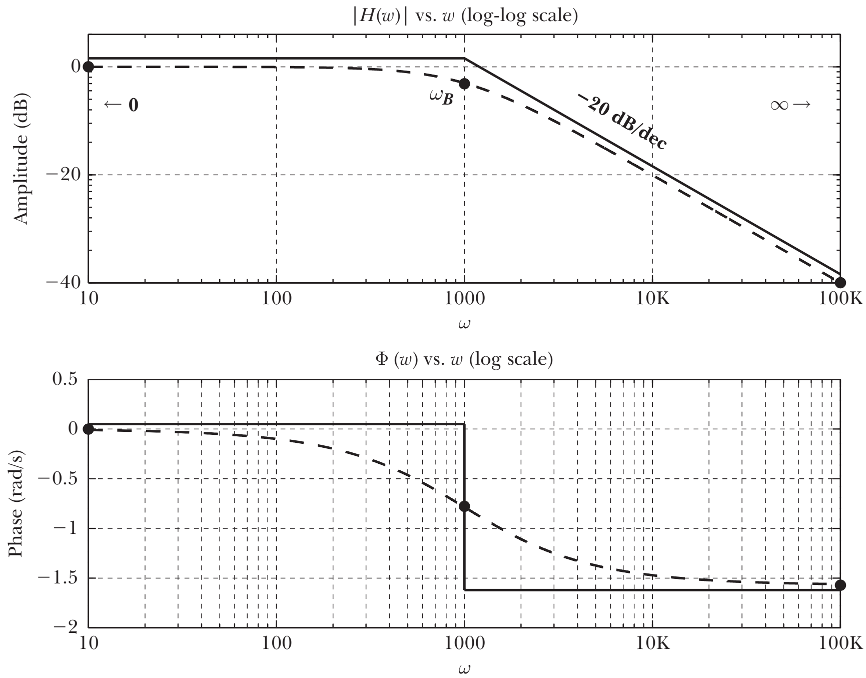



**FIGURE 7.16 Sketching the Lowpass Filter** The reference points shown are the limit as ($\omega \to 0$), ($\omega \to \infty$), and the computed breakpoint position, shown here at ($\omega_B = 1000$ rad/s). The dashed line on the upper plot shows magnitude $|H(\omega)|$ on a log-log graph. The dashed line on the lower graph shows the phase response $\Phi(\omega)$ on a log-linear scale. The corresponding straight-line approximations are sketched slightly offset from each line for clarity.

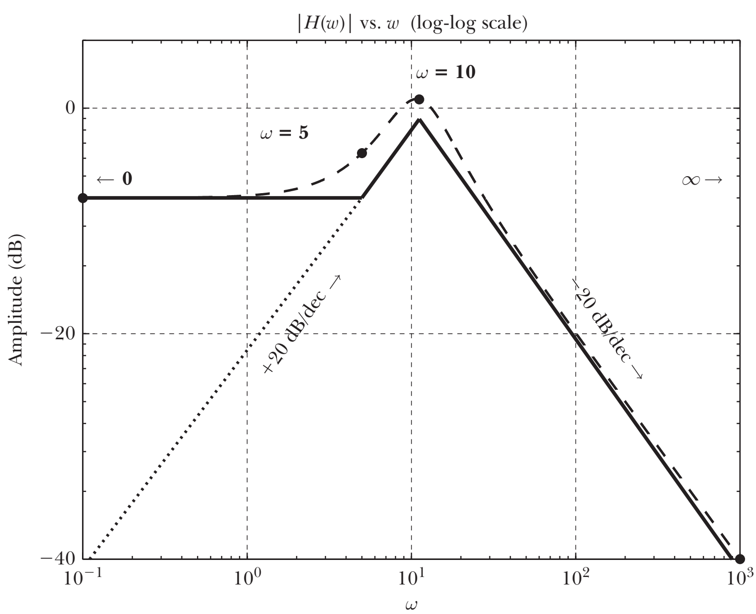



**FIGURE 7.17 Sketching the Bode Plot—Poles and Zeros** The dashed line shows the magnitude $|H(\omega)|$ on a log-log graph. The straight-line approximations can be sketched by hand. The reference points shown are the limit as ($\omega \to 0$), ($\omega \to \infty$), and the computed breakpoint positions, a zero at 5 rad/s, and a double pole at 10 rad/s.

By inspection, the numerator is zero for $s = -5$, defining a *zero* at $\omega = 5$ rad/s. The denominator may be factored to give roots at $s = -5 \pm j10$ defining a *double pole* at $\omega = \sqrt{5^2 + 10^2} = \sqrt{125} = 11.2$ rad/s. For low frequencies, $H(0) = 10$, while for high frequencies, $H(s \to \infty)$ goes down as $1/s$ or $-20$ dB/dec. The Bode diagram has several key points, as shown in Figure 7.17.

1. $\omega \to 0$, value $H(0) = 0.4$, starting slope is zero; starting phase = 0 deg;
2. $\omega = 5$, (ZERO) line turns up +20 dB/dec; phase changes by +90 deg;
3. $\omega = 11.2$, (POLE) line turns down $-20$ dB/dec; phase changes by $-90$ deg;
4. $\omega = 11.2$, (POLE) line turns down $-20$ dB/dec; phase changes by $-90$ deg;
5. $\omega \to \infty$, limiting slope = $-20$ dB/dec; limiting phase = $-90$ deg.

For clarity, the double pole is listed as two poles at the same frequency. The straight-line approximation to the Bode plot may now be sketched as in Figure 7.17. The actual form of $|H(\omega)|$ is shown as a dashed line for reference. Note that the peak indicates a resonance condition at $\omega = \sqrt{125} = 11.2$ rad/s.

**Laplace Transform Exercise 1: Calculating the Laplace Transform** Compute the Laplace transform of $h(t) = u(t)10e^{-5t}\cos(10t)$. This is a cosine with an amplitude that diminishes over time. The result may be expected to resemble Eqn. 7.10.

**Solution:** Before starting this problem, think about what the answer should resemble. The cosine function has a pair of poles at $\pm\omega_0 = 10$ rad/s and a zero at the origin. The decreasing exponential component serves to shift the cosine poles and zeros into the left-hand plane a distance $\sigma = 5$. The entire function is multiplied by 10 such that $h(0) = 10$.

The Laplace transform of $h(t)$ is given by:

$$H(s) = \int_0^\infty u(t)10e^{-5t}\cos(10t)e^{-st}\,dt \tag{7.23}$$

expressing $\cos(10t)$ in exponential form gives:

$$= 10\int_0^\infty e^{-5t}\left[\frac{e^{+j10t} + e^{-j10t}}{2}\right]e^{-st}\,dt \tag{7.24}$$

$$= 5\int_0^\infty e^{-5t}\left[e^{+j10t} + e^{-j10t}\right]e^{-st}\,dt \tag{7.25}$$

$$= 5\int_0^\infty e^{-5t}e^{+j10t}e^{-st}\,dt + 5\int_0^\infty e^{-5t}e^{-j10t}e^{-st}\,dt \tag{7.26}$$

$$= 5\int_0^\infty e^{-(s+5)t+j10t}\,dt + 5\int_0^\infty e^{-(s+5)t-j10t}\,dt \tag{7.27}$$

$$= 5\left[\frac{-1}{(s+5)+j10}\right]e^{-[(s+5)+j10]t}\Big|_{t=0}^{\infty} + 5\left[\frac{-1}{(s+5)-j10}\right]e^{-[(s+5)-j10]t}\Big|_{t=0}^{\infty}$$

the limit as $t \to \infty$ converges to zero only for $s > -5$, so the region of convergence is $s > -5$. The lower limit is 1 for $t = 0$, leaving two terms:

$$= 5\left[\frac{1}{(s+5)+j10}\right] + 5\left[\frac{1}{(s+5)-j10}\right] \tag{7.28}$$

rationalizing the denominator in each term leaves:

$$= 5\left[\frac{(s+5)-j10}{(s+5)^2+10^2} + \frac{(s+5)+j10}{(s+5)^2+10^2}\right] \tag{7.29}$$

$$H(s) = 5\left[\frac{2s+10}{(s+5)^2+10^2}\right] = 10\left[\frac{s+5}{(s+5)^2+100}\right] \tag{7.30}$$

which, as expected, resembles the cosine Laplace transform shifted left to $\sigma = -5$. The pole-zero diagram is obtained by observing when the numerator or the denominator is zero. As expected, the denominator resembles the form of Eqn. 7.10 where there are two poles located at $(\sigma = -5, \omega = \pm 10)$. The numerator is zero for $s = -5$, so there is a zero located at $(\sigma = -5, \omega = 0)$.

## 7.4.6 System Analysis in MATLAB

A system can be described and manipulated in MATLAB where the pole-zero diagram and Bode plots can be readily obtained using functions from the *Control System Toolbox*. For example:

1. `sys = tf(...)`—define a system named sys, then
2. `bode(sys)`—plot the transfer function magnitude and phase
3. `bodemag(sys)`—plot the transfer function magnitude only
4. `pzmap(sys)`—plot the transfer function pole-zero diagram
    a. `[p,z] = pzmap(sys)`—return any poles ($p$) and zeros ($z$)
    b. `pole(sys)`—returns the value of any poles
    c. `zero(sys)`—returns the value of any zeros
5. `impulse(sys)`—plot the corresponding impulse response
6. `step(sys)`—plot the corresponding step response
7. `lsim(sys,input,t)`—plot an arbitrary input response

A system defined by its transfer function $H(s)$ can use the `tf` function in one of two ways. For example, consider the transfer function from Eqn. 7.30:

$$H(s) = 5\left[\frac{2s+10}{(s+5)^2+10^2}\right]$$

To define this system, the variable $s$ can be assigned as a transfer function variable, after which $H(s)$ can be defined exactly as the equation is written.

```
s = tf('s');
sys = 5 * (2 * s + 10) / ((s + 5)^2+10^2)
Transfer function :

    10s + 50
-----------------
s^2+ 10 s + 125
```

It can be seen that MATLAB has identified sys as a transfer function and expanded the numerator and denominator into polynomials in terms of *s*. Alternatively, the transfer function can be specified directly in the form [*numerator*, *denominator*] where the three denominator terms in $s^N$ are in the form [1 10 125] while the numerator terms are [10 50]. In this case, *s* need not be defined; instead, using tf(num, den) gives:

```
sys = tf([10 50], [1 10 125])
Transfer function :

   10s + 50
--------------
s^2+ 10 s + 125
```

Using either method, the system object named sys has been established. The pole-zero diagram and the impulse and step responses may now be generated for this sys as shown below. In this example, three subplots are defined so that all three graphs can appear in the same figure.

```
figure(1);                   % point to Figure 1
clf                          % clear the figure if necessary
subplot(2,1,1);              % define first of two plots
pzmap(sys);                  % show the pole-zero diagram
axis([-11 +1 -15 +15 ]);     % rescale for visibiiity
subplot(2,1,2);              % define second of three plots
impulse(sys);                % show the impulse response
subplot(3,1,3);              % define third of three plots
step(sys);                   % show the step response
```

The result may be seen in Figure 7.18 where the axes of the pole-zero diagram have been extended slightly to better see the pole and zero positions. This solution is verified in the graph of the impulse function, where the value at $t = 0$ corresponds to the value $h(0) = 10$ in the original problem statement.

Similarly, the Bode plot for a system sys defined as above is obtained in MATLAB as in Figure 7.19:

```
figure(2);    % define a figure for Bode Plot
clf           % clear the figure if necessary
bode(sys);    % show the Bode plot (Freq and Phase)
grid on;      % embellish the Bode Plot
```

Alternatively, a system sys can be defined from a knowledge of the poles and zeros. Using the function zpk(zeros,poles,k), then the zero (at $s = -5$) and poles (at $s = -5 + 10i$ and $s = -5 - 10i$) lead to the same system as above. The factor *k* is an overall gain for the transfer function, as a zero at $\sigma = -5$ is found for any numerator of the form $(k[s+5])$; in this example, let $k = 10$:

```
% define the system from poles and zeros and gain (k=1)
sys = zpk([-5], [-5+10i -5+10i], 10)
Zero/pole/gain:

    10 (s + 5)
-----------------
(s^2+ 10s + 125)
```

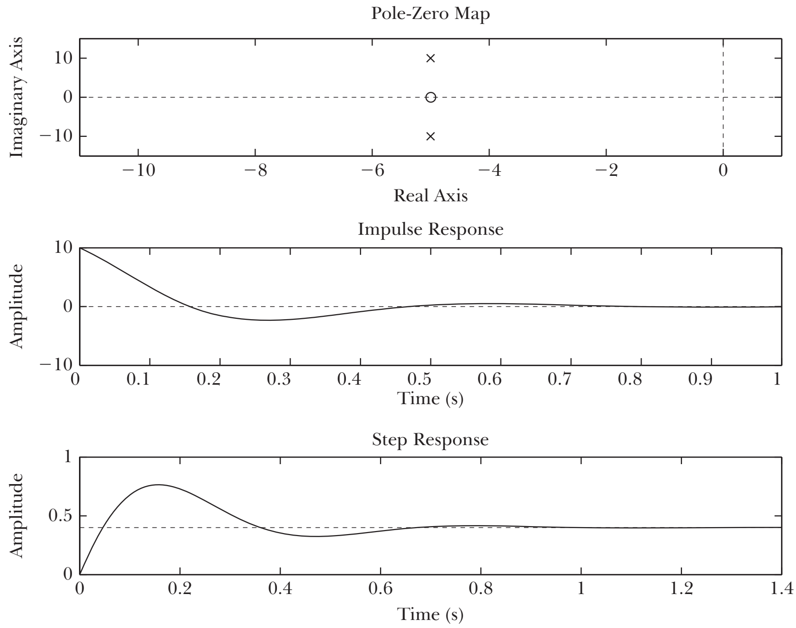



**FIGURE 7.18 Pole-Zero Diagram using MATLAB** The pole-zero diagram, impulse response, and step response are generated directly in MATLAB given the transfer function of Eqn. 7.30 defining a cosine with amplitude diminishing over time.

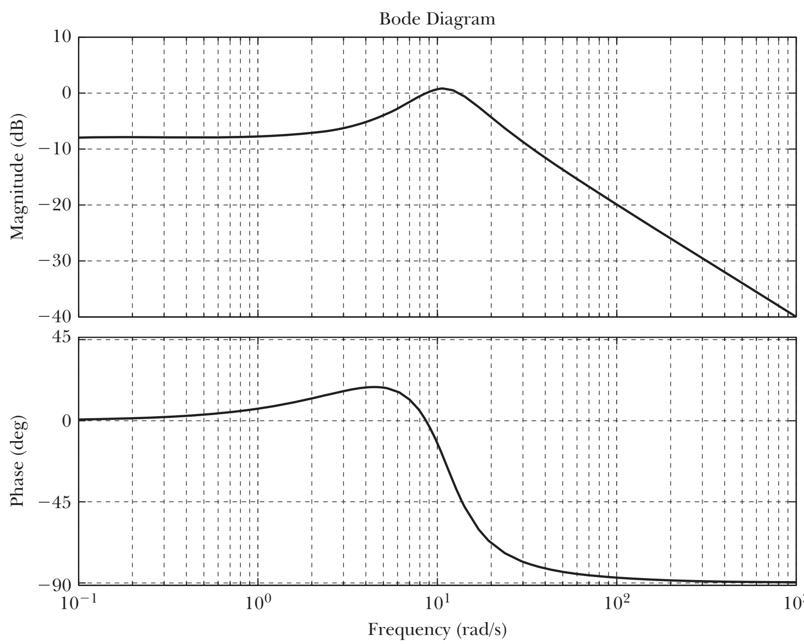



**FIGURE 7.19 Bode Plot using MATLAB** The frequency response (above) and phase response (below) are generated directly in MATLAB given the transfer function of Eqn. 7.30 defining a time domain $h(t)$ as a cosine with amplitude diminishing over time. This transfer function represents a second-order lowpass filter. Compare to Figure 7.17.

Defining a system model in this way allows the study of the effect as poles and zeros are moved about the $s$-plane. It will be seen in Section 7.9 that the same system model can be obtained by describing components and their interconnections.

# 7.5 Properties of the Laplace Transform

The properties of the Laplace transform are very similar to those of the Fourier transform. Stretching, scaling and shifting, convolution and multiplication, and the overall linearity of the transform make most of the following rules very similar to their Fourier transform counterparts.

**1.** Linearity

$$k_1 a_1(t) + k_2 a_2(t) \xleftrightarrow{\mathcal{L}} k_1 A_1(s) + k_2 A_2(s) \tag{7.31}$$

**2.** Scaling

$$a(kt) \xleftrightarrow{\mathcal{L}} \frac{1}{|k|} A\left(\frac{s}{k}\right) \tag{7.32}$$

**3.** Shifting

$$a(t-k) \xleftrightarrow{\mathcal{L}} e^{-sk} A(s) \tag{7.33}$$

**4.** Convolution $\xleftrightarrow{\mathcal{L}}$ Multiplication

$$a(t) * b(t) \xleftrightarrow{\mathcal{L}} A(s) \times B(s) \tag{7.34}$$

**5.** Differentiation (first order)

$$\frac{d}{dt} a(t) = a'(t) \xleftrightarrow{\mathcal{L}} sA(s) - a(0) \tag{7.35}$$

**6.** Differentiation (second order)

$$\frac{d^2}{dt} a(t) = a''(t) \xleftrightarrow{\mathcal{L}} s^2 A(s) - sa(0) - a'(0) \tag{7.36}$$

**7.** Integration

$$\int_0^t a(\tau)\, d\tau \xleftrightarrow{\mathcal{L}} \frac{1}{s} A(s) \tag{7.37}$$

Proof of the differentiation property is found in the following section.

# 7.6 Differential Equations

Fundamentally, the most useful Laplace transform property relates the derivative(s) of a function to simple algebraic relationships in the $s$-domain. This observation makes the transform extremely useful in solving differential equations or, alternatively, in analyzing linear systems without having to deal directly with the underlying differential equations.

Consider the Laplace transform of the time derivative of $f(t)$, written as $\mathcal{L}[f'(t)]$. From the definition of the Laplace transform:

$$\mathcal{L}\left[f'(t)\right] = \int_0^\infty f'(t)e^{-st}\,dt \tag{7.38}$$

Using integration by parts and recognizing that:

$$\int u\,dv = uv - \int v\,du, \text{ with : } \begin{array}{ll} u = s^{-st} & v = f(t) \\ du = -se^{-st}dt & dv = f'(t)dt \end{array}$$

gives:

$$\mathcal{L}\left[f'(t)\right] = s^{-st}f(t)|_{t=0}^{\infty} + s\int_0^\infty f(t)e^{-st}\,dt \tag{7.39}$$

where the first term is expected to go to zero for $t \to \infty$ and the remaining integral is recognized as the Laplace transform of $f(t)$, giving:

**THEOREM 7.1** *(Derivative)*

*If*

$$f(t) \overset{\mathcal{L}}{\longleftrightarrow} F(s)$$

*then*

$$f'(t) \overset{\mathcal{L}}{\longleftrightarrow} sF(s) - f(0)$$

The value $f(0)$ represents the *initial condition* of the function $f(t)$. By extension, the Laplace transform of the second derivative is given by:

**THEOREM 7.2** *(Second Derivative)*

*If*

$$f(t) \overset{\mathcal{L}}{\longleftrightarrow} F(s)$$

*then*

$$f''(t) \overset{\mathcal{L}}{\longleftrightarrow} s^2F(s) - sf(0) - f'(0)$$

and extensions to higher-order derivatives follow directly from these results.

## 7.6.1 Solving a Differential Equation

Consider the circuit of Figure 7.20 with $R = 1000\,\Omega$ and $C = 10\,\mu F$, and let the voltage across the capacitor be $v(t)$. It is clear that if the capacitor is discharged both the voltage and current in the circuit are zero and nothing happens; however, if the

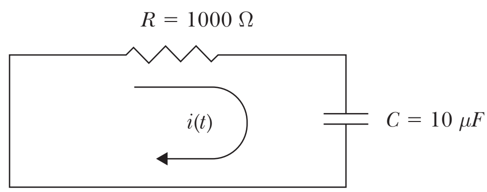



**FIGURE 7.20** **Solving a Differential Equation** Find the voltage $v(t)$ across the capacitor if the initial value $v(0) = 10$ V. The solution involves a first-order differential equation.

capacitor is initially charged then the nontrivial result ought to be an exponentially decreasing voltage $v(t)$ as the capacitor discharges to zero. The circuit behavior is modelled as a differential equation with initial conditions. Let the *initial value* of $v(t)$ be $v(0) = 10$ Volts.

Given that the voltage across the capacitor is $v(t)$, then the voltage across the resistor is $-v(t)$ since the total voltage around the loop must be zero. Let the current in the loop be $i(t)$ as shown, where the series components must have the same current passing through them. The current through the resistor is $i(t) = -v(t)/R$, and the current through the capacitor is $i(t) = Cdv/dt$, giving:

$$C\frac{dv}{dt} + \frac{1}{R}v(t) = 0, \quad \textit{with } v(0) = 10 \tag{7.40}$$

This differential relationship is identified as the *characteristic equation* of the series RC circuit. Its solution begins by taking the Laplace transform to give:

$$C[sV(s) - v(0)] + \frac{1}{R}V(s) = 0 \tag{7.41}$$

where the initial condition $v(0)$ comes from the Laplace transform of the derivative, Theorem 7.1.

$$V(s)\left[s + \frac{1}{RC}\right] = v(0) \tag{7.42}$$

$$V(s) = v(0)\left[\frac{1}{s + \frac{1}{RC}}\right] \tag{7.43}$$

The time-domain solution $v(t)$ to this problem is the inverse Laplace transform of this $V(s)$. It is not always straightforward to calculate the inverse transform, and the method of finding $v(t)$ for simple transforms is to recognize the form of $V(s)$ or to look it up in a table of Laplace transforms as found in Section 7.7

$$v(t) = v(0)e^{\frac{-1}{RC}t} \tag{7.44}$$

where the actual values may now be inserted:

$$v(t) = 10e^{\frac{-1}{(1000)(10\times 10^{-6})}t} = 10e^{-100t} \tag{7.45}$$

To check this result, first confirm that $v(0) = 10$, then compute $dv/dt$ to evaluate Eqn. 7.40 with the given values of ($R = 1000$, $C = 10^{-5}$):

$$C\frac{dv(t)}{dt} + \frac{1}{R}v(t) = C\,[-1000e^{-100t}] + \frac{1}{R}[10e^{-100t}] = 0, \quad \text{as expected.}$$

**Compound Interest** An interesting example of a first-order differential relationship is found in computing compound interest for a bank account. Let the balance in a bank account at year $t$ be $b(t)$ and the annual interest paid be 7 percent of this balance. The change in balance $(db/dt)$ at year $t$ is given by the differential equation:

$$\frac{db}{dt} = 0.07\, b(t), \quad \text{with } b(0) = \text{initial balance}$$

which, proceeding as above, leads to the Laplace transform:

$$B(s) = b(0)\left[\frac{1}{s - 0.07}\right]$$

where it can be noted that a pole-zero diagram would have a single pole at $(\sigma = +0.07)$ on the *right-hand* plane. Indeed, the bank balance would grow exponentially and without limit if no funds were withdrawn. If the initial value of $b(t)$ is \$100 deposited at time $t = 0$, then the balance $b(t)$ at year $t$ is:

$$b(t) = 100e^{+0.07t}$$

Note that after 10 years this balance is $b(10) = \$201$ and the original amount has doubled. A more conventional way to find this result would use the formula $\$100 \times 1.07^{10} = \$197$, the difference being that the differential equation describes interest compounded *continuously* rather than at year end. A closer correspondence would be found for accounts where daily interest is calculated.

### 7.6.2 Transfer Function as Differential Equations

The transfer function $H(s)$ relates input and output signals. As shown below, $x(t)$ is related to $y(t)$ through differential equations and where initial conditions may be included as required.

$$2\frac{d^2x}{dt} + 3\frac{dx}{dt} + 5x = \frac{dy}{dt} + 4y \tag{7.46}$$

in the $s$-domain:

$$[2s^2 + 3s + 5]\, X(s) = [s + 4]\, Y(s) \tag{7.47}$$

leaving

$$\frac{X(s)}{Y(s)} = \frac{s + 4}{2s^2 + 3s + 5} \tag{7.48}$$

## 7.7 Laplace Transform Pairs

A list of one-sided time-domain signals is shown below along with the corresponding Laplace transform for each. An inverse Laplace transform is often determined with reference to a table rather than by direct computation; in practice, it may be necessary to apply the Laplace transform properties and/or to factor a given Laplace transform into components that can be easily matched to $s$-domain functions as found here.

1. Unit Impulse

$$\delta(t) \overset{\mathcal{L}}{\longleftrightarrow} 1 \tag{7.49}$$

2. Unit Step

$$u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{1}{s} \tag{7.50}$$

3. Unit Ramp $(t > 0)$

$$tu(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{1}{s^2} \tag{7.51}$$

4. Exponential

$$e^{-at}\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{1}{s+a} \tag{7.52}$$

5. Sinusoid

$$\cos(\omega_0 t + \Phi)\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{s\cos(\Phi) - \omega_0 \sin(\Phi)}{s^2 + \omega_0^2} \tag{7.53}$$

6. Cosine

$$\cos(\omega_0 t)\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{s}{s^2 + \omega_0^2} = \frac{1}{2}\left[\frac{1}{s - j\omega_0} + \frac{1}{s + j\omega_0}\right] \tag{7.54}$$

7. Sine

$$\sin(\omega_0 t)\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{\omega_0}{s^2 + \omega_0^2} = \frac{1}{2j}\left[\frac{1}{s - j\omega_0} - \frac{1}{s + j\omega_0}\right] \tag{7.55}$$

8. Exponential $\times$ Cosine

$$e^{-at}\cos(\omega_0 t)\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{s+a}{(s+a)^2 + \omega_0^2} \tag{7.56}$$

9. Exponential $\times$ Sine

$$e^{-at}\sin(\omega_0 t)\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{\omega_0}{(s+a)^2 + \omega_0^2} \tag{7.57}$$

Hyperbolic formulas (*sinh* and *cosh*) can be derived from trigonometric identities by replacing $t$ by $jt$ [7]. For example, $sinh(t) = -j\sin(jt)$ and $cosh(t) = \cos(jt)$. These functions include real-valued exponentials.

10. Hyperbolic Cosine

$$cosh(at)\, u(t) = \frac{1}{2}[e^{+at} + e^{-at}]\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{s}{s^2 - a^2} \tag{7.58}$$

11. Hyperbolic Sine

$$sinh(at)\, u(t) = \frac{1}{2}[e^{+at} - e^{-at}]\, u(t) \overset{\mathcal{L}}{\longleftrightarrow} \frac{a}{s^2 - a^2} \tag{7.59}$$

$$\left[e^{-2t} + e^{-30t}\right] \quad u(t) \quad \stackrel{\mathcal{L}}{\longleftrightarrow}$$

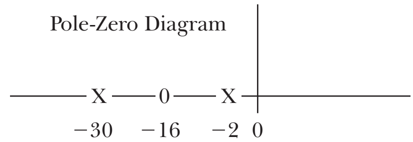



**FIGURE 7.21** The Laplace transform of the sum of two exponential functions contains two distinct poles as expected, plus a (new) zero located midway between the poles. This zero can be found in the sum of the component Laplace transforms, but is not evident from the individual pole-zero diagrams.

## 7.7.1 The Illustrated Laplace Transform

Appendix B presents a number of Laplace transforms shown as a pole-zero diagrams along with the corresponding mathematical form of each. Poles are shown as X while zeros appear as O. It is useful to refer to this appendix as confirmation of the properties of the Laplace transform and as an additional perspective on the above table of Laplace transform pairs.

Because the Laplace transform is linear, the Laplace transform of the sum of two signals $c(t) = a(t) + b(t)$ equals the sum of the transforms for $a(t)$ and for $b(t)$. On the other hand, the resulting pole-zero diagram is *not* expected to be the sum of pole-zero diagrams; instead, the new pole-zero diagram must be found directly from the Laplace transform $C(s)$. For example, let $a(t) = e^{-2t}$ and $b(t) = -30t$. The Laplace transform $A(s)$ has a single pole at $s = -2$, and $B(s)$ has a single pole at $s = -30$. The Laplace transform of the sum is given by:

$$[e^{-2t}u(t)] + [e^{-30t}]\, u(t) \stackrel{\mathcal{L}}{\longleftrightarrow} \frac{1}{[s+2]} + \frac{1}{[s+30]} = \frac{2s+32}{s^2+32s+60}$$

where it can be seen in the numerator that a zero has emerged at $s = -16$.

As seen in Figure 7.21, the resulting Laplace transform has poles at $s = -2$ and $s = -30$ as expected, plus a zero located halfway between the two poles.

In contrast, the Laplace transform of the *convolution* $a(t) * b(t)$ contains only the same poles (and/or zeros) associated with the Laplace transforms $A(s)$ and $B(s)$ as found directly in the product $A(s)B(s)$, or:

$$[e^{-2t}u(t)] * [e^{-30t}]\, u(t) \stackrel{\mathcal{L}}{\longleftrightarrow} \frac{1}{[s+2]} \times \frac{1}{[s+30]} = \frac{1}{s^2+32s+60}$$

# 7.8 Circuit Analysis with the Laplace Transform

The Laplace transform is indispensable to linear circuit analysis as many time-domain circuit calculations are greatly simplified when working in the *s*-domain. Circuit components that have differential or integral response functions (capacitors and inductors) can be treated as simple impedances in the *s*-domain, and the Laplace transform naturally includes the ability to specify initial conditions if necessary. Analyzing the system-level behavior of circuits is also made easier as the performance and overall characteristics of circuits emerge directly from the Laplace transform.

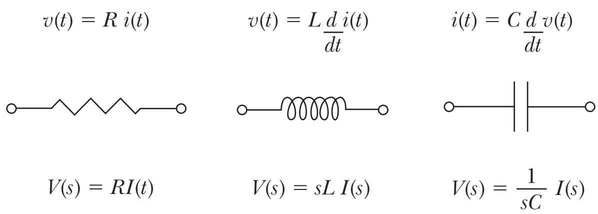



**FIGURE 7.22 Resistor, Inductor, and Capacitor** The time-domain voltage-current relationships (above the figures) governing three basic circuit components involve differential relationships that complicate advanced circuit analysis. In the *s*-domain (below the figures), the same expressions simplify to complex impedances that can be readily manipulated. Initial values are assumed to be zero.

Consider the ideal resistor, inductor, and capacitor shown in Figure 7.22. Significantly, the familiar resistor formula $V = IR$ is a straightforward linear relationship, implying that series and parallel resistor circuits can be simplified and analyzed by studying the current flows and voltage drops across components that are the same for both AC and DC signals. The situation is very different for inductors and capacitors. As their circuit symbols suggest, the ideal inductor is merely a piece of wire having no resistance, while an ideal capacitor consists of two plates that are not connected together and are essentially an open circuit. Of course, it is the transient and AC characteristics of these components that make them useful, and these relationships follow the time derivative of the voltage (capacitor) or the current (inductor). Consequently, full analysis of a circuit that includes inductors or capacitors inevitably involves deriving and solving differential equations that can quickly become quite cumbersome.

An inductor with inductance $L$ is governed by:

$$v(t) = L\frac{d}{dt}i(t)$$

From the Laplace transform pair in Eqn. 7.35 the *s*-domain version of this relationship is:

$$V(s) = Ls\, I(s)$$

Where the *impedance* (Z) of the inductor is $Ls$, and otherwise this can be recognized as $V = IR$ with $R$ replaced by $Ls$.

A capacitor with capacitance $C$ is governed by:

$$i(t) = C\frac{d}{dt}v(t)$$

or

$$v(t) = \frac{1}{C}\int_0^t i(\tau)\, d\tau$$

From the Laplace transform pair in Eqn. 7.37 the *s*-domain version of this relationship is:

$$V(s) = \frac{1}{Cs}\, I(s)$$

Where the *impedance* $(Z)$ of the capacitor is $1/sC$ and otherwise this can be recognized as $V = IR$ with $R$ replaced by $1/sC$.

Consequently, all three of these basic components reduce to simple impedances in the *s*-domain where each is defined by its *impedance* $Z$ and $V(s) = ZI(s)$. Several circuit configuration will now be examined to explore the advantages offered by working in the *s*-domain for circuit analysis.

## 7.8.1 Voltage Divider

Figure 7.23 illustrates a simple system composed of two resistors $(R1, R2)$. A voltage signal $v_{IN}(t)$ input on the left-hand side presents a voltage across the two resistors $(R1 + R2)$. The voltage signal $v_{OUT}(t)$ seen on the right-hand side is given by the voltage across the resistor $R2$. This circuit configuration can be identified as a *voltage divider* as the input voltage is divided across the two resistors in proportion to their relative resistances. If a voltage source $v(t)$ is connected to the input as shown, then the current at this input is $i(t) = v(t)/(R1 + R2)$. In turn, the voltage across $R1$ is given by $v_{R1}(t) = R1\, i(t) = v(t) \times R1/(R1 + R2)$, while the voltage across $R2$ is given by $v_{R2}(t) = R2\, i(t) = v(t) \times R2/(R1 + R2)$.

## 7.8.2 A First-Order Lowpass Filter

Figure 7.24 is a system that includes one resistor and one capacitor. This circuit may be recognized in the *s*-domain as the voltage divider of Figure 7.23 with $R2$ replaced by a capacitor.

In the *s*-domain, the voltage across $R$ is given by $V_R(s) = RI(s) = V(s) \times R/(R + 1/sC)$. The voltage across the capacitor is given by:

$$V_C(s) = V_{IN}\left[\frac{\frac{1}{sC}}{R + \frac{1}{sC}}\right]$$

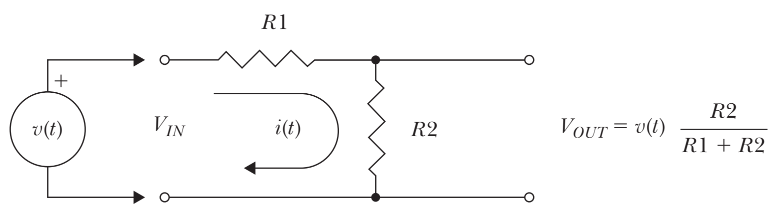



**FIGURE 7.23 A Voltage Divider** Two resistors in series carry the same current $i(t)$, and the voltage drop across each sums to the input voltage $v(t)$. The output voltage $V_{OUT}$ is a fraction of $V_{IN}$ as given by the ratio $\frac{R2}{R1+R2}$.

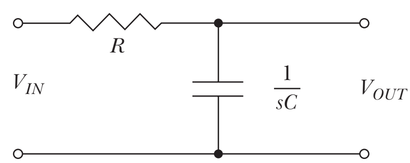



**FIGURE 7.24 A Lowpass Filter** A resistor $(R)$ and capacitor $(C)$ form a lowpass filter designed to impede high-frequency signal components. In the *s*-domain, this circuit is recognized as the *voltage divider* from Figure 7.23 using a capacitor in place of $R2$. The output voltage $V_{OUT}$ is a fraction of $V_{IN}$ as given by the ratio $\frac{1/sC}{R+1/sC}$.

Multiplying by $s/R$ top and bottom gives:

$$\frac{V_C(s)}{V_{in}} = \left[\frac{\frac{1}{RC}}{s + \frac{1}{RC}}\right]$$

in which the term in the denominator has $s$ alone. This is the $s$-domain transfer function $H(s)$ of the RC circuit, obtained directly from the circuit diagram. For simplicity, let $a = 1/RC$ (it is certain that $a > 0$) and:

$$\frac{V_C(s)}{V_{in}} = \left[\frac{a}{s + a}\right]$$

The corresponding pole-zero diagram in Figure 7.25 has a single pole at $(\sigma = -1/RC, \omega = 0)$ and no zeros.

As previously shown in Figure 7.10, this single pole defines a region of convergence that includes $\sigma = 0$, and therefore it is legitimate to continue and to determine the frequency response by setting $\sigma = 0$, giving:

$$H(j\omega) = \left[\frac{a}{j\omega + a}\right]$$

The goal is now to plot the magnitude and phase response of this $H(s)$. It will be shown that these graphs can be sketched by hand with little or no calculation once the $s$-domain transfer function $H(j\omega)$ is known.

Two important observations should always be made at this point. By checking the limiting conditions as $(\omega \to 0)$ and as $(\omega \to \infty)$, the overall behavior of the $|H(\omega)|$ and $\Phi(\omega)$ can be determined. These limiting values are generally easy to determine by inspection, and they serve to provide two key reference points to understand both $|H(\omega)|$ and $\Phi(\omega)$.

**1.** Let $\omega \to 0$ to determine the limiting response at zero frequency. In this case:

$$|H(\omega \to 0)| = \left[\frac{a}{0 + a}\right] = 1$$

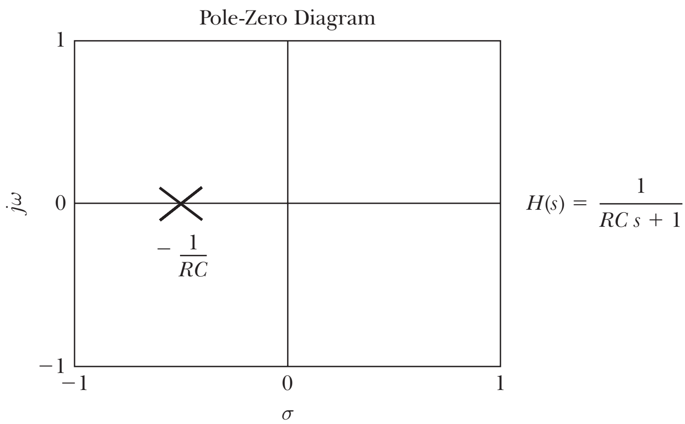



**FIGURE 7.25 First-Order Lowpass Filter Pole-Zero Diagram** The resistor $(R)$ and capacitor $(C)$ in Figure 7.24 form a lowpass filter designed to impede high-frequency signal components. In the $s$-domain, the transfer function $H(s)$ describes a single pole at $s = -1/RC$ (from the denominator).

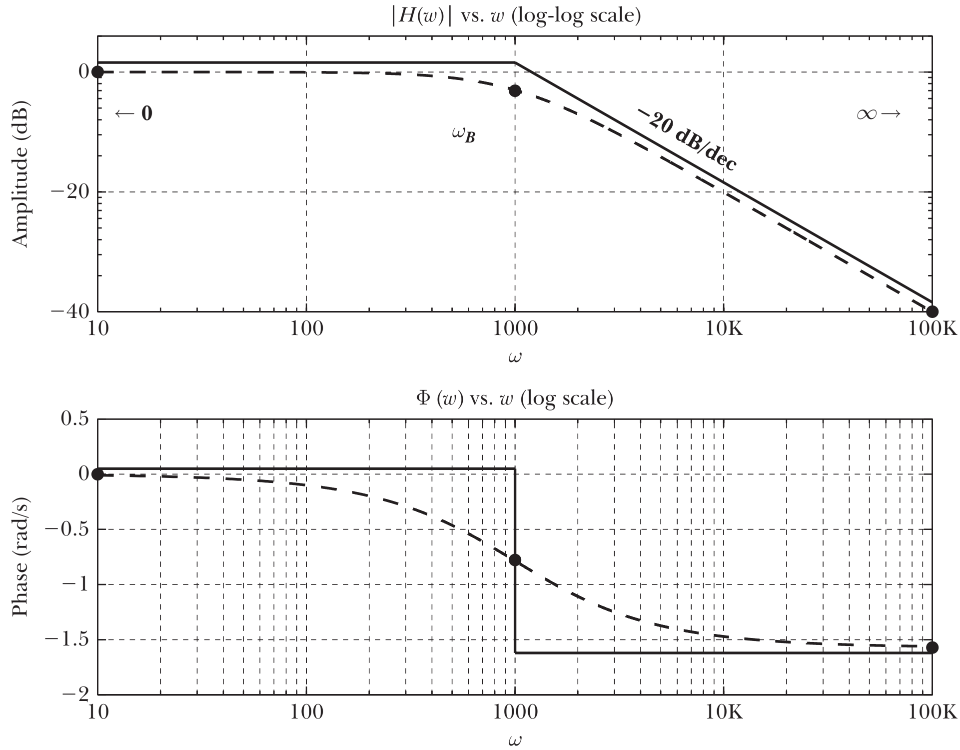



**FIGURE 7.26 Sketching the Lowpass Filter** The reference points shown are the limit as $(\omega \to 0)$, $(\omega \to \infty)$, and the computed breakpoint position, shown here at $(\omega = 1000 \text{ rad/s})$. The dashed line on the upper plot shows magnitude $|H(\omega)|$ on a log-log graph. The dashed line on the lower graph shows the phase response $\Phi(\omega)$ on a log-linear scale. The corresponding straight-line approximations are sketched slightly offset from each line for clarity.

showing that low-frequency components near zero pass unaffected. Also,

$$\Phi(\omega \to 0) = tan^{-1}\ (0) = 0$$

2. Let $\omega \to \infty$ to determine the limiting response at infinite frequency.

$$|H(\omega \to \infty)| = \left[\frac{a}{\infty + a}\right] = 0$$

as the denominator goes to infinity, the overall function tends to zero. This shows that high-frequency components are attenuated.

$$\Phi(\omega \to \infty) = tan^{-1}(-\infty) = -\frac{\pi}{2}$$

The above observations confirm that this circuit behaves as a lowpass filter, at least at the extremes. Incoming frequencies undergo a phase change of up to $-\pi/2\ rad/s = -90\ degrees$ at the highest frequencies.[4]

Finally, the pole position provides a third reference point to aid in sketching the frequency response of this filter. It can be noted that at the pole position $(\omega = a)$:

[4] Confirm this as for example, $tan^{-1}(-999999999)$ using a calculator.

$$|H(j\omega)| = \left[\frac{a^2}{\omega^2 + a^2}\right]^{\frac{1}{2}} = \frac{1}{\sqrt{2}} \quad for \;\; \omega = a$$

showing that an incoming frequency component $s(t)$ at this $\omega$ is reduced to amplitude $s(t)/\sqrt{2} = 0.707s(t)$. Also,

$$\Phi(\omega = a) = tan^{-1}(-1) = -\frac{\pi}{4}$$

For this filter, the squared output signal is reduced to one half at the pole frequency. This point is identified as the *half-power point* and defines the *cutoff frequency* for this filter.

Let $R = 10$ k$\Omega$ and $C = 0.1\,\mu$F. The breakpoint frequency is $\omega_B = 1/RC = 1000$ rad/s $\approx 160$ Hz as seen in Figure 7.26.

## 7.8.3 A First-Order Highpass Filter

Figure 7.27 is a system that includes one resistor and one inductor. This circuit may be recognized in the *s*-domain as the voltage divider of Figure 7.23 with $R2$ replaced by an inductor. Analysis proceeds as with the lowpass filter.

In the *s*-domain, the voltage across $R$ is given by $V_R(s) = RI(s) = V(s) \times R/(R + sL)$. The voltage across the inductor is given by:

$$V_C(s) = V_{IN}\left[\frac{sL}{R + sL}\right]$$

Multiplying by $1/L$ top and bottom gives:

$$\frac{V_C(s)}{V_{in}(s)} = \left[\frac{s}{s + \frac{R}{L}}\right]$$

in which the term in the denominator has $s$ alone. This is the *s*-domain transfer function $H(s)$ of the $RL$ circuit, obtained directly from the circuit diagram. The corresponding pole-zero diagram has a single pole at $(\sigma = -R/L, \omega = 0)$ and a zero at the origin $(\sigma = 0, \omega = 0)$. For simplicity, let $a = R/L$ (it is certain that $a > 0$),

$$\frac{V_C(s)}{V_{in}(s)} = \left[\frac{s}{s + a}\right]$$

The general form of $H(s)$ may be examined as:

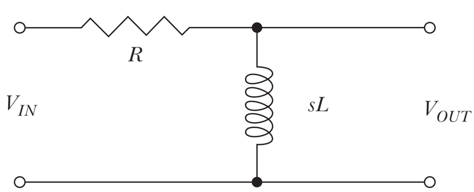



**FIGURE 7.27 A Highpass Filter** A resistor ($R$) and inductor ($L$) form a highpass filter designed to impede low-frequency signal components. In the *s*-domain, this circuit is recognized as the *voltage divider* from Figure 7.23 using an inductor in place of $R2$. The output voltage $V_{OUT}$ is a fraction of $V_{IN}$ as given by the ratio $\frac{sL}{R+sL}$.

$$H(s) = k\left[\frac{s}{s+a}\right], \quad \textit{with } k = 1$$

As in Figure 7.10, this single pole defines a region of convergence that includes $\sigma = 0$, and therefore it is legitimate to continue and to determine the frequency response by setting $\sigma = 0$.

$$H(j\omega) = k\left[\frac{j\omega}{j\omega + a}\right], \quad \textit{with } k = 1$$

## 7.8.4 A Second-Order Filter

Figure 7.27 includes one resistor and one capacitor and one inductor. This circuit may be recognized in the *s*-domain as a voltage divider with the output voltage measured across the capacitor. The circuit may also be configured to place the resistor or inductor at the output, as described below.

**Lowpass Filter** The voltage across the capacitor is given by:

$$V_{out} = V_{in}\left[\frac{\frac{1}{Cs}}{Ls + R + \frac{1}{Cs}}\right] \tag{7.60}$$

multiplying the numerator and denominator by $Cs$ gives:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{1}{LCs^2 + RCs + 1} \tag{7.61}$$

As a check, confirm that for $L \to 0$ (inductor replaced by a wire) this circuit and this $H(s)$ reduce to the first-order lowpass filter circuit discussed in Section 7.8.2. When $L = 0$, the term in $s^2$ goes to zero, leaving only the expected result.

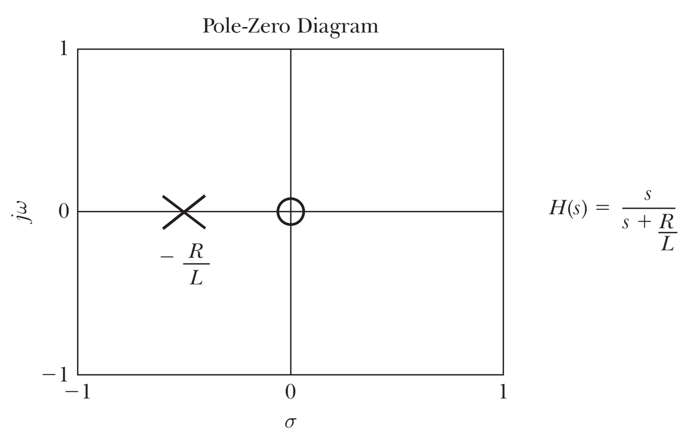



**FIGURE 7.28 First-Order Highpass Filter Pole-Zero Diagram** The resistor ($R$) and inductor ($L$) in Figure 7.27 form a highpass filter designed to impede low-frequency signal components. In the *s*-domain, the transfer function $H(s)$ describes a single pole at $s = -R/L$ (from the denominator) and a zero at $s = 0$ (from the numerator).

By inspection, this $H(s)$ has overall frequency-dependent properties as:

1. For low frequencies, as $s \to 0$, $H(s)$ goes to 1. The system has unit gain at DC.
2. For high frequencies, $H(s)$ goes down as $1/s^2$. In contrast, a first-order lowpass filter goes down as $1/s$.

This is identified as a *second-order system* from the presence of $s^2$ as the highest order of $s$ in the denominator. Overall this configuration acts as a lowpass filter that falls off with frequency at twice the rate of the first-order filter, as shown in Figure 7.31A. The exact behavior and the breakpoint frequencies can be very different depending on the denominator polynomial and the specific component values found in the circuit. The form of the numerator and denominator indicate that the pole-zero diagram will have two poles and no zeros.

**Bandpass Filter** The voltage across the resistor is given by:

$$V_{out} = V_{in}\left[\frac{R}{Ls + R + \frac{1}{Cs}}\right] \tag{7.62}$$

multiplying the numerator and denominator by $Cs$ gives:

$$H(s) = \frac{V_{out}}{V_i n} = \frac{RCs}{LCs^2 + RCs + 1} \tag{7.63}$$

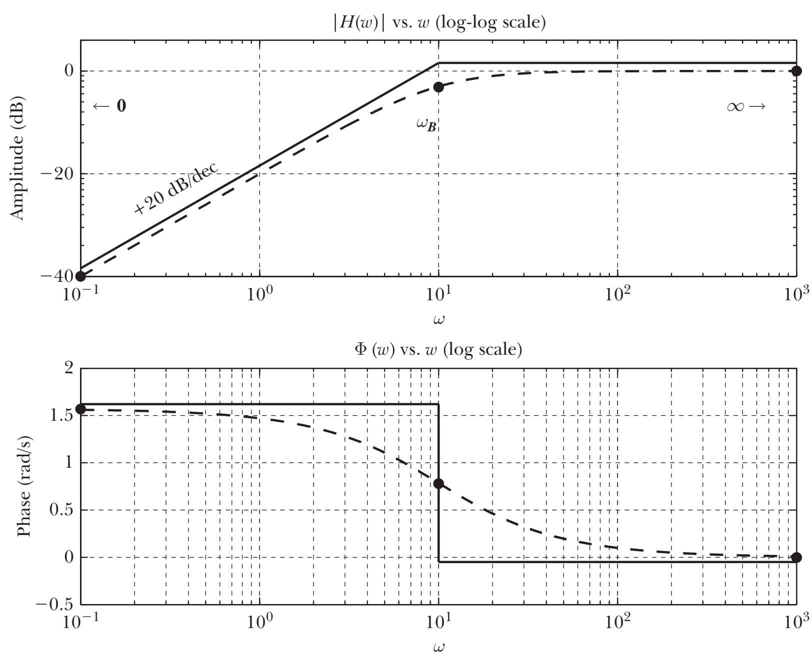



**FIGURE 7.29 Sketching the Highpass Filter** The reference points shown are the limit as $(\omega \to 0)$, $(\omega \to \infty)$, and the computed breakpoint position, shown here at ($\omega_B = 10$ rad/s). The dashed line on the upper plot shows magnitude $|H(\omega)|$ on a log-log graph. The dashed line on the lower graph shows the phase response $\Phi(\omega)$ on a log-linear scale. The corresponding straight-line approximations are sketched slightly offset from each line for clarity.

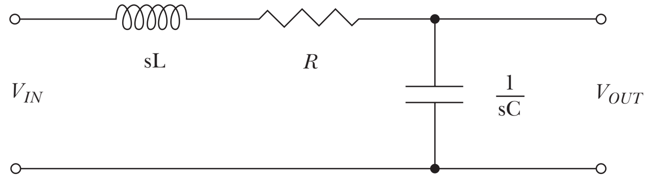



**FIGURE 7.30 A Second-Order Filter** A series resistor and inductor and capacitor form a lowpass filter designed to impede high-frequency signal components. In the *s*-domain, this circuit is recognized as the *voltage divider* from Figure 7.23 with the output voltage measured across the capacitor.

By inspection, this $H(s)$ has overall frequency-dependent properties as:

1. For low frequencies, as $s \to 0$, $H(s)$ goes to 0. The system has zero gain at DC.
2. For high frequencies, the highest terms in $s$ dominate and $H(s)$ goes down as $1/s$.

Overall this configuration acts as a bandpass filter that attenuates both low frequencies and high frequencies, as shown in Figure 7.31B. The exact behavior and the breakpoint frequencies can be very different depending on the denominator polynomial and the specific component values found in the circuit. The pole-zero diagram will have two poles. There is also a zero located at the origin where $s = 0$.

**Highpass Filter** The voltage across the inductor is given by:

$$V_{out} = V_{in}\left[\frac{Ls}{Ls + R + \frac{1}{Cs}}\right] \tag{7.64}$$

multiplying the numerator and denominator by $Cs$ gives:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{LCs^2}{LCs^2 + RCs + 1} \tag{7.65}$$

As a check, confirm that for $C \to \infty$ (capacitor replaced by a wire) this circuit and this $H(s)$ reduce to the first-order highpass filter circuit discussed in Section 7.8.3. As $C$ becomes large, the constant term in the denominator vanishes and the values of $C$ cancel, leaving the expected result.

By inspection, this $H(s)$ has overall frequency-dependent properties as:

1. For low frequencies, as $s \to 0$, $H(s)$ goes to 0. The system has zero gain at DC.
2. For high frequencies, the highest terms in $s$ dominate and $H(s)$ goes to 1.

Overall, this configuration acts as a highpass filter that attenuates low frequencies, as shown in Figure 7.31C. The exact behavior and the breakpoint frequencies can be very different depending on the denominator polynomial and the specific component values found in the circuit. The pole-zero diagram will have two poles. There is also a double zero located at the origin where $s = 0$.

Figure 7.31 summarizes the overall frequency response of the three filters corresponding to voltages measured across the capacitor, resistor, and inductor of the series RLC circuit. Ideally, all three curves coincide at the double breakpoint shown highlighted ($-6$ dB at $\omega = 1$ rad/s), where each Bode approximation line changes slope by $-40$ dB/dec. In practice, the behavior around this breakpoint frequency may vary considerably as discussed in the next section.

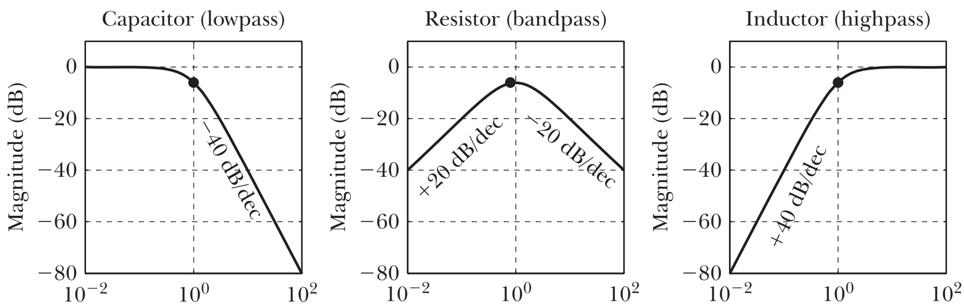



**FIGURE 7.31 Second-Order Filters** The expected overall frequency response for the second-order filters is shown, corresponding to voltages measured across the capacitor, resistor, and inductor of the series RLC circuit. In each case, the line slope changes by −40 dB/dec at the breakpoint.

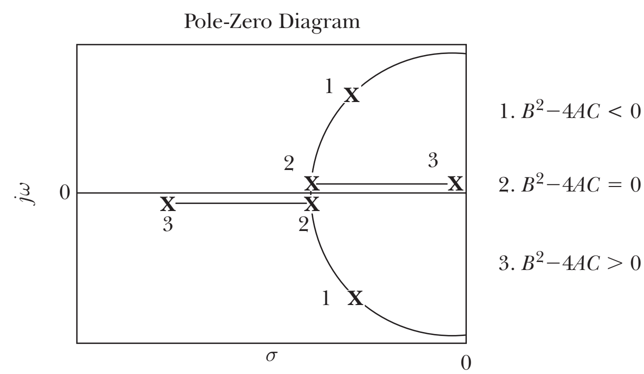



**FIGURE 7.32 Second-Order Pole Positions** In general, pole positions for the second-order filter depend on the roots of the quadratic equation $As^2 + Bs + C$. In this graph, $B$ is varied for constant $(A, C)$ to show three distinct cases where the roots reflect (1) complex paired poles, (2) a double pole, or (3) two poles on the horizontal axis. The corresponding impulse responses are shown in Figure 7.33.

**Analysis of a Second-Order System** All three circuits configurations above are identified as *second-order systems* from the presence of $s^2$ as the highest order in $s$ in the denominator. They all have the same denominator and therefore the same poles. In general, the denominator of a second-order system $H(s)$ is a quadratic equation $As^2 + Bs + C$ with two roots given by:

$$s = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}$$

where the terms $(A, B, C)$ are all real positive values derived from the real circuit component values $(R, L, C)$. This observation ensures that the value under the square root is always less than $B^2$ and therefore, the square root is always less than $B$, such that $Re(s) < 0$ and two poles must be located on the left half plane. Since the square root yields imaginary terms whenever $B^2 - 4AC$ is negative, the overall form of these roots and the corresponding pole positions now depends on the square root expression for three important cases labelled (1, 2, 3), as shown below and in Figure 7.32.

1. For $B^2 - 4AC < 0$, the roots are a pair of complex values. The poles are located above and below the real horizontal axis at positions:

$$s_1 = \frac{-B}{2A} - j\frac{\sqrt{4AC - B^2}}{2A} \quad \text{and} \quad s_2 = \frac{-B}{2A} + j\frac{\sqrt{4AC - B^2}}{2A}$$

where the value under the square root has been multiplied by $-1$ and the resulting $j$ is brought out in front. In the limit as $B \to 0$, these poles become:

$$s_1 = -j\sqrt{\frac{C}{A}} = -j\omega_R, \quad \text{and} \quad s_2 = +j\sqrt{\frac{C}{A}} = +j\omega_R$$

where $\omega_R$ is called the *undamped resonance frequency*. On a Bode plot, these poles correspond to a double breakpoint at frequency $\omega_B = \omega_R$, the same as for Case 2 below.

2. For $B^2 - 4AC = 0$, the term under the square root is zero. Two poles are located directly on the real axis at the same position:

$$s_1 = -B/2A, \quad \text{and} \quad s_2 = -B/2A$$

or, since $B^2 = 4AC$,

$$s_1 = -\sqrt{\frac{C}{A}}, \quad \text{and} \quad s_2 = -\sqrt{\frac{C}{A}}$$

On a Bode plot, these poles correspond to a double breakpoint at frequency $\omega_B = \omega_R$, the same as Case 1 above.

3. For $B^2 - 4AC > 0$, the roots are a pair of real values. Two poles are located directly on the real axis at positions:

$$s_1 = \frac{-B}{2A} - \frac{\sqrt{B^2 - 4AC}}{2A} \quad \text{and} \quad s_2 = \frac{-B}{2A} + \frac{\sqrt{B^2 - 4AC}}{2A}$$

where notably theses poles are real valued. On a Bode plot, these poles correspond to two breakpoints at frequencies $(s_1, s_2)$ corresponding to these two pole positions.

**TABLE 7.1**

**The *quality factor* (Q) and the *damping factor* ($\zeta$) are measures of system performance defined by the roots of the polynomial denominator of *H*(*s*). The two terms are inversely related such that high Q implies low $\zeta$.**

| | **CASE 1** $(B^2 - 4AC) < 0$ | **CASE 2** $(B^2 - 4AC) = 0$ | **CASE 3** $(B^2 - 4AC) > 0$ |
|---|---|---|---|
| | two complex poles | double real pole | two real poles |
| $\zeta = \sqrt{\frac{B^2}{4AC}}$ | $0 < \zeta < 1$ | $\zeta = 1$ | $\zeta > 1$ |
| $Q = \frac{1}{2\zeta}$ | $Q > \frac{1}{2}$ | $Q = \frac{1}{2}$ | $0 < Q < \frac{1}{2}$ |
| | Underdamped | Critically Damped | Overdamped |

These three conditions yield different results and are fundamental descriptors of system performance that are found directly in related system parameters. For each of the above conditions, the corresponding fraction $B^2/4AC$ is less than 1, equal to 1, or greater than 1, and its square root is a positive real number $\zeta$ called the *damping factor*.[5]

$$\zeta = \sqrt{\frac{B^2}{4AC}} \tag{7.66}$$

A related measure of the damping or resonant tendencies of a system is called the *quality factor* or *Q-factor* where $Q = 1/2\zeta$, or

$$Q = \frac{1}{2}\sqrt{\frac{4AC}{B^2}} = \sqrt{\frac{AC}{B^2}} \tag{7.67}$$

in which case, $Q$ is greater than 1/2, equal to 1/2, or less than 1/2 for each of the above conditions, as summarized in Table 7.1.

These pole positions define the form of the impulse response for each of the three conditions, as shown in Figure 7.33.

For $Q > 1/2$ (Case 1), it can be observed that the complex poles represent a sinusoidal impulse response that decreased exponentially over time as shown in Figure 7.33(1). This diminishing amplitude is called *damping*. As $B$ decreases (and $Q$

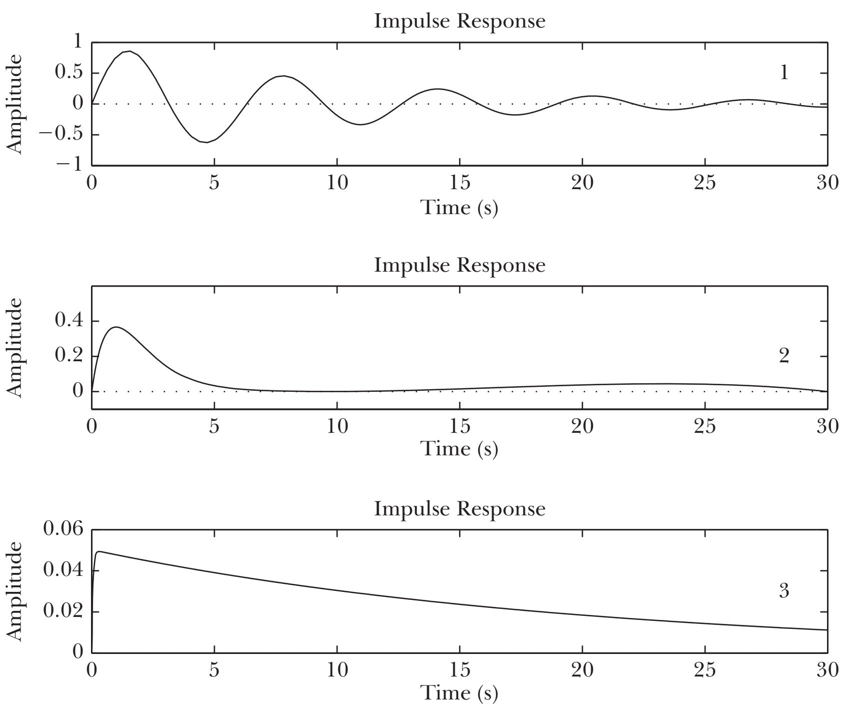



**FIGURE 7.33 Second-Order Impulse Reponses** The impulse response for the second-order lowpass filter is shown for various pole positions (1) for $Q = 5$, (2) $Q = 0.5$, and (3) $Q = 0.05$ as described in Figure 7.32.

[5] The choice of the lowercase Greek letter *zeta* ($\zeta$) follows common usage.

increases), the tendency to oscillate grows as the poles move closer to the vertical axis; this movement is accompanied by an increasingly resonant (peak) behavior in the frequency response $|H(j\omega)|$. A higher $Q$ implies a sharper resonance peak. In the limit as $B \rightarrow 0$, the poles lie directly on the vertical axis, and the *undamped natural frequency* is identified where the system would oscillate forever with constant amplitude:

$$s = \frac{0 \pm j\sqrt{0 + 4AC}}{2A} = \pm j\sqrt{\frac{C}{A}} = \pm j\omega_R$$

therefore:

$$\omega_R = \sqrt{\frac{C}{A}} \tag{7.68}$$

For $Q = 1/2$ (Case 2), it can be observed that the poles move to meet at the same position on the horizontal axis when the square root is zero (2) and:

$$s = \frac{-B}{2A} \tag{7.69}$$

such that the breakpoint is found at $\omega_B = B/2A$. At this breakpoint, the signal level is down $-6dB$ for this double pole. With poles on the horizontal axis, there is no oscillatory component to the impulse response, as shown in Figure 7.33(2). The condition $Q = 1/2$ or $\zeta = 1$ represents an ideal of sorts in which the impulse response rises and returns to zero rapidly and with no oscillation.

For $Q < 1/2$ (Case 3), two separate poles lie on the horizontal axis. The impulse response as shown in as shown in Figure 7.33(3) has no oscillatory component, and its return to zero is governed by the longest time constant (the largest pole position $\sigma$). This is the overdamped condition.

In a second-order *mechanical* system, the impulse responses of Figure 7.33 may be associated with a swinging door that can be pushed open from either side but closes itself automatically using a spring mechanism. In Case 1 ($Q > 1/2$, underdamped), if the door is kicked open (impulse input), it may first open wide and then rapidly swing back and forth several times before finally settling in the closed position. In Case 3 ($Q > 1/2$, overdamped), the opened door may very slowly swing back to the closed position over several seconds. Finally, in Case 2 ($Q = 1/2$, ideally damped), the door quickly returns to the closed position with no perceptible oscillations. A door designer would aim for $Q = 1/2$ in the quest for perfect performance.

**Series RLC Circuit Analysis** In the series RLC circuit of Figure 7.30 described above, the common denominator $LCs^2 + RCs + 1$ in the lowpass, bandpass, and highpass filters corresponds to the general second-order equation $As^2 + Bs + C$, with $A = LC$, $B = RC$, $C = 1$. Therefore, each filter has the same poles and the same resonant frequency and differs only in the zeros of the transfer function. From the above analysis, the resonant frequency for Cases 1 and 2 equals $\omega_R = \sqrt{\frac{C}{A}}$ or with this circuit:

$$\omega_R = \sqrt{\frac{C}{A}} \rightarrow \omega_R = \sqrt{\frac{1}{LC}}$$

$$\zeta = \sqrt{\frac{B^2}{4AC}} \rightarrow \zeta = \sqrt{\frac{(RC)^2}{4(LC)}} = \frac{R}{2}\sqrt{\frac{C}{L}}$$

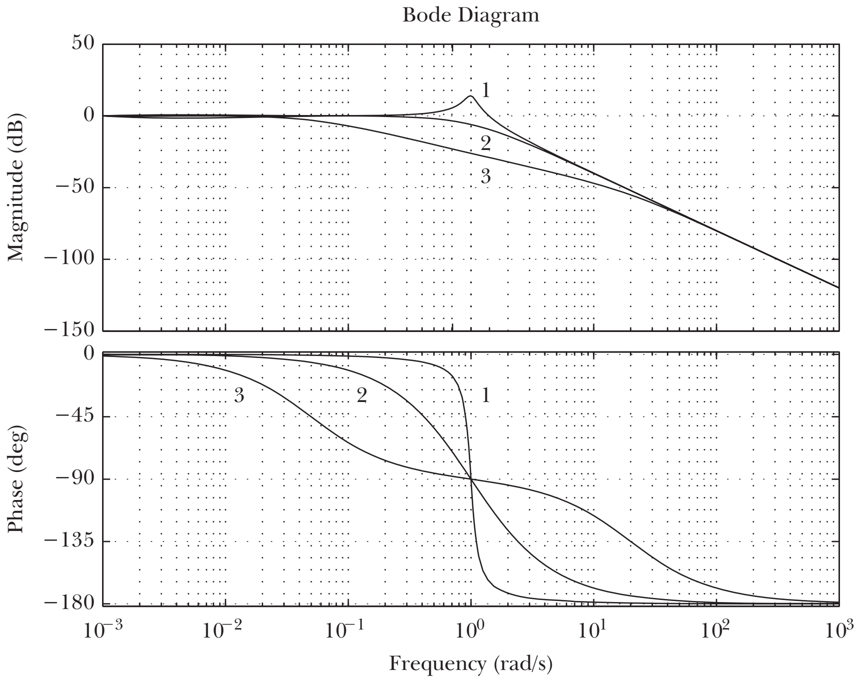



**FIGURE 7.34 Second-Order Bode Plots** The Bode plots showing frequency and phase response for the second-order lowpass filter are shown for various pole positions (1) for $Q = 5$, (2) $Q = 0.5$, and (3) $Q = 0.05$, as described in Figure 7.32. All three curves coincide at the extremes, with unit gain at the origin and the high-frequency slope being $-40$ dB/dec. A resonance point is evident in curve (1).

$$Q = \frac{1}{2\zeta} = \frac{1}{R}\sqrt{\frac{L}{C}}$$

In Figure 7.30, the output voltage (the voltage across the capacitor) is described by a second-order lowpass filter as Eqn. 7.61, or:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{1}{LCs^2 + RCs + 1}$$

which can be simplified for demonstration using the component values $L = 1$, $C = 1$, leaving $H(s)$ as a function of $R$:

$$H(s) = \frac{V_{out}}{V_{in}} = \frac{1}{s^2 + Rs + 1}$$

in which the $Q$-factor depends only on the resistance $R$. The resonance frequency is $\omega_R = 1$ rad/s. The Bode plot shown in Figure 7.34 shows the effect of different Q-values by varying $R$ corresponding to the three cases above.

## 7.9 State Variable Analysis

The study of complex systems can be simplified using a powerful matrix-based method known as *state variable analysis*. Modelling a system in this way lends itself well to numerical computations and the use of MATLAB tools to study the system's

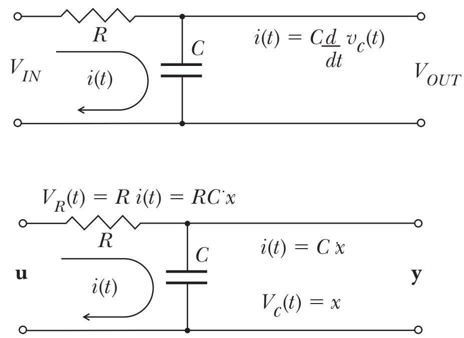



**FIGURE 7.35 State Variable Analysis RC** A series resistor and capacitor form a first-order lowpass filter. The single energy storage component (capacitor) containing the differential voltage term is the focus of state variable analysis. This circuit can be analyzed (bottom) by defining a *state variable* named $x$ describing the voltage across the capacitor.

behavior. In this technique, a system is described using a number of first-order differential equations, regardless of how complex the system may be overall. To this end, independent differential elements are represented by *state variables,* which represent the changing state of the system over time. In practice, the state variable approach can be used for many systems including nonlinear systems; that is, systems having initial conditions and those having multiple inputs and multiple outputs. The method is first illustrated using a first-order RC filter circuit, then with a second-order series RLC lowpass circuit from the previous section, after which it may be generalized to other systems.

## 7.9.1 State Variable Analysis—First-Order System

In state variable analysis, differential elements or energy storage components are modelled using a general *state variable,* often labelled $x$ regardless of whether $x$ represents a time-varying current or voltage or some other quantity, which may or may not be observable in the actual system. Similarly, inputs are generally labelled $\mathbf{u}$ and outputs labelled $\mathbf{y}$. Consider the simple RC circuit of Figure 7.35, where the fundamental differential equation describing the current and voltage relationship in the capacitor is shown. Let the voltage across the capacitor be the state variable $x$, then the current $i(t)$ through the capacitor (and the resistor) is given by $C\dot{x}$, where $\dot{x}$ is the time derivative of $x$. The voltage across the resistor $V_R = Ri(t)$ is then $V_R = RC\dot{x}$. The circuit is shown in state variable form in the lower part of Figure 7.35. The voltage around the first loop gives:

$$\mathbf{u} = v(t) = V_C + V_R = x + RC\dot{x} \tag{7.70}$$

which can be written as:

$$\dot{x} = \frac{-1}{RC}x + \frac{1}{RC}\mathbf{u} \tag{7.71}$$

This equation summarizes the dynamics of the system and includes the input $\mathbf{u}$; however, it does not describe the output signal $\mathbf{y}$, which is given by a separate equation as the voltage across the capacitor, or:

$$\mathbf{y} = V_C = x \tag{7.72}$$

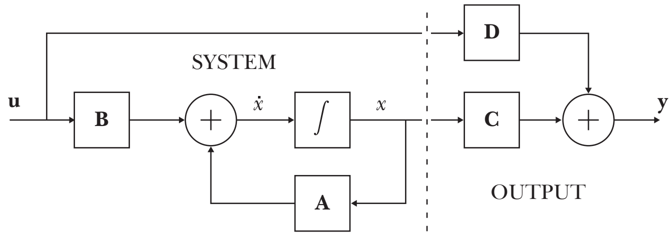



**FIGURE 7.36 State Variable Block Diagram—General Form** By assigning state variables $\dot{x}$ and $x$ at the input and output of the central integrator block, the system state equation can be identified here as $\dot{x} = \mathbf{A}x + \mathbf{Bu}$, while the output state equation is $\mathbf{y} = \mathbf{C}x + \mathbf{Du}$.

These two equations completely describe this first-order system and in this form can be readily generalized to more complex systems.

The general form of the state variable representation corresponds to the system block diagram Figure 7.36, where the variables $(A, B, C, D)$ identify four separate gain (multiplier) blocks. In this figure, the system definition blocks and the output connections are shown separated by a dashed line. The figure is built around an integrator block with input $\dot{x}$ and output $x$. Observe that the term **D** directly connects the input and the output, and in particular, if the other terms are zero, the system describes a simple amplifier with gain **D**; in this case, as in many systems of interest, $\mathbf{D} = 0$. These two sections of the block diagram correspond to the equations:

$$\dot{x} = \mathbf{A}x + \mathbf{Bu} \tag{7.73}$$

and

$$\mathbf{y} = \mathbf{C}x + \mathbf{Du} \tag{7.74}$$

The first-order circuit of Figure 7.35 is obtained when $\mathbf{A} = \frac{-1}{RC}$, $\mathbf{B} = \frac{1}{RC}$, $\mathbf{C} = 1$, and $\mathbf{D} = 0$ as found in Eqn. 7.71 and Eqn. 7.72.

## 7.9.2 First-Order State Space Analysis with MATLAB

System modelling using state variables is supported by the MATLAB *Control System Toolbox* using the function `ss(a,b,c,d)`, which maps the four state system terms into a system object. From the first-order circuit of Figure 7.35, let $R = 100\Omega$ and $C = 100\mu F$, then $a = -100$, $b = 100$, $c = 1$, and $d = 0$.

```
                        % define the state matrices (A,B,C,D)
a = -100;
b = 100;
c = 1;
d = 0;
sys = ss(a, b, c, d);% define the system
tf(sys)                 % find the corresponding transfer function
Transfer function:
  100
-------
s + 100
bode(sys)               % see the Bode plot for this system
```

State variable analysis provides a systematic way to describe a system and to perform numerical calculations. This transfer function was obtained directly from the terms $(a, b, c, d)$ in the standard state equations. Once the system object is defined in MATLAB, the Bode plot (not shown) obtained with `bode(sys)` will confirm that this is a first-order lowpass filter. Other available functions include `pzmap(sys)` to see the pole-zero diagram, `impulse(sys)` to see the impulse response, and `step(sys)` to see the step response, as seen in Section 7.4.6 where `sys` was defined either from the transfer function or from the poles and zeros.

From the circuit in Figure 7.35, the lowpass filter is obtained when the output is measured across the capacitor; however, if the output is instead taken across the resistor, then a first-order *highpass* filter should be obtained. By inspection, the resistor voltage is given by the difference between the input voltage and the voltage across the capacitor, $V_R = v(t) - V_C$, or

$$V_R = \mathbf{y} = \mathbf{u} - x \tag{7.75}$$

which can be identified in the standard output equation $\mathbf{y} = \mathbf{C}x + \mathbf{Du}$ with $\mathbf{C} = -1$ and $\mathbf{D} = 1$. Observe that only the output equation is modified; the circuit and the system dynamics are unchanged. The new MATLAB calculation becomes:

```
                        % define the state matrices (a,b,c,d)
a = -100;
b = 100;
c = -1;
d = 1;
sys = ss(a, b, c, d);   % define the system
tf(sys)                 % find the corresponding transfer function
Transfer function:
    s
--------
s + 100

bode(sys)               % see the Bode plot for this system
```

The expected transfer function is obtained. The Bode plot will confirm that this is a first-order highpass filter. This method can now be extended to higher-order systems, where the terms $(a, b, c, d)$ become matrices.

### 7.9.3 State Variable Analysis—Second-Order System

Consider the RLC circuit of Figure 7.37 in which the three components are shown with their fundamental current and voltage relationships. The assignment of state variables to this circuit begins by identifying the two differential components (capacitor and inductor), which serve to define two state variables. Let the first variable $x_i$ be the current through the inductor; this is also the current through the resistor and the capacitor. Let the second state variable $x_v$ be the voltage across the capacitor; this is also the output voltage of the circuit. The specific assignment of state variables will determine the exact form of the following analysis. Let the time derivatives of the variables $(x_i, x_v)$ be represented as $(\dot{x}_i, \dot{x}_v)$, respectively.

In general, the state variables must be chosen such that the total number of state variables will equal the number of independent energy-storage elements (capacitors

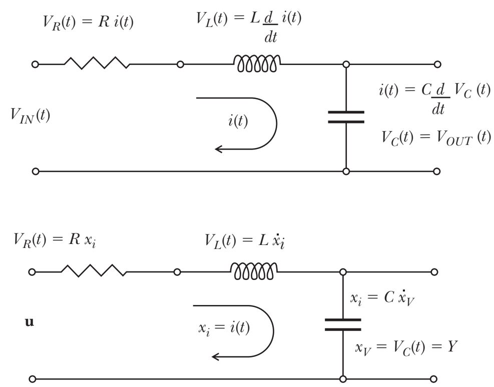



**FIGURE 7.37 State Variable Analysis RLC** A series resistor and inductor and capacitor form a second-order lowpass filter. In state variable representation, this circuit is analyzed through *state variables* describing its energy storage components (capacitor and inductor). In the lower figure, the variable $x_i$ represents the current through the inductor and the variable $x_v$ is the voltage across the capacitor.

and inductors). For a system of order $N$, there will be $N$ state variables, and it is customary to use the labels $(x_1, x_2 \cdots x_N)$; the labels $(x_i, x_v)$ are used in this example for clarity. It is also commonplace to use the symbol **u** to describe the input to a system and the symbol **y** to describe its output. In the lower part of Figure 7.37 the RLC circuit is relabelled in terms of the state variables. In this case, the input $\mathbf{u} = v(t)$ and the voltage around the loop described by the three components is given by:

$$v(t) = \mathbf{u} = L\dot{x}_i + Rx_i + x_v \tag{7.76}$$

also,

$$C\dot{x}_v = x_i \tag{7.77}$$

Rearranged in terms of $(\dot{x}_i, \dot{x}_v)$, these two equations become:

$$\dot{x}_i = -\frac{R}{L}x_i - \frac{1}{L}x_v + \frac{1}{L}\mathbf{u} \tag{7.78}$$

and

$$\dot{x}_v = \frac{1}{C}x_i \tag{7.79}$$

These two first-order differential equations now describe this second-order linear system. A primary goal of state variable analysis is to reduce high-order differential equations to multiple first-order equations, one for each state variable. Another goal is to express these equations in matrix form as:

$$\begin{bmatrix} \dot{x}_v \\ \dot{x}_i \end{bmatrix} = \begin{bmatrix} 0 & \dfrac{1}{C} \\ -\dfrac{1}{L} & -\dfrac{R}{L} \end{bmatrix} \begin{bmatrix} x_v \\ x_i \end{bmatrix} + \begin{bmatrix} 0 \\ \dfrac{1}{L} \end{bmatrix} \mathbf{u} \tag{7.80}$$

To complete the model, the output **y** is defined as the voltage across the capacitor and shown in matrix form as:

$$\mathbf{y} = x_v = [1 \quad 0] \begin{bmatrix} x_v \\ x_i \end{bmatrix} \tag{7.81}$$

## 7.9.4 Matrix Form of the State Space Equations

The two matrix equations (Eqns. 7.80, 7.81) compactly represent this series RLC circuit in a state variable formulation. This result may be recognized in the standard general form of these two equations as:

$$\dot{x} = \mathbf{A}x + \mathbf{Bu} \tag{7.82}$$

$$\mathbf{y} = \mathbf{C}x + \mathbf{Du} \tag{7.83}$$

where $x$ and $\dot{x}$ are $(2 \times 1)$ column vectors and $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ are identified as the matrices:

$$\mathbf{A} = \begin{bmatrix} 0 & \dfrac{1}{C} \\ -\dfrac{1}{L} & -\dfrac{R}{L} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \dfrac{1}{L} \end{bmatrix}, \quad \mathbf{C} = [1 \quad 0], \quad \mathbf{D} = 0 \tag{7.84}$$

Equation 7.82 describes the system configuration. For a system of order $N$, the matrix **A** is an $(N \times N)$ matrix based on the system parameters, while the matrix **B** is an $(N \times 1)$ matrix that incorporates the system inputs **u**. In turn, Eqn. 7.83 describes the system output **y**. In many physical systems, as in this example, the term **D** equals zero. Observe that these same equations could be reformulated and **u** and **y** defined as column vectors if a system included multiple inputs (sources) and multiple outputs (MIMO). The present discussion is limited to *single input, single output* (SISO) systems.

The formulation of the state equations as a general system model reflects the block diagram of Figure 7.36, where the terms $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ now correspond to matrices describing a high-order feedback integration circuit. Significantly, each term $x_n$ corresponds to a first-order system, such that higher-order systems are described as coupled first-order systems. The block diagram now describes a high-order system in which each state variable defines a separate integrator block and the matrix **A** essentially connects a set of coupled first-order differential equations. The upper path may be omitted when $\mathbf{D} = 0$.

For a second-order system specifically, the block diagram in Figure 7.38 relates to the matrix equations as:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \mathbf{u} \tag{7.85}$$

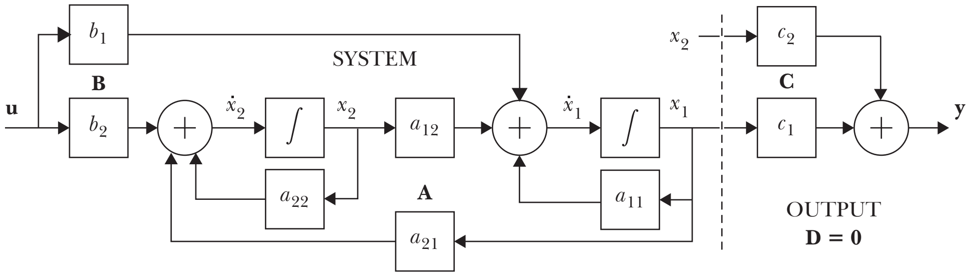



**FIGURE 7.38 Second-Order State Variable Block Diagram** By assigning state variables $\dot{x}$ and $x$ at the input and output of the central integrator block, the system state equation can be identified here as $\dot{x} = \mathbf{A}x + \mathbf{Bu}$, while the output state equation is $\mathbf{y} = \mathbf{C}x + \mathbf{Du}$.

and

$$\mathbf{y} = [\,c_1 \quad c_2\,] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{7.86}$$

with $\mathbf{D} = 0$, which may be omitted.

## 7.9.5 Second-Order State Space Analysis with MATLAB

Higher-order system modelling using state space analyis uses the function `ss(a,b,c,d)` where the state variables are matrices. From the second-order circuit of Figure 7.37, let $R = 100\ \Omega$ and $C = 100\ \mu F$ and $L = 1\ H$, then from Eqn. 7.84:

$$\mathbf{A} = \begin{bmatrix} 0 & 10000 \\ -1 & -100 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{C} = [1 \quad 0], \quad \mathbf{D} = 0 \tag{7.87}$$

In MATLAB, this becomes:

```
                          % define the state matrices (A,B,C,D)
a = [0 10000; -1 -100];
b = [0 ; 1];
cw = [1 0];
d = 0;
sys = ss(a, b, c, d);% define the system
tf(sys)                   % find the corresponding transfer function
Transfer function:
      1e04
-----------------
s^2+ 100 s + 1e04
bode(sys)                 % see the Bode plot for this system
>> pole(sys)              % find the pole positions
ans =
-50.0000 +86.6025i
-50.0000 -86.6025i
```

The transfer function describes a second-order lowpass filter circuit with cutoff frequency $f_c = 100$ rad/s as seen in Figure 7.39 (note the 180-degree phase shift and $-40$ db/decade downward slope). The poles of the system matrix are found with pole(sys) as shown; these correspond to the roots of the transfer function denominator. This could be confirmed in a pole-zero diagram obtained with pzmap(sys).

### 7.9.6 Differential Equation

If state variables $(x_1, x_2 \cdots x_n)$ are defined recursively such that $x_n = \dot{x}_{n-1}$, then any high-order differential equation can be represented by coupled first-order equations. Consider the following example of a second-order ordinary differential equation.

$$2\frac{d^2v(t)}{dt} + 3\frac{dv(t)}{dt} + 5v(t) = s(t) \tag{7.88}$$

Let $x_1 = v(t)$, then $x_2 = \dot{x}_1$ and the equation becomes:

$$2\dot{x}_2 + 3x_2 + 5x_1 = \mathbf{u} \tag{7.89}$$

which defines the two terms $(\dot{x}_1, \dot{x}_2)$, rearranging as:

$$\dot{x}_2 = \frac{1}{2}\mathbf{u} - \frac{3}{2}x_2 - \frac{5}{2}x_1 \tag{7.90}$$

and

$$\dot{x}_1 = x_2 \tag{7.91}$$

and in matrix form, the state equation is:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{5}{2} & -\frac{3}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix} \mathbf{u} \tag{7.92}$$

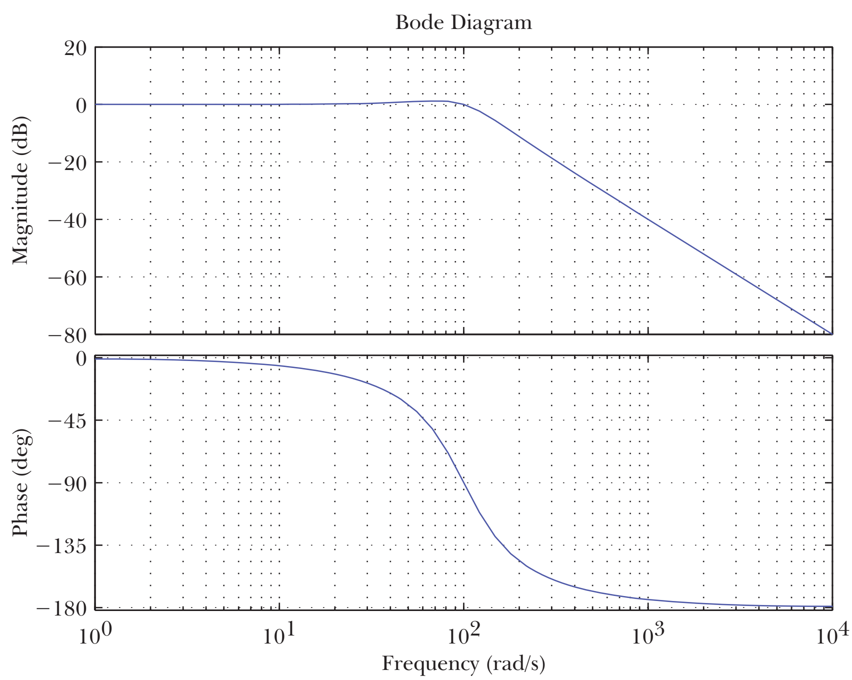



**FIGURE 7.39 Bode Diagram—Second-Order Lowpass Filter** The circuit of Figure 7.37 described by a state space model is solved in MATLAB to produce this Bode plot.

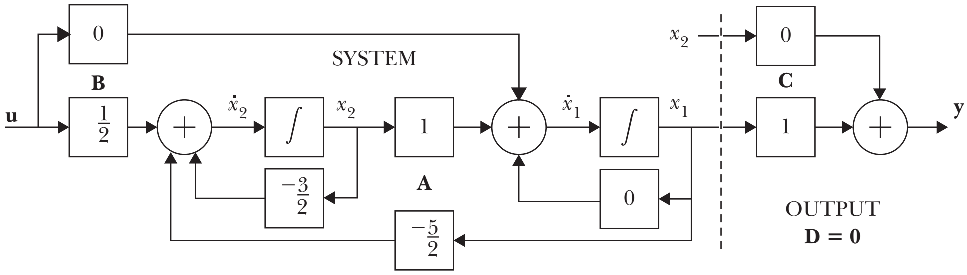



**FIGURE 7.40 State Variable Block Diagram—Second-Order System** By assigning state variables $\dot{x}_2$ and $x_2$ and $\dot{x}_1$ and $x_1$ at the inputs and outputs of the central integrator blocks, the terms in the system state matrix (Eqn. 7.92) and the output system matrix (Eqn. 7.93) can be identified.

and the output equation is given by $s(t) = x_1 = \mathbf{y}$:

$$\mathbf{y} = [1 \quad 0]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{7.93}$$

Summarizing for the general state equations and output equations as shown in Figure 7.40 are:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{5}{2} & -\frac{3}{2} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}, \quad \mathbf{C} = [1 \quad 0], \quad \mathbf{D} = [0] \tag{7.94}$$

By extension, the equation

$$2\frac{d^2v(t)}{dt} + 3\frac{dv(t)}{dt} + 5v(t) = \frac{ds(t)}{dt} + 4s(t) \tag{7.95}$$

differs only on the right-hand side and with reference to Figure 7.40, the state variable matrices will differ only in the output as $\mathbf{C} = [4\ 1]$.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{5}{2} & -\frac{3}{2} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}, \quad \mathbf{C} = [4 \quad 1], \quad \mathbf{D} = [0] \tag{7.96}$$

## 7.9.7 State Space and Transfer Functions with MATLAB

In MATLAB, a system may be described by its state space model, from which the transfer function and pole-zero diagram may be obtained. It is also possible to convert a transfer function into a state space model. For example, given the differential equation of Eqn. 7.95, the matrices from Eqn. 7.96 can be entered in MATLAB to define a system object variable and then obtain the transfer function.

```
                    % define the state matrices (A,B,C,D)
a = [0 1 ; -5/2 -3/2 ];
b = [0 ; 1/2 ];
c = [4 ; 1 ];
```

```
d = [0];                 % this term may also be simply specified as 0
sys = ss(a, b, c, d);    % define the system
tf(sys)                  % find the corresponding transfer function
Transfer function:
    0.5 s + 2
------------------
s^2+ 1.5 s + 2.5
```

This transfer function has been arranged to have a unit multiplier in the highest-order denominator term. MATLAB may also be used to obtain the state space matrices given the transfer function. Because the state variable description is not unique, the results may differ from the above; however, the same transfer function must be obtained in any case. For example, the transfer function of Eqn. 7.95 can be entered directly using the MATLAB function `tf(num,den)`, and the resulting matrices $(a, b, c, d)$ will differ somewhat from the above definitions. To confirm the validity of this result, the original transfer function is found from the state matrices as shown below.

```
              % define transfer function as tf(num, den)
transfer = tf([1 4],[2 3 5])
Transfer function:
     s + 4
-----------------
2 s^2+ 3 s + 5
              % find state space matrices from the transfer function
sys = ss(transfer)
a =
       x1     x2
x1     -1.5   -1.25
x2      2      0
b =
       u1
x1     1
x2     0
c=
       x1     x2
y1     0.5    1
d =
       u1
y1     0
Continuous-time model.
              % check the result, convert back to transfer function
tf(sys)
Transfer function:
    0.5 s + 2
------------------
s^2+ 1.5 s + 2.5
```

As expected, this is the same transfer function obtained in the previous example, although the definitions of $(x_1, x_2)$ and the individual matrices are assigned differently in the MATLAB solution as seen in Figure 7.41. The individual matrices

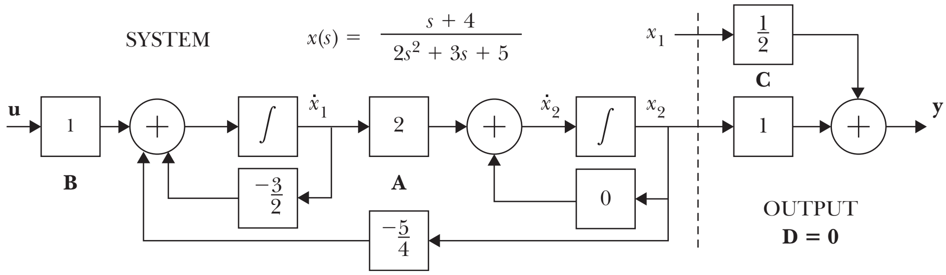



**FIGURE 7.41 State Variables Obtained from a Transfer Function** The values in this state space model were obtained in MATLAB by providing the transfer function. The text solution differs in that the state variables $(x_1, x_2)$ are reordered and the state equations do not contain the same values for $(A, B, C, D)$. A state space representation is not unique.

$(A, B, C, D)$ defining the state space object `sys` can be can be isolated as `sys.a`, `sys.b`, `sys.c`, and `sys.d`, respectively.

# 7.10 Conclusions

The Laplace transform resembles the Fourier transform incorporating a real exponential component. State space modelling is a powerful way of describing a high-order system in matrix form that is readily analyzed using numerical techniques.

# End-of-Chapter Exercises

**7.1** Write the simplified system equation describing the block diagram in Figure 7.42.

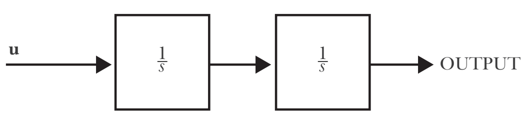



**FIGURE 7.42** Figure for Question 7.1.

**7.2** Write the simplified system equation describing the block diagram in Figure 7.43.

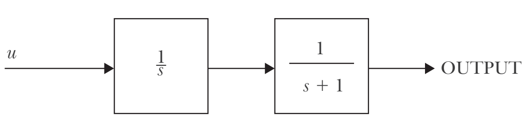



**FIGURE 7.43** Figure for Question 7.2.

**7.3** Write the simplified system equation describing the block diagram in Figure 7.44.

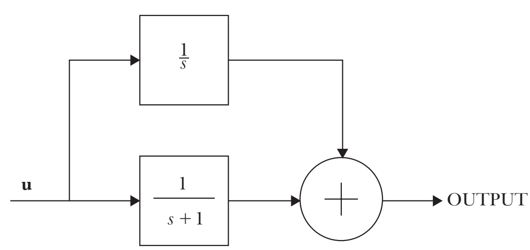



**FIGURE 7.44** Figure for Question 7.3.

**7.4** Write the simplified system equation describing the block diagram in Figure 7.45.

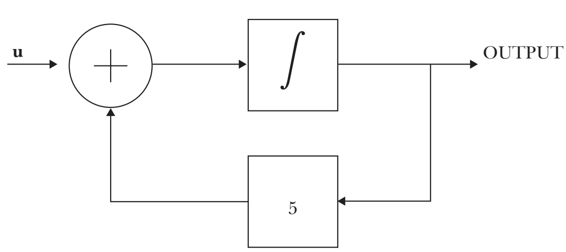



**FIGURE 7.45** Figure for Question 7.4.

**7.5** Sketch the pole-zero diagram corresponding to each of the following impulse response functions defined for $t > 0$.

**(a)** $h(t) = 10e^{-5t}$
**(b)** $h(t) = 10e^{-5t}\cos(10t)$
**(c)** $h(t) = 10e^{-5t}\sin(10t)$

**7.6** Sketch the pole-zero diagram corresponding to each of the following transfer functions $X(s)$. Indicate the region of convergence in each case.

**(a)** $X(s) = \frac{s^2}{s+4}$
**(b)** $X(s) = \frac{1}{s^2-16}$
**(c)** $X(s) = \frac{1}{s^2+16}$

**7.7** Sketch the pole-zero diagram corresponding to each of the following transfer functions $X(s)$. Indicate the region of convergence in each case.

**(a)** $X(s) = \frac{1}{s+4}$
**(b)** $X(s) = \frac{1}{s-4}$
**(c)** $X(s) = \frac{s}{s+4}$

**7.8** Find the Fourier transform $|H(j\omega)|$ for each of the following transfer functions $X(s)$. Which of these transfer functions do not lead to a valid Fourier transform? Sketch a Bode plot (frequency and phase) for each of the valid cases.

**(a)** $X(s) = \frac{1}{s+4}$
**(b)** $X(s) = \frac{1}{s-4}$
**(c)** $X(s) = \frac{s}{s+4}$

**7.9** Find the Fourier transform $H|(j\omega)|$ for each of the following transfer functions $X(s)$. Which of these transfer functions do not lead to a valid Fourier transform? Sketch a Bode plot (frequency and phase) for each of the valid cases.

**(a)** $X(s) = \frac{s^2}{s+4}$
**(b)** $X(s) = \frac{1}{s^2-16}$
**(c)** $X(s) = \frac{1}{s^2+16}$

**7.10** Write the inverse Laplace transform $x(t)$ corresponding to each of the following transfer functions $X(s)$. The table of Laplace transforms may be consulted to determine the correct form of $x(t)$.

**(a)** $X(s) = \frac{1}{s+4}$
**(b)** $X(s) = \frac{1}{s-4}$
**(c)** $X(s) = \frac{s}{s+4}$

**7.11** Write the inverse Laplace transform $x(t)$ corresponding to each of the following transfer functions $X(s)$. The table of Laplace transforms may be consulted to determine the correct form of $x(t)$.

**(a)** $X(s) = \frac{s^2}{s+4}$
**(b)** $X(s) = \frac{1}{s^2-16}$
**(c)** $X(s) = \frac{1}{s^2+16}$

**7.12** Write the differential equation corresponding to each of the following transfer functions $X(s)$.

**(a)** $X(s) = \frac{1}{s+4}$
**(b)** $X(s) = \frac{1}{s-4}$
**(c)** $X(s) = \frac{s}{s+4}$

**7.13** Write the differential equation corresponding to each of the following transfer functions $X(s)$.

**(a)** $X(s) = \frac{s}{s+4}$
**(b)** $X(s) = \frac{1}{s^2-16}$
**(c)** $X(s) = \frac{1}{s^2+16}$

**7.14** Consider the pole-zero diagram of Figure 7.46.

**(a)** Is this system stable? Explain.
**(b)** Write the corresponding Laplace transform.
**(c)** Describe the system response with the aid of a sketch.

**7.15** Consider the pole-zero diagram of Figure 7.46.

**(a)** Use MATLAB to define this system as the variable `sys`.
**(b)** Plot the impulse response.
**(c)** Plot the step response.
**(d)** Plot the Bode diagram for this system.

**7.16** Consider the pole-zero diagram of Figure 7.47.

**(a)** Is this system stable? Explain.
**(b)** Redraw the diagram and show the region of convergence.
**(c)** Write the corresponding Laplace transform.
**(d)** Describe the system response with the aid of a sketch.

**7.17** Consider the pole-zero diagram of Figure 7.47.

**(a)** Use MATLAB to define this system as the variable `sys`.
**(b)** Plot the impulse response.
**(c)** Plot the step response.

**7.18** Consider the pole-zero diagram of Figure 7.48.

**(a)** Is this system stable? Explain.
**(b)** Redraw the diagram and show the region of convergence.
**(c)** Write the corresponding Laplace transform.
**(d)** Describe the system response with the aid of a sketch.

**7.19** Consider the pole-zero diagram of Figure 7.48.

**(a)** Use MATLAB to define this system as the variable `sys`.
**(b)** Plot the impulse response.
**(c)** Plot the step response.

**7.20** Consider the pole-zero diagram of Figure 7.49.

**(a)** Is this system stable? Explain.
**(b)** Write the corresponding Laplace transform.

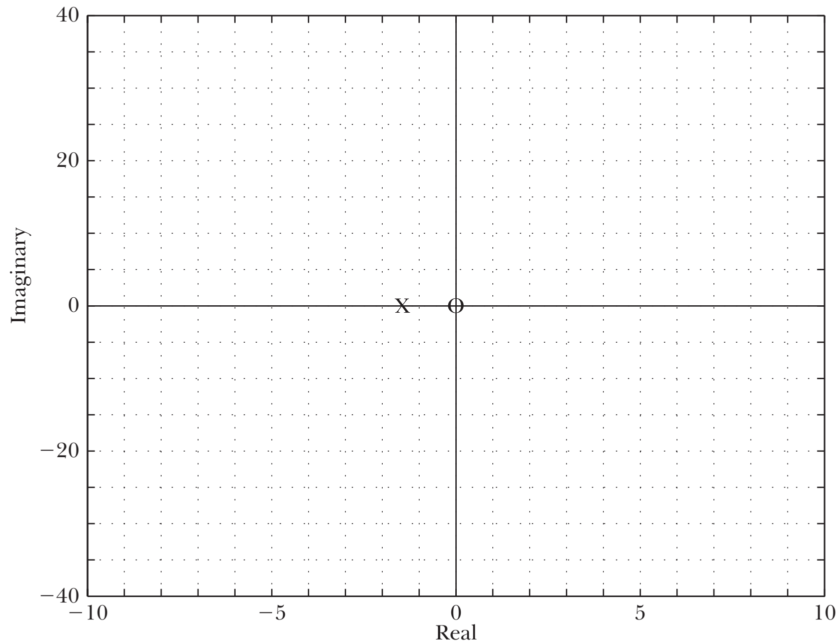



**FIGURE 7.46** Diagram for Questions 7.14 and 7.17.

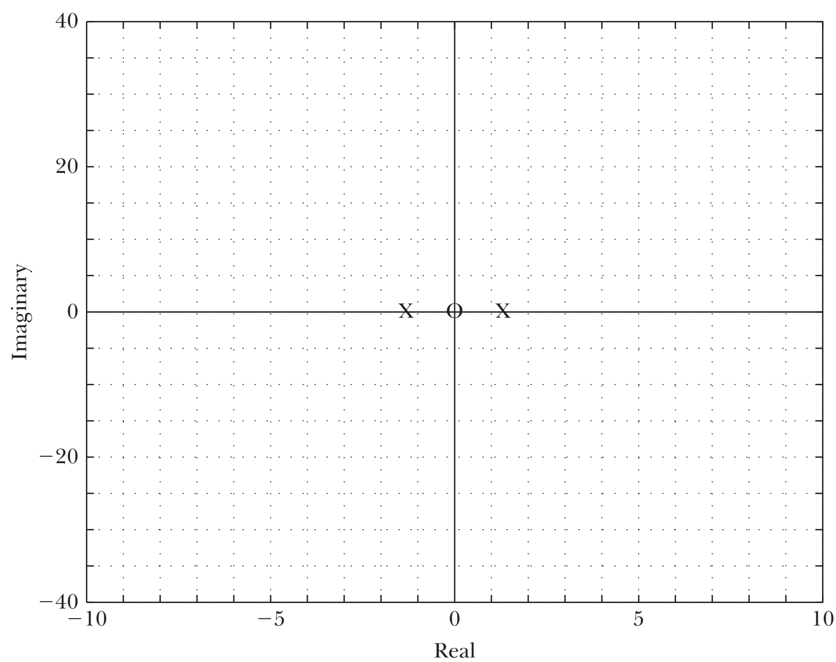



**FIGURE 7.47** Diagram for Questions 7.16 and 7.17.

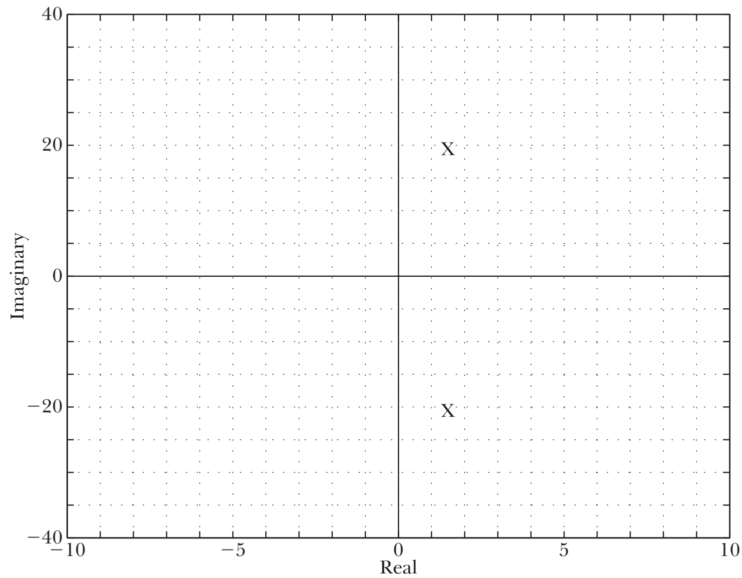



**FIGURE 7.48** Diagram for Questions 7.18 and 7.19.

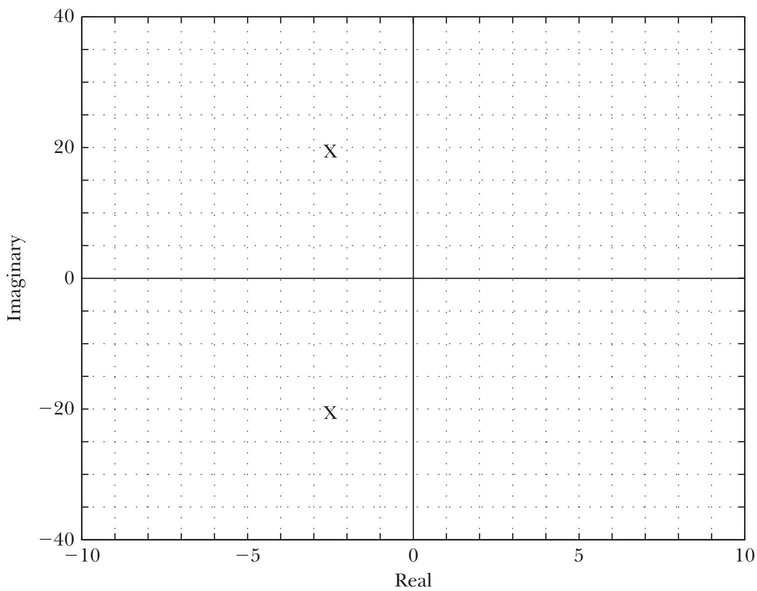



**FIGURE 7.49** Diagram for Questions 7.20 and 7.21.

**(c)** Describe the system response with the aid of a sketch.

**(d)** How would the system response vary if the poles are moved closer to the vertical axis?

**7.21** Consider the pole-zero diagram of Figure 7.49.

**(a)** Use MATLAB to define this system as the variable `sys`.

**(b)** Plot the impulse response.

**(c)** Plot the step response.

**(d)** Plot the Bode diagram for this system.

**7.22** Consider the Laplace transform $x(t) \xleftrightarrow{\mathcal{L}} 5/(s+3)^2$ and use the transform properties to determine the Laplace transform of the following variations on $x(t)$:

(a) $x(2t)$
(b) $x(2+t)$
(c) $2x(t)$

**7.23** Consider the Laplace transform $x(t) \xleftrightarrow{\mathcal{L}} 5/(s+3)^2$ and use the transform properties to determine the Laplace transform of the following variations on $x(t)$:

(a) $x(t/2)$
(b) $x(2+t/2)$
(c) $e^{2t}x(t)$

**7.24** Sketch the pole-zero diagram for the signal $X(s) = \frac{s+4}{s+4}$. Is it always possible to cancel the numerator and denominator in a case such as this?

**7.25** Consider the pole-zero diagram in Figure 7.50 as produced in MATLAB using the `pzmap()` command. From this diagram, determine the transfer function, then use MATLAB to plot the Bode diagram. Explain the response of this system in terms of amplitude and phase.

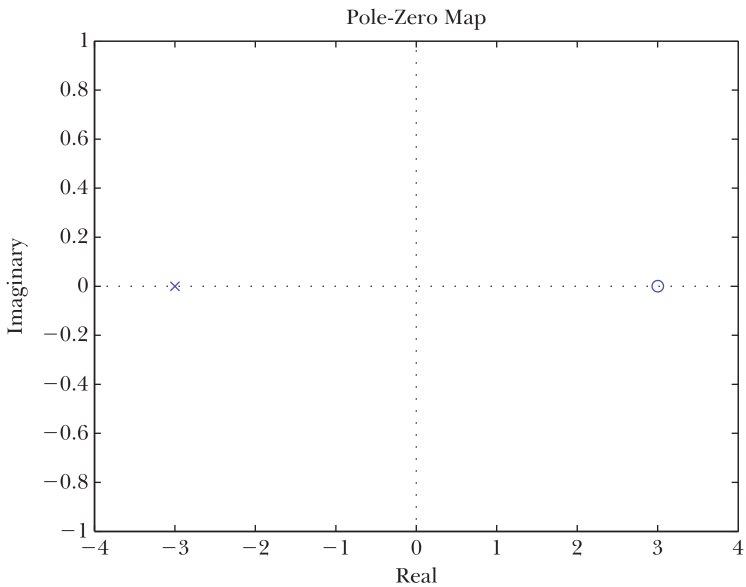



**FIGURE 7.50** Diagram for Question 7.25.

**7.26** Consider the circuit of Figure 7.51 with $R = 100\ \Omega$ and $C = 10\ \mu F$.

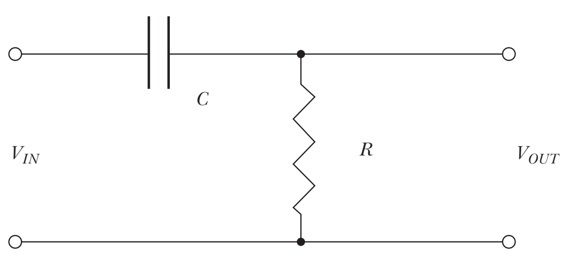



**FIGURE 7.51** Diagram for Questions 7.26 and 7.27.

(a) Find the Laplace transform relating the input and output voltages.
(b) Sketch the Bode plots describing the frequency and phase responses of this circuit.
(c) What is the overall behavior of this circuit?

**7.27** Consider the circuit of Figure 7.51 with $R = 20\ \text{k}\Omega$ and $C = 0.01\ \mu F$.

(a) Find the Laplace transform relating the input and output voltages.
(b) Sketch the Bode plots describing the frequency and phase responses of this circuit.
(c) What is the overall behavior of this circuit?

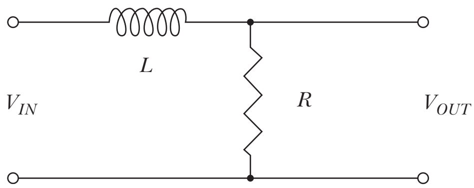



**FIGURE 7.52** Diagram for Questions 7.28 and 7.29.

**7.28** Consider the circuit of Figure 7.52 with $R = 100\ \Omega$ and $L = 1\ mH$.

**(a)** Find the Laplace transform relating the input and output voltages.
**(b)** Sketch the Bode plots describing the frequency and phase responses of this circuit.
**(c)** What is the overall behavior of this circuit?

**7.29** Consider the circuit of Figure 7.52 with $R = 20\ \text{k}\Omega$ and $L = 3\ H$.

**(a)** Find the Laplace transform relating the input and output voltages.
**(b)** Sketch the Bode plots describing the frequency and phase responses of this circuit.
**(c)** What is the overall behavior of this circuit?

**7.30** Consider the simple pendulum of Figure 7.53 in which a mass $m$ kg is at the end of a massless rod of length $L$ meters. The mass is acted upon by a constant downward force $mg$ due to gravity ($g = 9.8\ \text{m/s}^2$). The mass is constrained to move along an arc of radius $L$ where the pendulum position can be described as $\theta(t)$ radians. The equations of motion as shown include a factor ($k$) describing friction or damping as a function of angular velocity. *The system as shown is nonlinear but can be made linear for small* $\theta$ *by the approximation sin* $\theta = \theta$.

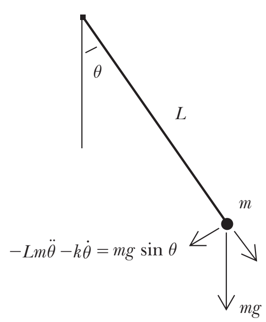



**FIGURE 7.53** Diagram for Questions 7.30 and 7.31.

The system may be considered initially hanging vertically $\theta(0) = 0$ and at rest $\dot{\theta}(0) = 0$ until some small input $u$ arrives to displace the pendulum. The equation of motion as a second-order differential equation is:

$$Lm\ddot{\theta} + k\dot{\theta} + mg\sin(\theta) = u$$

**(a)** Let $\sin\theta = \theta$ as above and find the Laplace transform of the system differential equation.
**(b)** From the Laplace transform write the transfer function in terms of $(L, k, g)$.
**(c)** From the transfer function, identify the poles and zeros for $k = 0$ and $L = 1$.
**(d)** Let $k = 0$ (no damping) and $L = 1$ and $m = 1$ use MATLAB to define a system named `sysA`.
**(e)** Plot the pole-zero diagram, Bode plot, and impulse response.
**(f)** Let $k = 1$ (damping) and $L = 0.5$ and $m = 1$ use MATLAB to define a system named `sysB`.
**(g)** Plot the pole-zero diagram and the impulse response.
**(h)** Plot the Bode diagram and identify the natural frequency of this system.

**7.31** Consider the simple pendulum of Figure 7.53 as in Question 7.30 and its analysis as a state space model.

**(a)** Let $x_1 = \theta(t)$ and $x_2 = \dot{\theta}(t)$ and set up a state space model for the simple pendulum. The system may be considered initially hanging vertically $\theta(0) = 0$ and at rest $\dot{\theta}(0) = 0$ until some small input $u$ arrives to displace the pendulum.
**(b)** Let $k = 0$ (no damping) and $L = 1$ and $m = 1$ and use MATLAB to define a state space system named `sysC`.
**(c)** Use the function `tf(sysC)` to see the transfer function for the system.
**(d)** Plot the pole-zero diagram and the impulse response.
**(e)** Plot the Bode diagram and identify the resonant frequency of this system.
**(f)** Does the resonant frequency depend on the mass $m$?
**(g)** Let $k = 1$ (damped) and $L = 0.5$ and $m = 1$, and use MATLAB to define a state space system named `sysD`.
**(h)** Plot the pole-zero diagram and the impulse response.
**(i)** Plot the Bode diagram and identify the natural frequency of this system.

CHAPTER 8

# Discrete Signals

## 8.1 Introduction

A signal described by a continuous function of time has a unique value at every definable instant. For example, given the continuous cosine signal $c(t) = \cos(2\pi f_0 t)$, its value at an arbitrary time such as $t = 5.00012$ s can be determined by computing $c(5.00012)$ with the aid of a pocket calculator. This chapter deals with discrete time signals of the form $\{c(nT)\}$ for constant $T$ and integer $n$, which do not exist at every point in time, but which are generally representative of an underlying continuous signal $c(t)$.

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Describe the difference between discrete time and continuous time signals
- Compute a suitable sampling rate for a bandlimited signal
- Explain how a signal can be recovered from its samples
- Explain the concept of aliasing and the use of an anti-aliasing filter
- Describe the discrete time Fourier transform (DTFT) and discrete Fourier transform (DFT)
- Use MATLAB to find the Fourier transform of sampled signals
- Apply the DFT to frequency-domain filtering
- Explain time-domain filtering in terms of delays

## 8.2 Discrete Time vs. Continuous Time Signals

Consider the task of plotting by hand some continuous signal $s(t)$. It is usually not necessary to compute $s(t)$ for every point on a graph. Instead, the desired shape of $s(t)$ can be reproduced by drawing a smooth line through a set of sample points $\{s[n]\}$ computed at regular intervals of $T$ s along the time axis. To this end, a table might be constructed showing the values of $s[n]$ as shown below.

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $\cdots$ |
|---|---|---|---|---|---|---|---|---|---|
| $s[n]$ | 1.00 | 0.95 | 0.81 | 0.59 | 0.31 | 0.00 | $-0.31$ | $-0.59$ | $\cdots$ |

In the same way, a computer program might be written or MATLAB may be used to create the graph, first by filling a data array with values $s[n]$, then by fitting a smooth curve to the points. In either case, the final (continuous) graph of $s(t)$ is based upon a relatively small number of stored points, as shown in Figure 8.1.

In constructing a graph, it is not necessary to compute every point in the function for every small fraction of its existence, yet the resulting curve is expected to describe a function for every point in time. This procedure effectively describes the important process of signal sampling and reconstruction, since *only the points in the*

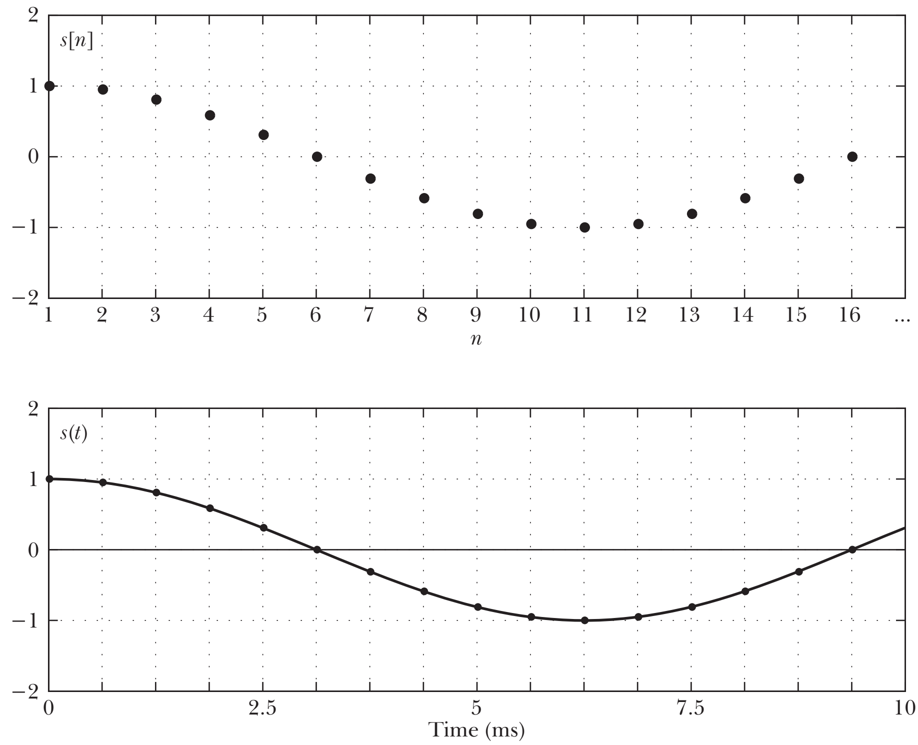



**FIGURE 8.1** **Sampling and Reconstruction** Graphs produced by hand or by computer generally begin with a set of points from a table after which a smooth curve is drawn through the points to complete a continuous shape. The intent of drawing a graph in this way is that the final curve should represent an underlying continuous function $s(t)$.

*table are required to reproduce the original signal.* Because the discrete and continuous versions of a signal are related, the manipulation of discrete time signals naturally reflects the properties and behavior of continuous time signals presented in previous chapters.

## 8.2.1 Digital Signal Processing

The study of discrete signals is especially important because it is necessary to use an array of sampled points whenever a digital computer is used to facilitate signals analysis. The processing of discrete signals by computer, or *digital signal processing* (DSP), is increasingly employed to replace traditional analog approaches in modern systems. For example, long-distance telephone calls are typically transmitted over digital links as discrete time values of amplitude measured at 8000 samples per second. Similarly, an audio compact disk (CD) stores music digitally at 44,100 samples per second, and a CD player reconstructs the original signal waveforms from the sample values stored on the disk.

Generally, digital signal processing refers to the *real-time* processing of signals as in the above examples. In this context, the term *real-time* implies that a very fast specialized computer is used to performed operations on samples of a continuous signal as quickly as those samples are acquired. Digital signal processing provides an attractive alternative to the use of analog system components. It has already been observed that the overall behavior of a system is fully described by its impulse response function. Thus, just as a first-order lowpass filter might be built using a resistor and a capacitor (RC) or a resistor and an inductor (RL), it is also possible to perform this operation numerically using digital signal processing. After sampling an input signal, the discrete points can be manipulated mathematically to perform first-order lowpass filtering, and the filtered signal can be reconstructed from the resulting discrete signal. If a system has the desired response function, it should make no difference whether the circuit inside is built from RC, RL, or DSP.

Given that a DSP-based circuit can reproduce the behavior of a hardware-only design, the advantages of DSP can outweigh the limitations of traditional approaches. Computer-based designs avoid the problems of dealing with the tolerance of real components or values, which may change with temperature or humidity or time. When component behavior is specified mathematically, it is possible to manufacture multiple units with identical performance specifications. New or modified design objectives are easily met by changing parameters without the need to order new parts. Indeed, the same circuit can perform different functions at different times, which is the case for *adaptive filters* that can vary their characteristics over time or in response to changing conditions.

## 8.3 A Discrete Time Signal

A discrete time signal is a sequence of values or *samples* taken at specific times, usually from a continuous signal $s(t)$. Generally, the time increment $T$ between each sample is constant, and this increment defines a fixed *sampling rate*, $f_s = 1/T$ Hz. The resulting discrete time signal is of the form $s[n]$ where $n$ is an integer.

**DEFINITION 8.1 Discrete Time Signal**

*If a signal* $\mathrm{s[n]}$ *has values* $\mathrm{s[n] = s(nT)}$ *defined only at discrete points of time* $\mathrm{t = nT}$ *for constant* $\mathrm{T}$ *and integer* $\mathrm{n}$*, then* $\mathrm{s[n]}$ *is called a* discrete time signal.

The discrete time signal can be written as $s[n]$ using square brackets to distinguish it from the continuous time signal $s(t)$. The signal $s[n]$ may be regarded as an array of values $s[n] = s(nT)$. Since $s[n]$ is a function of $n$ alone, the important factor $T$ that describes the time between each value must be separately recorded if the signal $s[n]$ is to faithfully represent the $s(t)$ from which it was acquired. For clarity, discrete time signals will be labelled either $s(nT)$ or $s[n]$ as may be appropriate.

### 8.3.1 A Periodic Discrete Time Signal

**DEFINITION 8.2 Periodic Discrete Time Signal**

*Given a signal* $\mathrm{s[n]}$ *if, for all* $\mathrm{n}$ *and for constant* $\mathrm{N}$,

$$s[n + N] = s[n]$$

*then* $\mathrm{s[n]}$ *is a* periodic *signal, and* $\mathrm{N}$ *is its* period.

An important issue arises when sampling periodic continuous time signals to create periodic discrete time signals. Consider a continuous time periodic signal $s(t)$ and its discrete time version $s[n]$. By definition, $s(t)$ is periodic with period $T_0$ if $s(t) = s(t + T_0)$. Similarly, the corresponding $s[n]$ is periodic with period $N$ if $s[n] = s[n + N]$. In practice, this demands that the signal period $T_0$ be an exact multiple of samples $n$, or $T_0 = kn$ for some integer $k$. While this condition can be met in theory, it is rarely true in practical signals.

## 8.4 Data Collection and Sampling Rate

Before going on to study discrete time signals, it is important to examine closely the question of signal sampling and reconstruction. The most fundamental issue concerns the minimum sampling rate required to faithfully reproduce a signal from discrete samples.

### 8.4.1 The Selection of a Sampling Rate

Suppose that the local airport weather office requires an instrumentation package to monitor daily temperature readings. An embedded computer-based system is to be designed that will gather temperature readings and record them for later analysis. How often should readings be taken to faithfully follow temperature variations? Give this question some thought before reading on.

A colleague suggests 30 times a second yet, somehow, that rate intuitively seems far too high; it does not make good sense to record the temperature 1800 times a minute or over 100,000 times every hour. The engineering *gut feeling* about this suggestion is based on the sure knowledge that weather does not change that quickly. This acquisition rate would quickly fill a computer file with temperature data in which most readings were identical to the previous ones. The fact is that variations in outside temperature take place over periods of minutes or hours, rather than fractions of seconds. By the same reasoning, once-a-week or once-a-month readings would not produce very useful temperature records, although some statistics may eventually be compiled from the results.

The most fundamental question in sampling any signal concerns the choice of sampling rate. An appropriate sampling rate can be determined only if something is known about the signal to be measured. In the case of outdoor temperature, this choice is based upon the expected behavior of the weather. If this underlying assumption is violated (for example, if the outside temperature suddenly and mysteriously begins to oscillate up and down ten degrees every half second), the sampling operation will fail.

Thirty outdoor temperature readings per second are far too many. Once-a-minute readings are probably still too many. Yet, with once-a-minute readings, it is easy to reconstruct with a high degree of confidence a table of temperature readings every second within that minute and every $1/30$ s therein. In other words, the exact temperature at every instant of the day (every microsecond, if required) can be interpolated from readings taken at suitable intervals. No one would argue that anything was lost by not choosing a $1/30$ s spacing, because that data can always be reconstructed if necessary, using only the minute-by-minute recordings and the confidence that *weather doesn't change that quickly*. The latter statement really implies that outside temperature variations are dominated by low-frequency components.

This observation begins to quantify the gut feeling that has directed this discussion. In summary, where there are no high-frequency components, a smaller sampling rate can safely be chosen.

## 8.4.2 Bandlimited Signal

The frequency components in a signal reflect how rapidly the signal values can change. A signal that varies slowly over time would have only low-frequency components. Conversely, a signal constrained to have only low-frequency components would never change more rapidly than the maximum frequency component present. Such a signal is said to be *bandlimited*, as shown in Figure 8.2. In the frequency domain, a bandlimited signal $s(t)$ is defined as having no components beyond a maximum frequency limit.

**DEFINITION 8.3 Bandlimited Signal**

*If*

$$s(t) \xleftrightarrow{\mathcal{F}} S(f)$$

*and*

$$S(f) = 0 \text{ for all } |f| > f_{max}$$

*then* s(t) *is called a* bandlimited signal.

Practical signals are ultimately bandlimited by the fact that cables, amplifiers, microphones, human ears, or any system through which signals must travel, cannot pass all frequencies up to infinity. Any signal $s(t)$ can be forced to be bandlimited by passing the signal through a suitable lowpass filter. In every case, some of the power in the signal may be lost, and the shape or appearance of the signal will be altered by any loss of frequency components. In many cases, the resulting loss will be negligible. For example, no practical difference would be found when listening to an audio signal that is bandlimited to remove components above 20 kHz, since human ears cannot detect the missing frequencies anyway. In fact, AM broadcast stations limit audio frequencies to under 5 kHz, and telephone circuits limit the passage of voice signals to under 4 kHz.

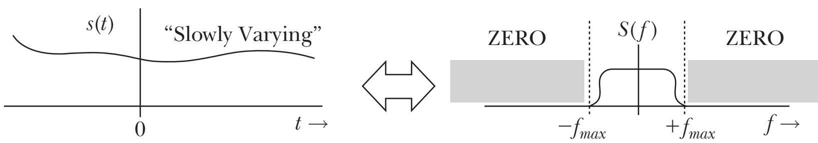



**FIGURE 8.2 Bandlimited Signal** A signal $s(t)$ is *bandlimited* if its spectrum $S(f)$ is zero for all $|f| > f_{max}$.

## 8.4.3 Theory of Sampling

Given some knowledge of the system to be measured (in particular, knowing the most rapid change expected, or the highest frequency component present), a suitable sampling rate can be determined from which the exact behavior of the system for every instant of time can be interpolated. In other words, *a bandlimited signal can be reproduced exactly from samples taken at appropriate intervals.* The corresponding *sampling rate* or *sampling frequency* is typically expressed in units of hertz.

A close examination of the process of sampling and signal reconstruction will reveal what is an appropriate sample rate for a given signal $s(t)$. Throughout this chapter it will be assumed that samples are taken at regular intervals. The analysis of variable sampling rates is beyond the scope of this discussion.

## 8.4.4 The Sampling Function

Sampling a signal $s(t)$ consists of measuring its amplitude at regular intervals determined by the sampling rate. Mathematically, this can be accomplished by multiplying $s(t)$ by a *sampling function,* which serves to extract a series of amplitude values. Consider samples taken every $T$ s (sampling rate = $1/T$ Hz). The simplest sampling function is the familiar impulse train, written as $\{\delta(t-nT)|n \in I\}$ and other more practical sampling functions will follow the same general behavior.

The product $\delta(t-nT) \times s(t)$ is zero everywhere except at every point $nT$; for a given value $n$, the area of the impulse $\delta(t-nT)s(nT)$ located at $nT$ is the discrete value $s[n]$. Recall that an impulse function is usually sketched with a height that reflects its area (all impulses being of infinite amplitude). A sketch of the sampled signal therefore resembles an impulse train with a envelope reflecting the instantaneous value of $s(t)$ at the position of each impulse. These values form the signal samples and could be recorded or transmitted digitally for later reproduction.

The goal of sampling theory is to reproduce the original, continuous signal $s(t)$, starting with only these samples. This will be possible only if $s(t)$ is bandlimited and a suitable sampling rate has been chosen.

In the frequency domain, the signal $s(t)$ has a unique Fourier transform $S(f)$. In other words, knowing $S(f)$ is as good as having $s(t)$ in its original continuous form. Consequently, if $S(f)$ can be identified in, and isolated from, the Fourier transform of the sampled signal, then the goal of reproducing $s(t)$ from those samples will have been accomplished.

Figure 8.3 illustrates the sampling process. Recall that the Fourier series of an impulse train with impulses spaced every $T$ s as $\delta(t-nT)$ resembles another impulse train in frequency as $\delta(f-n/T)$ with components spaced at $1/T$ Hz. Sampling $s(t)$ in the time domain by multiplication with the impulse train corresponds to the frequency-domain convolution of the transformed sample train with $S(f)$.

The sampling process creates, in the time domain, discrete signal values obtained at regular time intervals ($T$). In the frequency domain, the spectrum consists of multiple copies of $S(f)$, which appear spaced at frequency intervals of $1/T$ Hz, as shown in Figure 8.3.

It is important to remember that when processing sampled signals, only these sampled waveforms are available and the original signal $s(t)$ is unknown. For example, these samples may be read in a CD player, from which it is necessary to re-create the continuous musical waveform. In practice, recovering the original $s(t)$ from the sample values $s[n]$ is a relatively simply operation.

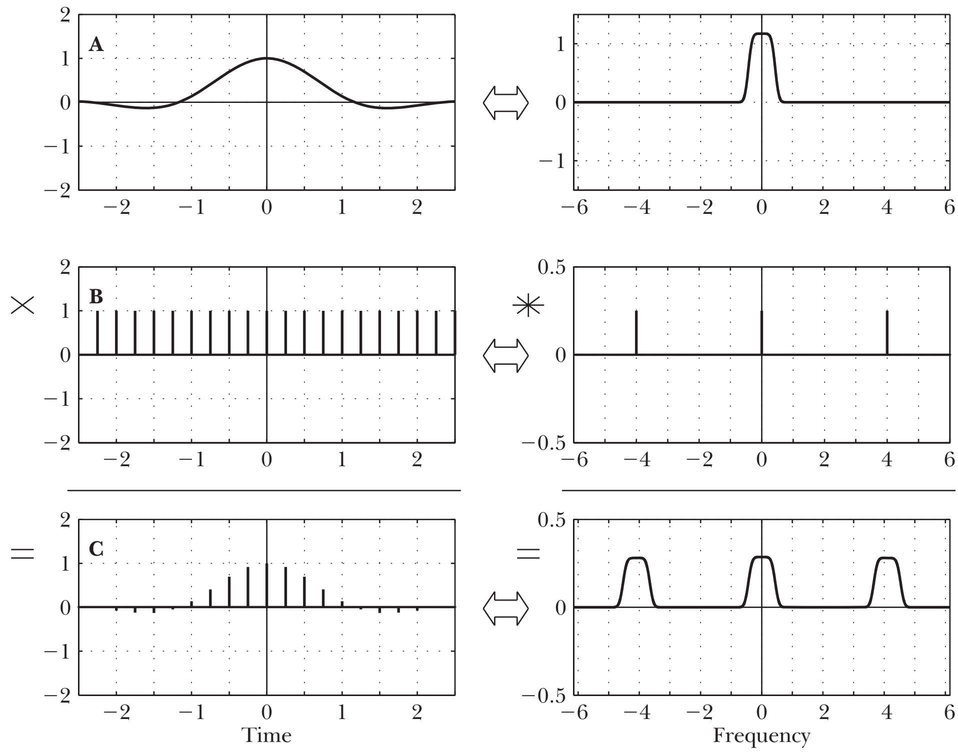



**FIGURE 8.3 Impulse Train Sampling Function** A bandlimited signal **A** is sampled by multiplication with the impulse train in **B** to give the regularly spaced samples **C**. In the frequency domain, the spectrum close to the origin is identical for both the original signal and the samples.

## 8.4.5 Recovering a Waveform from Samples

Consider the signal in the frequency domain resulting from sampling every $T$ seconds the continuous time signal $s(t)$, as shown in Figure 8.3C. Multiple copies of $S(f)$ lie spaced at intervals of $1/T$ along the frequency axis. Observe that one copy of $S(f)$ lies alone at the origin. An ideal lowpass filter can isolate this $S(f)$ from the other copies, and the goal of reproducing the exact original signal from samples will have been accomplished. In other words, a lowpass filter is all that is required to recover the original signal from the samples.

## 8.4.6 A Practical Sampling Signal

While sampling with an impulse train nicely illustrates the principles of sampling theory, it is clear that an ideal impulse train cannot be generated in the lab, and a

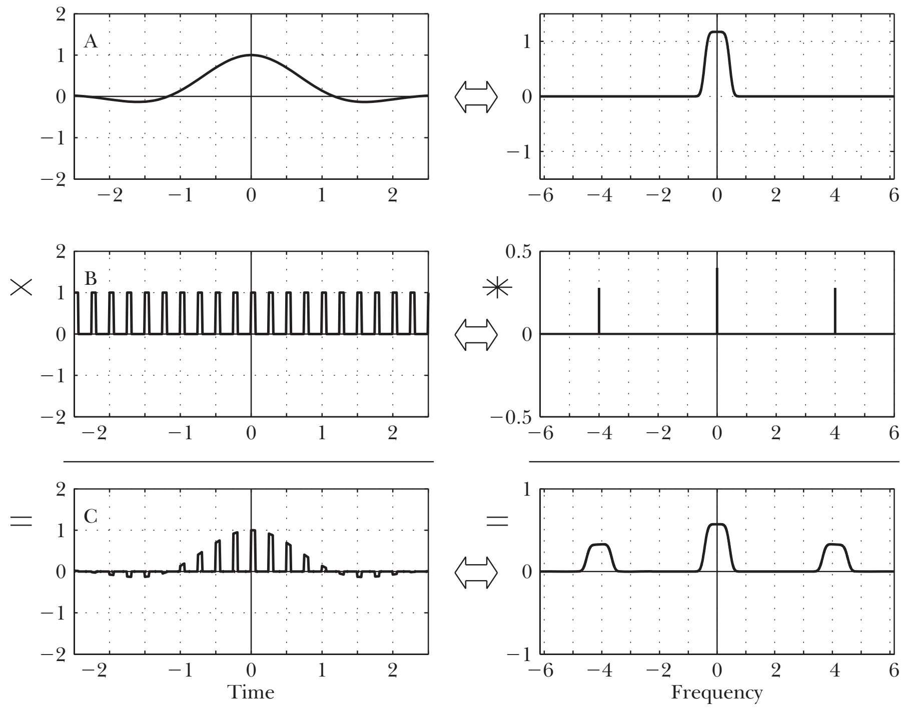



**FIGURE 8.4 Pulse Train Sampling Function** A bandlimited signal **A** is sampled by multiplication with the pulse train in **B** to give the regularly spaced samples **C**. In the frequency domain, the spectrum close to the origin is identical for both the original signal and the samples.

real sampling function inevitably has some width. One practical sampling circuit consists of a switch that is opened and closed for short periods of time. This sampling function can then be modelled as a pulse train, as shown in Figure 8.4.

When this pulse sampling function is multiplied by the continuous signal $s(t)$, the sampled signal looks like Figure 8.4C. It is noteworthy that this change in sampling function compared to Figure 8.3 has little effect on the reconstruction process. A lowpass filter is again used to reconstruct the original signal. This is because the Fourier transform of the pulse train with period $T$ s is an infinite series of impulses spaced at $1/T$ Hz. Since the only component of interest lies at the origin $(f = 0)$, the non-zero frequency components will not affect the outcome. If anything, the higher-order frequency components will generally become smaller at higher frequencies, and therefore more easily filtered out. This is logical since the wider sampling pulses lead to narrower gaps chopped from the original continuous signal, and consequently reconstruction ought to be easier.

## 8.4.7 Minimum Sampling Rate

In the examples above, the frequency components were well separated, and isolating one copy of $S(f)$ at the origin was straightforward. The separation between copies is given by the sampling rate $1/T$ Hz, such that copies of $S(f)$ will be wider apart in frequency for higher sampling rates and closer together for lower sampling rates. As seen in Figure 8.5, the desired copy of $S(f)$ near the origin will not be isolated or alone if the adjacent copies of $S(f)$ interfere (this starts to happen if the sampling rate is too low). The importance of a suitably high sampling rate can now be seen. As the sampling rate goes higher, adjacent copies of $S(f)$ move conveniently away from the origin and it becomes easier to isolate the components near $f = 0$. This is consistent with the example of airport temperature readings; if many extra readings are taken (higher sampling rate), there is no difficulty interpolating between those readings. Also note the importance of $s(t)$ being a bandlimited signal; without this condition, there could always be some frequency-domain overlap to prevent the reliable isolation of $S(f)$.

It is important to determine the lowest possible sampling rate, and how is it related to the bandwidth of $s(t)$. By inspection of the frequency-domain sketch, for a bandlimited signal $s(t)$ having a maximum frequency $f_{max}$, the next copy of $S(f)$ will be guaranteed not to overlap only if the chosen sampling rate is greater than $2f_{max}$.

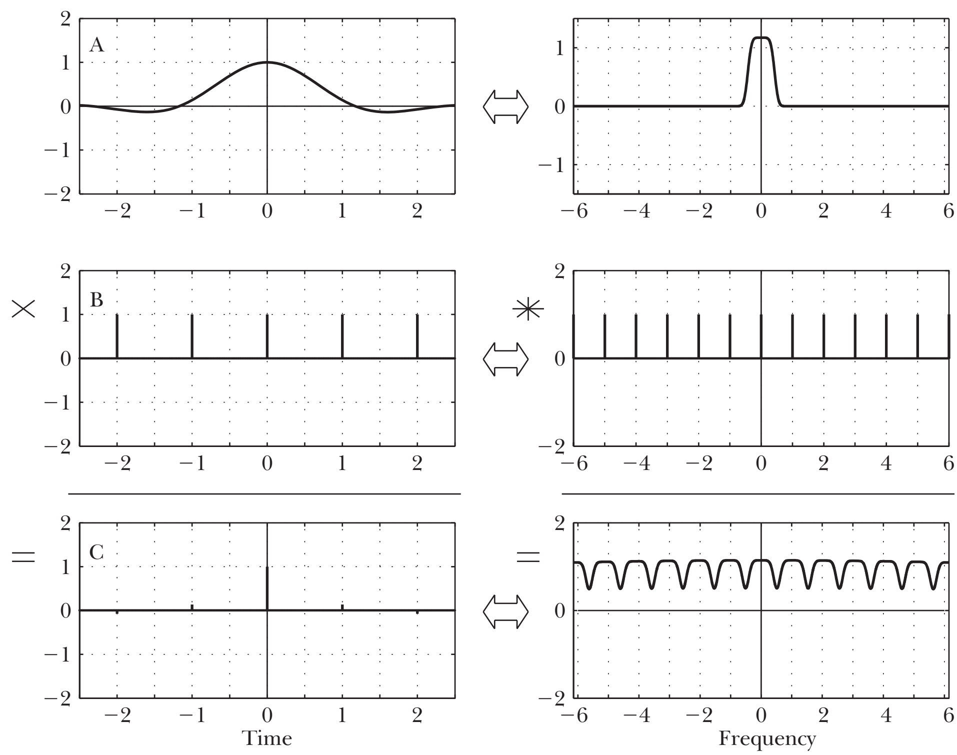



**FIGURE 8.5 Undersampling** A bandlimited signal **A** is sampled by multiplication with the pulse train in **B** to give the regularly spaced samples **C**. Because the samples **B** are too far apart, the frequency-domain copies of **A** overlap in **C** making it impossible to recover the original signal.

## 8.4.8 Nyquist Sampling Rate

If $s(t)$ is a bandlimited signal, then $s(t)$ can be exactly reproduced from samples taken at appropriate intervals. For this to be possible, the lowest acceptable sampling rate $f_s$ must be greater than $2f_{max}$, where $f_{max}$ is the highest frequency components present in the signal $s(t)$. This result is known as the *Nyquist sampling rate*.[1]

**THEOREM 8.1** *(Sampling Theorem)*

*If* $\mathrm{s(t)}$ *is a bandlimited signal with maximum frequency* $\mathrm{f_{max}}$ *Hz, then* $\mathrm{s(t)}$ *can be exactly reproduced from samples* $\mathrm{s(nT)}$ *for constant* $\mathrm{T}$ *and integer* $\mathrm{n}$ *provided that the sampling rate* $\mathrm{f_s = 1/T}$ *is chosen such that*

$$f_s > 2f_{max}$$

*The frequency* $\mathrm{2f_{max}}$ *is called the* Nyquist sampling rate *or* Nyquist rate.

**EXAMPLE 8.1 (Sampling Rate)**

*Determine the minimum sampling rate that can be used for the signal*

$$s(t) = \cos(10\pi t) + \sin(4\pi t)$$

**Solution:**
The signal $s(t)$ has a cosine component at 5 Hz and a sine component at 2 Hz. Therefore, the signal $s(t)$ is bandlimited to $f_{max} = 5$ Hz. From the sampling theorem, the minimum sampling rate is $f_s > 2f_{max}$, so a sampling rate $f_s > 10$ Hz must be chosen.

## 8.4.9 The Nyquist Sampling Rate Is a Theoretical Minimum

While the minimum sample rate guarantees that the original signal can be reconstructed, it does not really provide many points with which to work. Consider Figure 8.6, showing only the samples from a (1 Hz) bandlimited waveform sampled at 2.3 Hz. At a glance, it is not immediately obvious what the original waveform is, yet sampling theory states that it should be possible to reconstruct the entire signal, knowing only that the highest frequency present is less than one-half of this sampling rate. This is the real challenge of reconstructing a signal from samples. Take a moment to guess what the original signal looked like.

Given only the samples, the sampling rate may be determined by inspection (by measuring the spacing between sample points), and it may be assumed that sampling was done correctly and therefore the highest frequency component is less than one-half the sampling frequency. This observation eliminates many candidate

[1] Also called *Nyquist-Shannon sampling theorem.*

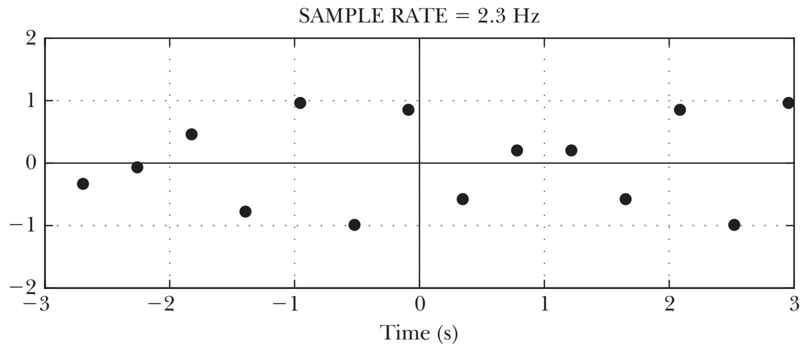



**FIGURE 8.6 Samples taken at 2.3 Hz rate** Reconstructing a sampled signal begins with only the samples. It is not immediately obvious what continuous signal was sampled here.

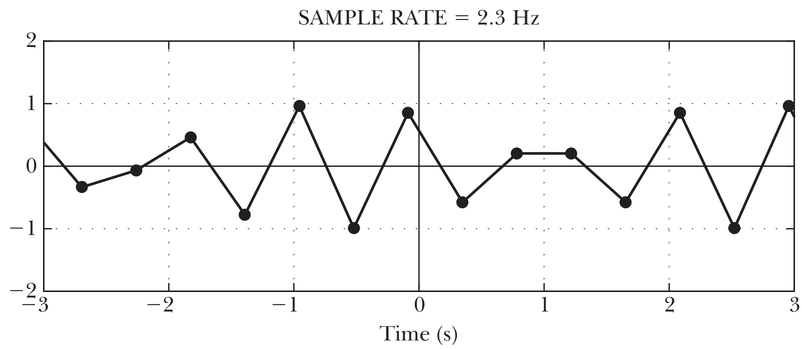



**FIGURE 8.7 Reconstructing a signal from samples** Connecting the sample points from Figure 8.6 as shown here could not be correct as any such sharp edges in the original signal would require much higher frequency components than the sample rate suggests.

waveforms that might match the sample points. For example, the reconstructed waveform shown in Figure 8.7 could not be correct, since the presence of sharp corners (rapid changes) suggests many high-frequency components, which would violate the sampling rule.

When the correct sampled waveform is superimposed in Figure 8.8, it is obvious that these were samples of a 1 Hz cosine. The sample rate was slightly higher than the Nyquist rate, and the match is definitive. No lower-frequency cosine would fit these points, and any higher-frequency cosines would not be eligible candidates under the Nyquist rule. This is the continuous waveform that would come about if the samples were passed through an ideal lowpass filter with a cutoff frequency of 2.3/2 Hz.

Finally, it can be noted that if the sampling rate is a few times higher than the minimum suggest by Nyquist, reconstruction becomes much easier. It is not difficult to guess the answer in the same example if the sampling rate is four times higher, as shown in Figure 8.9. Here, the 1 Hz cosine is sampled at 9.2 Hz.

Taking more samples than is absolutely necessary is known as *oversampling*. It is not always possible or practical to use a higher sampling rate, and the Nyquist rate remains an important lower limit to the sampling operation. It was noted that an

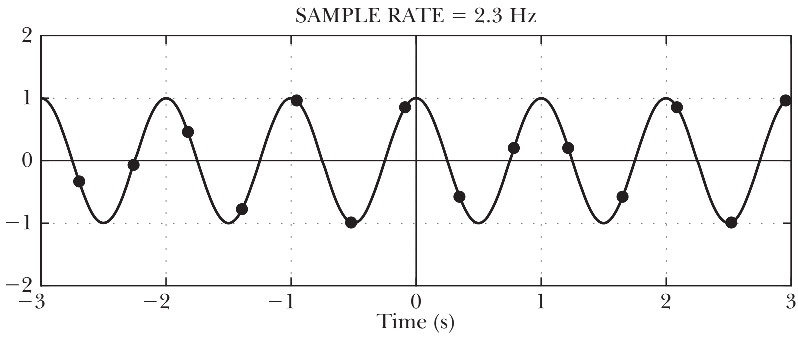



**FIGURE 8.8 The correct answer** The samples in Figure 8.6 are of a 1 Hz cosine sampled at 2.3 Hz.

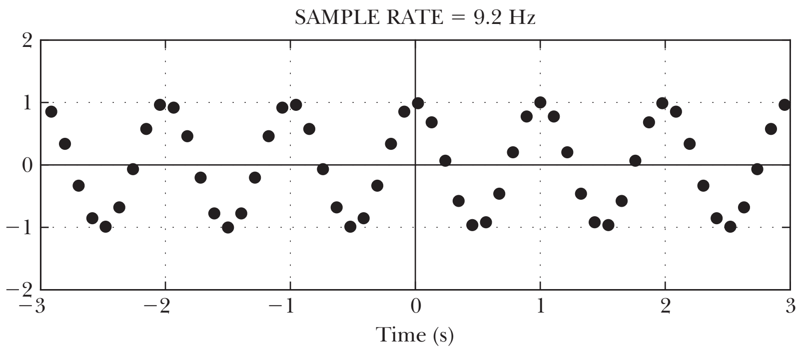



**FIGURE 8.9 Oversampling** Samples of a 1 Hz cosine, taken at 9.2 Hz, much higher than the minimum sampling rate.

audio CD holds music sampled at 44,100 Hz. This rate allows reproduction of musical content exceeding 20,000 Hz, which is at the limit of human hearing; the sampling rate is chosen to maximize the capacity of the storage medium.

## 8.4.10 Sampling Rate and Alias Frequency

If the sampling rate is increased, it becomes easier to reconstruct the original signal. As samples are taken closer together in time, copies of the Fourier transform of $S(f)$ spaced equally along the frequency axis will spread apart, making it easier to isolate a single copy at the origin, and thereby recover the original signal in its entirety. This is desirable in practice, since a perfect lowpass filter can never be constructed to exactly extract the original signal. In fact, sampling rates at several times this minimum value are commonly chosen in real DSP applications. Note that a near-perfect filter would be necessary if a signal was to be recovered from samples taken at a rate just above the Nyquist limit.

The Nyquist Sampling Theorem is frequently (mis)quoted as something like, "You must sample at twice the frequency of the signal...." There are two problems with such a statement. First, unless dealing with pure sinusoids, the highest frequency component of a signal must be considered, not just the fundamental frequency. Second, the minimum sampling rate in the theorem is strictly greater than the highest-frequency component. For example, a 1 Hz sine wave $s(t) = \sin(2\pi t)$ sampled at exactly 2 Hz, can yield samples $s(n\pi) = 0$ for all integer $n$.

Conversely, if the sampling rate is decreased, the copies begin to overlap, and recovery of the original signal becomes impossible. If the sampling rate is many times too small, then multiple overlapped copies of $S(f)$ in the frequency domain will result in complete failure of the sampling operation. A troublesome result of sampling at less than the required frequency or *undersampling*, can be observed with a simple example.

Consider the signal $a(t) = \cos(2\pi t)$. Since this cosine has a frequency of 1 Hz, a sampling rate greater than 2 Hz must be used for reliable reproduction. To see what would happen, suppose that a too low sample rate of 1.7 Hz is deliberately chosen. Next, consider the signal $b(t) = \cos(1.4\pi t)$, which has a frequency of 0.7 Hz and for which the sample rate of 1.7 Hz is perfectly adequate. The resulting samples from both $a(t)$ and $b(t)$ may be computed with a calculator and are shown in Table 8.1 and in Figure 8.10B.

This sampling operation seems correct; however, it can be observed that the samples for $b(t)$ are identical to those for $a(t)$. Figure 8.10 shows both $a(t)$ and $b(t)$ and the samples. It is clear that *these same samples could have come from either cosine signal.* Moreover, there is no way to uniquely identify the original signal from these samples because a too-low sampling rate was chosen for $a(t)$. Note that many higher-frequency cosines could also match these sample points, but no lower-frequency cosine passes through all the sample points, and the lowest frequency signal $b(t)$ would always be interpreted as being the *true* original signal. The problem is especially significant, because using a too-low sample rate did not produce an obvious error, yet it did produce a totally incorrect result.

All observations suggest that the original signal was $b(t)$ and not $a(t)$. Recall that only the samples can be consulted at this stage; there is no means of going back to check the original appearance of $a(t)$. With no other information, the signal $b(t)$ effectively masquerades as the actual signal that was sampled. This effect is called *aliasing*, and the signal $b(t)$ is called the *alias frequency*.

**TABLE 8.1**

**Undersampling** A bandlimited signal given by $a(t) = \cos(2\pi t)$ is sampled at a (too low) sample rate of 1.7 Hz ($T = 1/f_s = 0.5882$ s) The samples also match the signal $b(t) = \cos(1.4\pi t)$. Signal $b(t)$ is called the *alias frequency*. See Figure 8.10.

| n | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| nT | 0 | 0.5882 | 1.1765 | 1.7647 | 2.3529 | 2.9412 | 3.5294 |
| a(nT) | 1.0000 | −0.8502 | 0.4457 | 0.0923 | −0.6026 | 0.9325 | −0.9830 |
| b(nT) | 1.0000 | −0.8502 | 0.4457 | 0.0923 | −0.6026 | 0.9325 | −0.9830 |

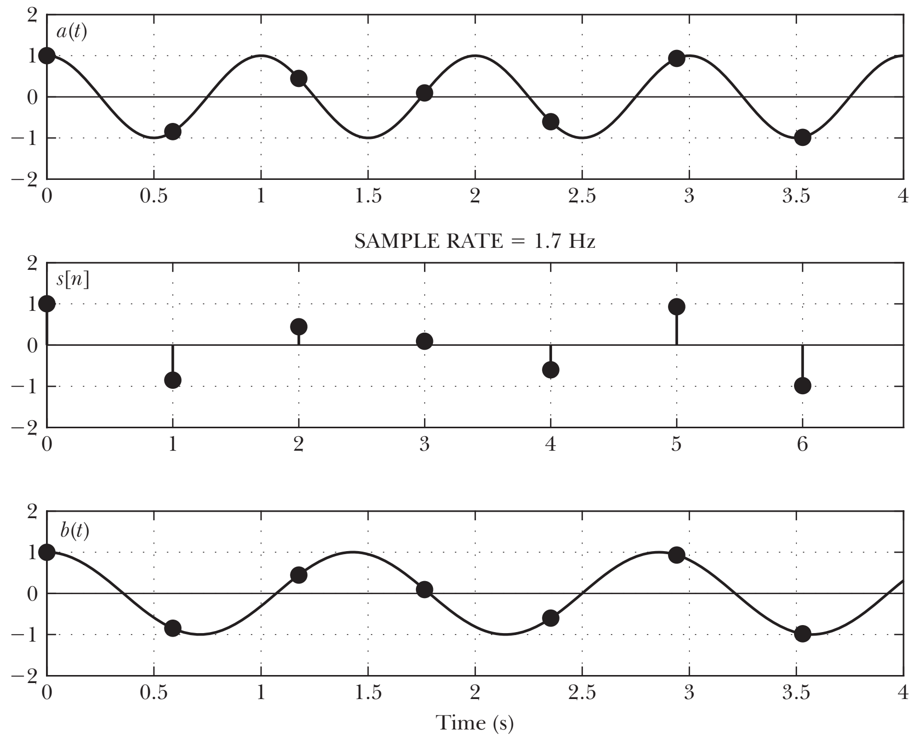



**FIGURE 8.10** **Undersampling** A bandlimited signal $a(t) = \cos(2\pi t)$ is sampled at a (too low) sample rate of 1.7 Hz. The samples $s[n]$ also match the signal $b(t) = \cos(1.4\pi t)$. Signal $b(t)$ is called the *alias* signal. See Table 8.1.

## 8.4.11 Practical Aliasing

Alias components can be identified in a number of practical situations as found here.

1. The classic example of aliasing concerns the visual effect observed in old western movies when a stage coach or covered wagon begins to roll forward. As the wheels begin to turn, the spokes at first appear to describe a backwards motion! Without seeing the entire scene, it could only be concluded that the wagon was, in fact, moving backwards. This problem occurs because motion pictures represent a sampling system, where still photographs taken at 24 frames per second are flashed back in rapid sequence, creating the illusion of continuous motion. The original, continuous motion scene is accurately observed by moviegoers only if objects being filmed do not move faster than the Nyquist sampling rate demands. As the wagon begins to move, the spokes move in front of the camera faster than the sampling rate permits for accurate reproduction. The result is a vivid illustration of aliasing, and a stark reminder of what happens when this sampling rule is ignored.

   In this example, the sampling rate (24 frames per second) works well for most movie action. (The minimum frame rate is based on human physiology and the desire to minimize flicker. Human eyes form a lowpass filter that effectively reproduces the original continuous movement from the sampled photos.) The sampling process fails when an unexpected input signal arrives (the fast-moving wheels). This problem could be avoided by increasing the sampling rate (but shooting at 100 frames per second would waste a lot of film[2]) or by eliminating fast-moving objects (do not shoot any scenes with wagon wheels). Moviemakers choose instead to let the audience decide what is real, and what is aliasing.

The discussion using a single sinusoid can be extended to all signals, since the Fourier transform is a decomposition into orthogonal sinusoidal components. When a sample rate is chosen, it is important to consider the highest-frequency *component* present in the input signal.

2. Consider a series of photographs of a clock tower taken at intervals once every 55 minutes. Consecutive images would show the minute hand moving backwards five minutes, while the hour hand moves ahead normally almost one hour. If these (sampled) images were played back quickly, the apparent motion of the clock hands would be most peculiar, illustrating, like the wagon wheels, another example of aliasing.

   In this example with two periodic components, the hour hand is moving slowly enough to be accurately tracked at this sampling rate, while aliasing affects representation of the minute hand. Like the Fourier series components of a signal, the highest-frequency component present in a signal must be considered when choosing a sampling rate.

   The problem created by alias frequencies is a serious one, since the incorrect frequency component that is recovered falls within the range of acceptable frequencies for the sampling rate used. There is no way to tell, looking only at the samples, if the result is the true original signal or an artifact caused by aliasing. In the above example, someone unfamiliar with clocks would not notice any problem, and if the minute hand in a defective clock tower actually was moving backwards, the exact same pictures would be taken as in the example above.

3. Consider a signal $s(t) = \cos(18\pi t)$ to be manipulated in MATLAB and a time variable $t$ defined to have elements spanning the interval $[0, 10]$ seconds as shown below. This straightforward MATLAB task gives a surprising and undesirable outcome as seen when the function is plotted in Figure 8.11.

```
t = (0 : 0.1 : 10); % 101 points spanning 0 to 10 s
s = cos(18 * pi * t);
plot(t, s);
grid on;
```

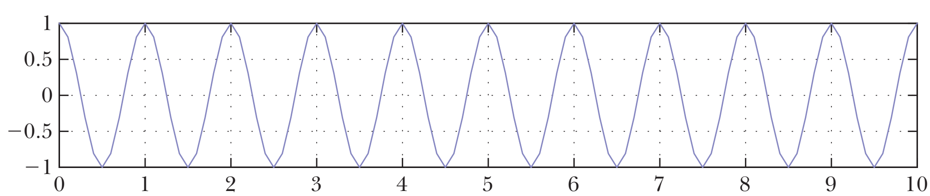



**FIGURE 8.11 MATLAB Alias Pitfall** The $s(t) = \cos(18\pi t)$ is defined in MATLAB on 101 points spanning 0 to 10 s. The effective sampling rate is only 10 Hz and the 9 Hz cosine signal appears as this 1 Hz alias frequency when plotted.

[2] Of course, special movie cameras designed to capture very fast events (such as a bullet piercing a balloon) must use extremely high frame rates.

By specifying a time axis with points every 0.1 seconds, a *sampling system* has effectively been created with a sampling rate of 10 Hz. Since the cosine has a frequency of 9 Hz, the alias frequency emerges when the function is plotted. As in the previous examples, there is no way to tell from the graph, or in calculations that use $s(t)$, that this was not actually a 1 Hz cosine.

This particular aliasing example is especially important as it could potentially arise any time a computer is used to process signals. In every case, the number of stored values spanning the dependent variable must implicitly represent a sampling rate sufficiently large for the signals encountered. In the above example, if 1000 points are used, or if the same points span 1 s instead of 10 s as `t = (0:0.01:1)`, then the effective sampling rate is 100 Hz, and there is no aliasing problem with this 9 Hz signal.

## 8.4.12 Analysis of Aliasing

If the undersampling operation is studied carefully in the frequency domain, the exact result can be predicted. In Figure 8.12, a 1 Hz cosine signal $s(t)$ is properly sampled at 3 Hz by multiplying $s(t)$ by an impulse train sampling function. In the

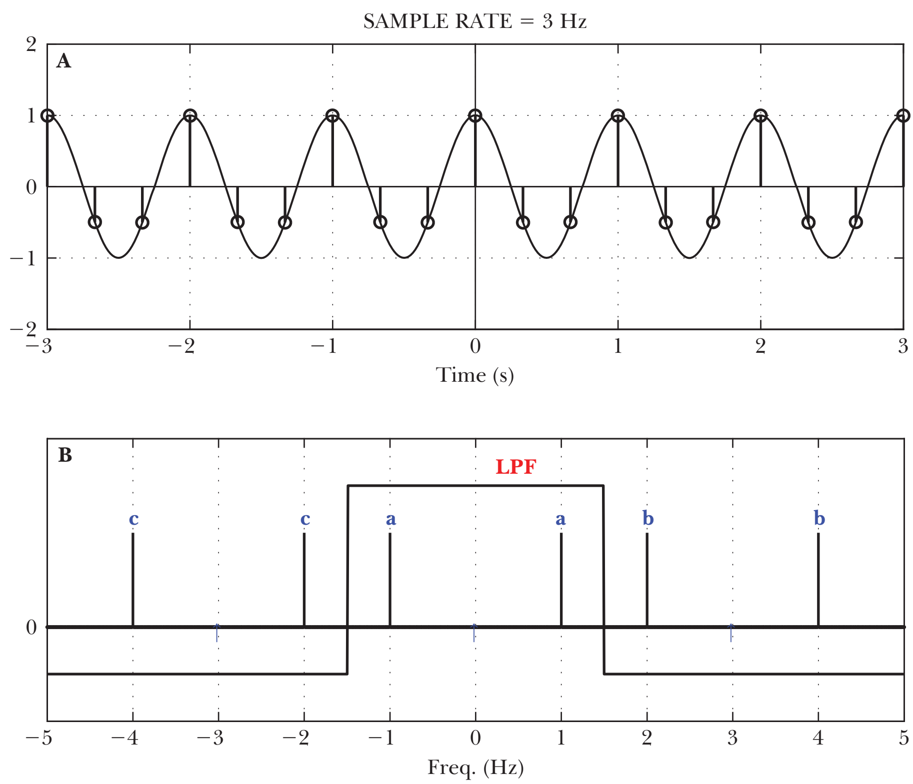



**FIGURE 8.12 Sampling and Reconstruction** A bandlimited signal **A** given by $s(t) = \cos(2\pi t)$ is sampled at a sample rate of 3 Hz. In the frequency domain **B** one copy of the original signal $S(f)$ is recovered by a lowpass filter with cutoff at one-half the sampling rate. Compare to Figure 8.13.

frequency domain, copies of $S(f)$ lie along the frequency axis, spaced every 3 Hz. The three copies close to the origin are labelled (a, b, c) for clarity. A lowpass filter with a cutoff frequency of one-half the sampling rate (1.5 Hz) is able to isolate the original $S(f)$, and the original continuous signal $s(t)$ is exactly recovered from the sampled signal.

In Figure 8.13, a 1 Hz cosine signal $s(t)$ is undersampled at 1.5 Hz by multiplying $s(t)$ by an impulse train sampling function. In the frequency domain, copies of $S(f)$ lie along the frequency axis, spaced every 1.5 Hz. The three copies close to the origin are labelled (a, b, c) for clarity; however, with this too-low sampling rate it is clear that, unlike in Figure 8.12, the copies overlap each other. A lowpass filter with a cutoff frequency of one-half the sampling rate (0.75 Hz) is not able to isolate the original $S(f)$; however, it does isolate a cosine at 0.5 Hz. (Of course, the labelling (a, b, c) has no bearing on this outcome.) This is the alias frequency.

Sampling a 1 Hz cosine at 1.5 Hz resulted in a 0.5 Hz alias frequency. This leads to an important conclusion, namely, *the alias frequency is the difference between the sampling frequency and the true cosine frequency*. This can be seen because the aliased cosine impulses lie at the difference between the spacing of each copy (the sampling rate) and span of each cosine pair (the cosine frequency).

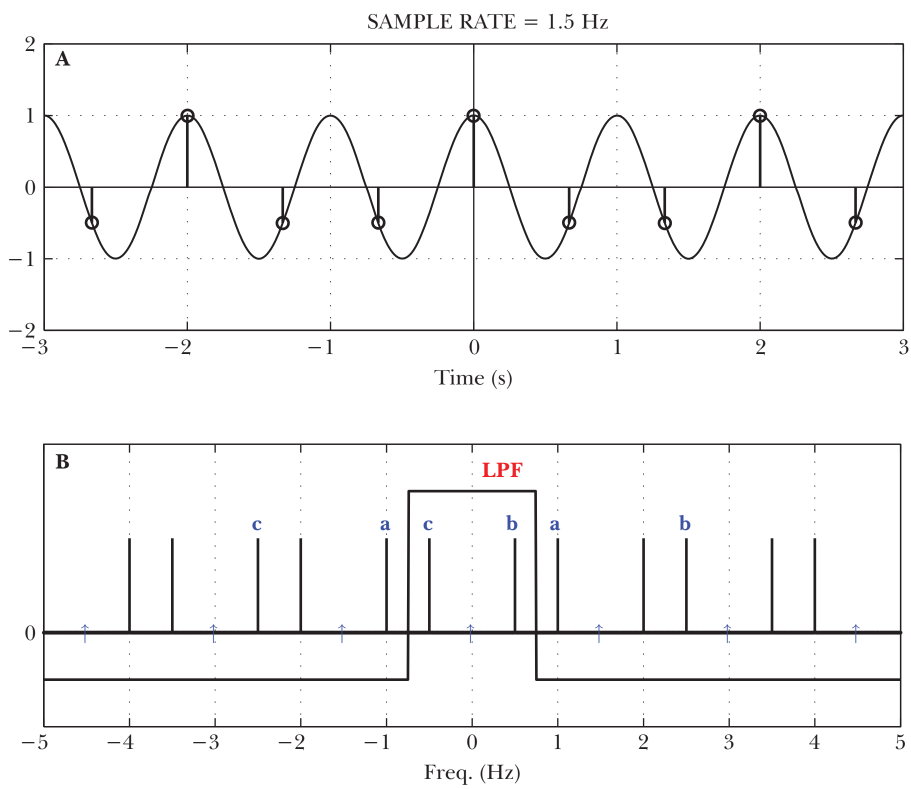



**FIGURE 8.13 Undersampling and Reconstruction** A bandlimited signal **A** given by $s(t) = \cos(2\pi t)$ is sampled at a (too low) sample rate of 1.5 Hz. In the frequency domain **B**, the signal recovered by a lowpass filter with cutoff at one-half the sampling rate is a cosine at 0.5 Hz. This is the alias frequency. Compare to Figure 8.12.

**EXAMPLE 8.2 (Alias Frequency)**

*Determine the alias frequency if a 1 Hz cosine is sampled at 0.25 Hz.*

**Solution:**
The difference between the sampling rate (0.25 Hz) and the too-high frequency component (1 Hz) gives the alias frequency (0.75 Hz).

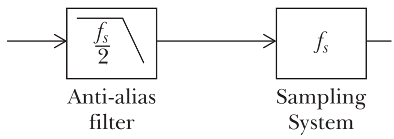



**FIGURE 8.14 Anti-Alias Filter** Even when a sampling rate $f_s$ is carefully chosen based on the expected bandwidth of an input signal $s(t)$, it is important to guarantee the sampling conditions by adding an $f_s/2$ Hz lowpass filter to the input.

## 8.4.13 Anti-Alias Filter

If the sampling rate is properly chosen, considering the highest-frequency component expected in an input signal, there is still the potential for trouble if an unexpected input signal component appears at the input of a sampling system. To avoid this possibility, a well-designed sampling system will avoid this possibility by incorporating a lowpass filter at the input to cut off any too high-frequency components. Such a filter is called an *anti-alias filter*, as shown in Figure 8.14.

For a sampling rate of $f_s$ Hz, the highest expected frequency is $f_s/2$ Hz. Consequently, the sampler should be preceded by a lowpass filter that passes only frequencies less than $f_s/2$ Hz, as shown in Figure 8.14. To be useful, the anti-alias filter must be a separate circuit installed before the sampling system as shown; it cannot be a digital filter implementation that itself depends on the same sampling rate as the input system.

It is surprisingly easy to violate the Nyquist rule in practice. Consider a system having a 2000 Hz sampling rate and designed to sample frequencies up to 1000 Hz. What could go wrong?

1. It might seem acceptable to use this system to sample an 800 Hz square wave; however, only the fundamental frequency of this input signal is 800 Hz, and all the harmonic components at higher frequencies would violate the sampling condition and lead to alias frequencies.
2. The same system might be used to sample an 800 Hz cosine that is squared before sampling. Squaring a sinusoidal signal doubles its fundamental frequency and, once again, violates the sampling condition.

   Moreover, squaring the 800 Hz cosine numerically *after* sampling and during calculations also violates the sampling rule. The choice of a suitable sampling rate must be respected throughout the signal-processing operation.

A sound understanding of signals theory is necessary to appreciate and to avoid these potential problems.

# 8.5 Introduction to Digital Filtering

If a bandlimited signal $s(t)$ is sampled at a suitable rate, then mathematical operations performed on those samples can constitute the response function of a sampled system. Such operations are performed on sampled values spaced by integral multiples of the sampling period. Consequently, it is useful to study the behavior of a discrete time response function $h(nT)$. It will be seen that any desired transfer function can be developed from a discrete time response function.

## 8.5.1 Impulse Response Function

In the time domain, a response function $h(t)$ is convolved with an input signal $s(t)$ to produce an output signal. If $h(t)$ takes the form of a single unit impulse $\delta(t-T)$, then the resulting convolution $s(t)*h(t-T)$ leads to the output $s(t-T)$, which is a single copy of $s(t)$ delayed by time $T$.

In general, response functions based on one or more impulses can also be described as *response functions based on delays*, or *response functions based on echoes* (since delayed copies of $s(t)$ really amount to echoes in time). This same process could be sketched in a block diagram by defining a *delay block*, as shown in Figure 8.15. Consequently, the following discussion of signal delays relates closely to the concept of a discrete time impulse response.

## 8.5.2 A Simple Discrete Response Function

Consider a linear system where a delay block combines a signal with a delayed copy occurring exactly $T$ s in the past, as shown in Figure 8.16.

If an impulse is input to this linear system, the output would consist of the impulse, followed $T$ s later by a copy of the impulse. By inspection, the causal response function is $h(t) = \delta(t) + \delta(t-T)$, as shown below Figure 8.17.

What is the effect of this response function $h(t)$ on an input signal? It is useful to think about how this system would behave in the time domain, as every point in an input signal is added to the point $T$ s in the past to create the output signal.


**FIGURE 8.15 Delay Block** The input signal $a(t)$ passes through an delay block to emerge $T$ s later as the output signal $a(t-T)$.

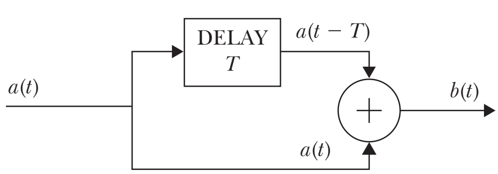



**FIGURE 8.16 System with Delay Block** This simple system includes a delay block such that the output $b(t) = a(t) + a(t-T)$. A cosine of period $2T$ is completely blocked by this system.

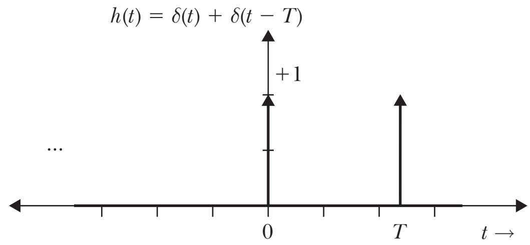



**FIGURE 8.17** The system impulse response from Figure 8.16 is $h(t) = \delta(t) + \delta(t - T)$.

1. For a constant (DC) input, the steady-state output will be doubled in amplitude, since the onset of a constant input signal would give a constant output (at $t = 0$) followed immediately by a second copy (at $t = T$) added to the first. Therefore, this circuit amplifies a signal at zero frequency.
2. Next, consider a cosine input signal $s(t) = \cos(\pi t/T)$. This cosine has period $2T$, and the sum of any two values of $s(t)$ spaced by one-half its period will exactly cancel. This circuit completely blocks a sinusoid with frequency $1/2T$ Hz. Similarly, any frequency components at odd integer multiples of $1/2T$ Hz will be blocked.
3. A cosine input with double the above frequency $s(2t) = \cos(2\pi t/T)$ will have exactly one period between the two impulses, such that the signal and its copy will add to double the cosine amplitude, like the DC component. A careful examination of the convolution of $h(t)$ and $s(2t)$ will reveal that the output cosine is also inverted. Similarly, frequencies at even integer multiples of $s(t)$ will all be doubled in amplitude.

Examination of the above system has revealed the overall behavior resulting from the use of a delay of $T$ seconds in the circuit. The system appears to block all frequencies at exactly $(2n + 1)/2T$ Hz, for integer $n$, and doubles the amplitude of all frequency components at exactly $n/T$ Hz, including the DC component. What of the other frequencies? These results can now be confirmed and extended by determining the frequency response $|H(f)|$ of the circuit.

The impulse response is

$$h(t) = \delta(t) + \delta(t - T)$$

and by the time-shifting theorem:

$$H(f) = 1 + e^{-j2\pi fT}$$

The transfer function $H(f)$ will give the frequency response $|H(f)|$ of this system. By inspection, $h(t)$ consists of two impulses spaced by $T$. If the response function is shifted left by $T/2$ s, $h(t + T/2)$ looks like Figure 8.18.

The shifted impulse response is

$$\begin{aligned} h\left(t + \frac{T}{2}\right) &= \delta\left(t + \frac{T}{2}\right) + \delta\left(t - T + \frac{T}{2}\right) \\ &= \delta\left(t + \frac{T}{2}\right) + \delta\left(t - \frac{T}{2}\right) \end{aligned}$$

and by the time-shifting theorem:

$$H(f) = [e^{+j2\pi fT/2} + e^{-j2\pi fT/2}] = 2\cos(\pi fT)$$

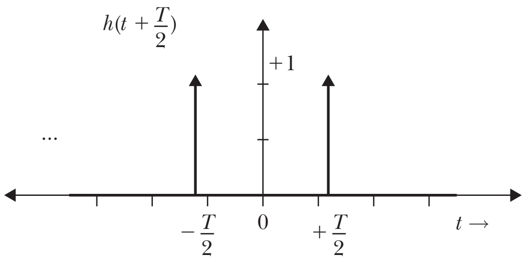



**FIGURE 8.18** The system impulse response from Figure 8.16 is shifted to give $h(t+\frac{T}{2}) = \delta(t+\frac{T}{2}) + \delta(t-\frac{T}{2})$.

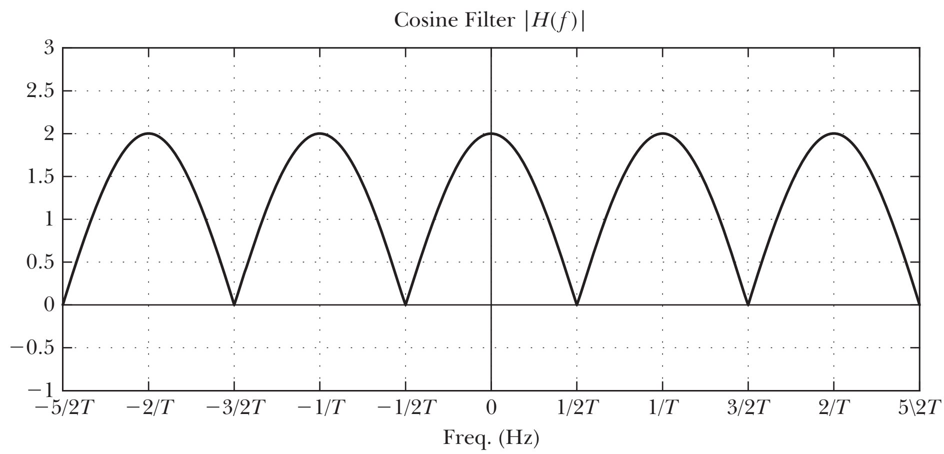



**FIGURE 8.19 Discrete Cosine Filter Transfer Function (System with Delay Block)** The magnitude $|H(f)|$ shows that all frequencies at odd multiples of $1/2T$ Hz are completely blocked by this system.

Except for the effect of the time shift, the Fourier transform $H(f)$ is a cosine shape in the frequency domain with period $2/T$ Hz. The magnitude $|H(f)|$, which does not depend on shift, is shown in Figure 8.19, or

$$|H(f)| = 2|\cos(\pi f T)|$$

This figure shows how the delay block implements a simple filter.

This circuit is known as a *cosine filter*. This function must be multiplied in the frequency domain by the Fourier transform of an input signal. As predicted, it can be seen that the system would double the amplitude of any DC component and completely block components at odd-numbered multiples of $f_0/2$. This frequency-domain behavior is confirmed by the time-domain observations made earlier.

## 8.5.3 Delay Blocks Are a Natural Consequence of Sampling

In constructing the above system, it was assumed that a delay block circuit would somehow be constructed. In practice, no such circuit need ever be built. Consider a sampling system with samples taken every $T$ s. In effect, all the above system really does is add each sample to the previous sampled point collected $T$ s earlier.

The cosine filter may be constructed mathematically through the addition of sampled points, and the response function will be exactly the same as that analyzed above.

If the sampled points are stored in a computer as a vector `sample[x]`, then the following code segment written in a high-level programming style would filter the stored waveform using the above $|H(f)|$.

```
for x = 1 to total{output[x] = sample[x] + sample[x - 1];}
```

Furthermore, if such a sampling system employs a sampling interval of $T$ s, it can be assumed that *no frequency components greater than or equal to* $1/2T$ *are ever expected* in this system. In such a bandlimited system, the cosine filter only applies to frequencies less than $1/2T$ Hz. In other words, this system represents a kind of lowpass filter, favoring near-DC frequencies. The practice of modelling discrete time response functions for sampled systems is fundamental to digital signal processing.

## 8.5.4 General Digital Filtering

The simple delay block circuit of Figure 8.15 adds an input signal to a delayed copy of itself, leading to the cosine filter. By adding additional delay blocks, each with a different weight combining to form the output signal, a variety of different response functions can be created, allowing the design of customized (configurable) digital filters for any application.

In a more general sense, the filter above may be extended to give the causal system in Figure 8.20.

The time-domain impulse response is sketched in Figure 8.21 and written as:

$$h(t) = \sum_{n=0}^{4} a_n \delta(t - nT)$$

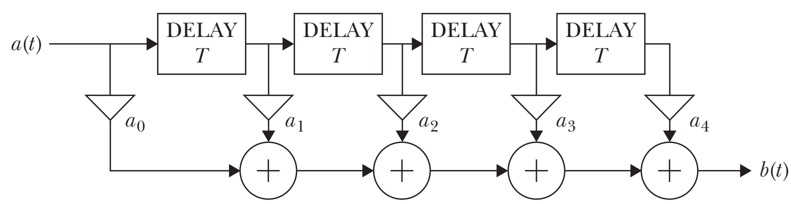



**FIGURE 8.20 General Delay Block System** This configurable system includes multiple delay blocks with variable weighting $a_n$. The system could be extended to incorporate many stages. The system of Figure 8.16 can be constructed by setting $a_0 = 1$ and $a_1 = 1$ and all other $a_n = 0$.

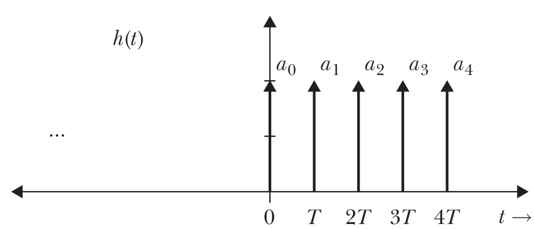



**FIGURE 8.21** System impulse response $h(t)$ for a configurable filter. The system behavior can be modified by varying each $a_N$.

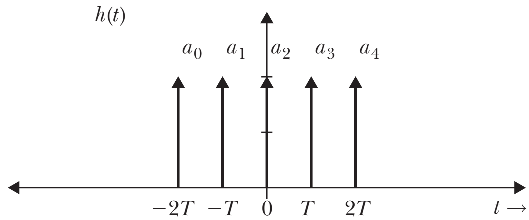



**FIGURE 8.22** Shifted system impulse response $h(t + 2T)$ from Figure 8.21 for a configurable filter. This sketch describes the Fourier series components for a periodic frequency-domain signal of period $1/T$ Hz.

and by the time-shifting theorem:

$$H(f) = \sum_{n=0}^{4} a_n e^{-j2\pi fnT}$$

since each shifted impulse corresponds to a frequency term in $e^{-j2\pi fnT}$.

Shifting the above impulse response by $2T$ s as $h(t+2T)$ gives Figure 8.22.

It may now be observed that the signal $h(t + 2T)$ in Figure 8.22 describes the terms in a Fourier series representation of a periodic frequency-domain waveform. In the limit, as more delay blocks are added to the system, the appearance of the frequency-domain waveform can be varied by adjusting $a_n$. Furthermore, since the frequency interval of interest is limited to one period (in Hz) by the sampling rules, it is now possible to imagine the design of digital filters of arbitrary shapes, working on sampled values of an input signal.

Note that shifting $h(t)$ made the result $h(t + 2T)$ noncausal; however, the only difference between the filter formed by the original, causal, $h(t)$ and the shifted version is a time delay, or a linear phase shift in the associated transfer function. In summary:

1. Digital filters operating on sampled signals can be made any shape by choosing appropriate values of $a_n$, and;
2. Appropriate values of $a_n$ can be chosen by computing the (inverse) Fourier series components of the desired filter shape in the frequency domain.

## 8.5.5 The Fourier Transform of Sampled Signals

A sampled signal $s(t)$ consists of discrete values $s(nT)$ taken at regular intervals $T$ s, and the Fourier transform $S(f)$ of the signal $s(nT)$ will be a periodic function of frequency with period $1/T$ Hz.

$$S(f) = \int_{-\infty}^{+\infty} \sum_{n=-\infty}^{\infty} \delta(t - nT)\, s(t) e^{-j2\pi ft}\, dt \tag{8.1}$$

This resulting signal $S(f)$ is called the the *discrete time Fourier transform* (DTFT). In Figure 8.23, an even unit-area triangle is sampled at $T = 0.1$ Hz to create the discrete time signal $s(nT)$. The Fourier transform of this signal is the continuous function $\text{sinc}^2(f)$ repeated periodically every $1/T$ Hz as shown in the lower part of Figure 8.23.

It has been shown that a sufficiently high sampling rate can be chosen to completely represent a bandlimited signal $s(t)$ so this $S(f)$ is an accurate frequency-domain representation wherein one period found at the origin $f = 0$ will be the

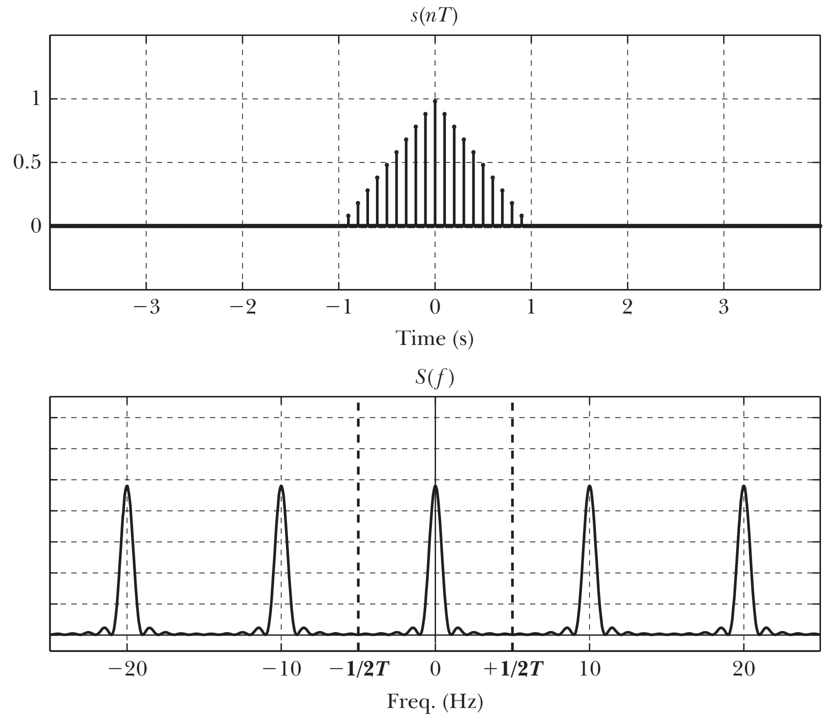



**FIGURE 8.23 Discrete Time Fourier Transform (DTFT)** The Fourier transform of the discrete time unit-area triangle $s(nT)$ is a continuous function of frequency $S(f)$ with period $1/T$ Hz. The single $\mathrm{sinc}^2(f)$ at the origin in the frequency domain is recognized as the Fourier transform of the original (unsampled) triangle.

Fourier transform of the original signal $s(t)$ before sampling.[3] Since this DTFT integral includes an impulse function, it may be simplified directly to give:

$$S(f) = \sum_{n=-\infty}^{\infty} s(nT)e^{-j2\pi fnT} \tag{8.2}$$

It follows that the inverse DTFT relates $S(f)$ back the the original samples by taking the inverse Fourier transform over one complete period of $S(f)$. Using the limits shown in Figure 8.23:

$$s(nT) = \int_{-T/2}^{+T/2} S(f)e^{+j2\pi fnT}\,df \tag{8.3}$$

If the signal $s(t)$ is nonperiodic, as above, then the DTFT will be a continuous function of frequency $S(f)$. If the signal $s(t)$ is periodic with period $(1/F_0$ s), then the DTFT will be a discrete frequency signal with components at multiples of the fundamental frequency $F_0$ Hz. The next section deals with the most practical case of signals where both the time- and frequency-domain are sampled, implying periodic signals in both domains regardless of the actual periodicity of the original signals.

[3]While not strictly bandlimited, most of the energy in the example triangle $s(t)$ is concentrated in frequencies close to $f = 0$.

## 8.5.6 The Discrete Fourier Transform (DFT)

The discrete Fourier transform (DFT) is a Fourier transform in which both the time- and frequency-domain signals are discrete (sampled). In practice, the DFT will necessarily use a fixed number of samples ($N$) taken from a (usually) continuous waveform. (For example, a vector in MATLAB may contain $N = 128$ samples taken from a much longer time-domain waveform.) The properties relating discrete signals in the time and frequency domains follow the same basic rules as continuous signals.

It has been shown that the Fourier transform $S(f)$ of a periodic signal $s(t)$ has the form of an infinite series of impulses spaced at the (implicit) period of the time-domain signal. Similarly, the inverse Fourier transform of a signal that is periodic in frequency would yield a time-domain signal consisting of regularly spaced impulses. It has also been shown that the Fourier transform of a sampled signal is periodic in frequency. This property was demonstrated in the preceding section on sampling and reconstruction, where *multiple copies spaced at the sampling rate* characterized the Fourier transform of a sampled signal. It follows that the Fourier transform of a sampled periodic signal will be both periodic and discrete, as shown in Figure 8.24. The result is periodic because the input signal is discrete and discrete because the input signal is periodic.

One of the motivations for sampling signals and for studying discrete time signals was the observation that continuous signals must be sampled if they are to be stored and processed within a digital computer. This fact is also implicit in all calculations performed with MATLAB. Just as a time-domain signal must be stored in discrete form, its Fourier transform must be computed on samples and stored as a sampled signal in the frequency domain. It follows that both the time- and frequency-domain versions of a sampled signal are both discrete and periodic. If the Fourier transform is computed on $N$ samples, the underlying implication is that the samples represent one period of a waveform that goes on for all time. Figure 8.25 shows a close-up of Figure 8.25 in which the Fourier transform is computed on 16 samples starting at the origin.

It should also be noted that computations performed on any real signal must take place on a limited time segment of the signal. For example, ten thousand samples of a continuous signal sampled at 1000 Hz represent only a ten-second segment extracted from a much longer original signal. This effect may be likened to multiplying

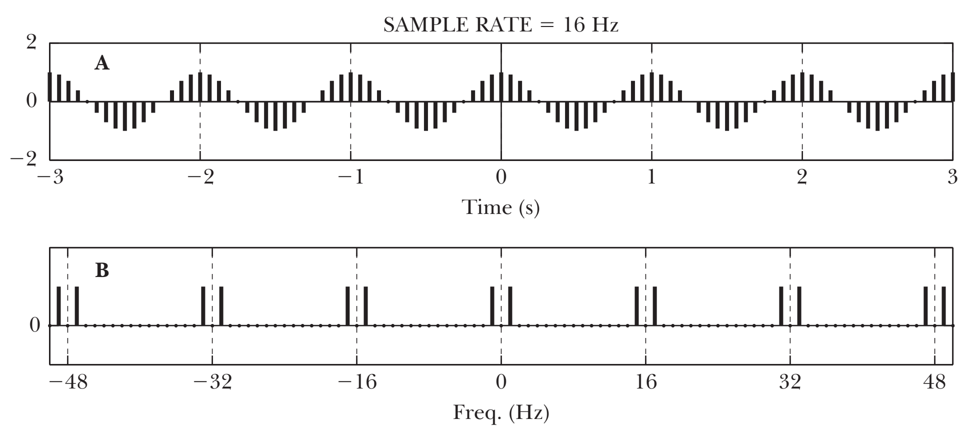



**FIGURE 8.24** **A 1 Hz cosine sampled at 16 Hz** The Fourier transform of a discrete periodic signal is also periodic and discrete.

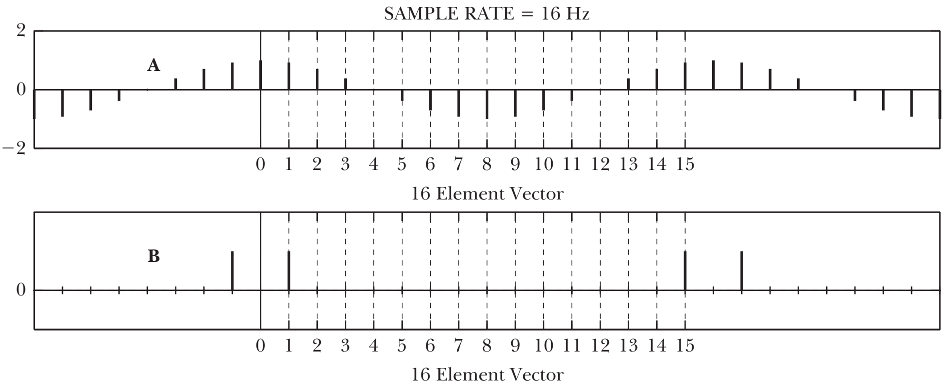



**FIGURE 8.25 A 1 Hz cosine sampled at 16 Hz—Sample Vector** A close-up view of Figure 8.24 shows how the corresponding stored samples each span sixteen elements starting at the origin. Both the time- and frequency-domains signals are implicitly periodic.

a continuous signal by a ten-second-wide rectangle, effectively cutting off everything that passed before, or that came after, the sampling process.

Recall that when computing the Fourier series components of a periodic signal, it is only necessary to perform calculations on a single period for the approximation to be valid for the entire periodic signal. Furthermore, the issue of computing the Fourier series on nonperiodic signals was conveniently handled by considering a single pulse, for instance, as part of a pulse train having an exceedingly long period. Isolating a single period of a signal is much like multiplying a periodic signal by a rectangle one period wide.

## 8.5.7 A Discrete Fourier Series

Consider the complex Fourier series representation of a signal. Each of up to $N$ frequency components found at integer multiples of the fundamental frequency is represented by a pair of complex values, $C_{\pm N}$ plus the DC component $C_0$. Like the samples of a time-domain signal, these values can be stored in a vector, in this case requiring $2N + 1$ storage elements. The lowest-frequency component ($C_1$) represents the fundamental frequency $f_0$, corresponding to a single period of the time-domain signal. The highest-frequency component ($C_{\pm N}$) represents the highest frequency ($Nf_0$) present in this approximation to the time-domain signal. In other words, the corresponding time-domain signal has components only up to $Nf_0$ Hz.

The Fourier series components in Figure 8.26 have a maximum frequency of $Nf_0$ where $N = 4$. By inspection, this Fourier series corresponds to a time-domain signal that is periodic with a period of $1/f_0$ seconds. The signal of interest in the time domain spans a full period of $T_0 = 1/f_0$ seconds, as the Fourier series is always computed over one complete period.

Given a time-domain signal that is discrete, the maximum frequency component is known to be $Nf_0$ Hz, since this signal is sampled at a rate of at least $2Nf_0$ Hz. The total span in samples of one period of the time-domain signal is the width in seconds ($1/f_0$) multiplied by the number of Hz ($2Nf_0$). In other words, a discrete time-domain signal spanning $2N$ samples corresponds to a two-sided complex Fourier series representation containing $2N$ values spanning frequencies from $-Nf_0$ to $+Nf_0$ Hz, plus the DC component.

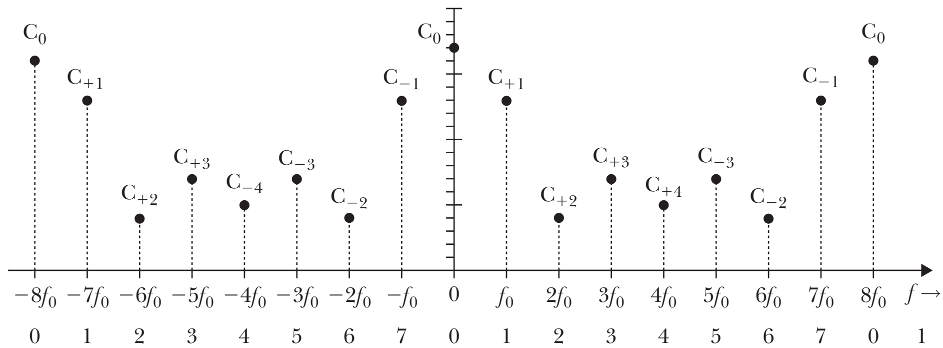



**FIGURE 8.26 Discrete Fourier Series** One period of a discrete frequency-domain signal spans $N = 8$ points, where the eight unique components up to frequency $4f_0$ Hz are numbered for reference. Note that the component numbered 4 overlaps with adjacent copies of this periodic signal. The exact frequency of each component is irrelevant to the calculation and is known only from the sampling rate $1/f_0$ sec.

Finally, note that the assumption that the time-domain signal is discrete implies that the frequency-domain signal must be periodic. The samples in the time domain are taken at a rate of $2Nf_0$ Hz, so one period is $2Nf_0$ Hz wide. In other words, the spectrum in Figure 8.26 represents only *a single period* of a periodic frequency-domain signal. If the next copy of the spectrum is located at $f = 2Nf_0$, then the component at exactly $f = Nf_0$ is located exactly on the boundary between each period. In Figure 8.26, the components numbered 4 overlap with the adjacent copies.

For example, one period of a time-domain signal stored as 2048 consecutive samples would give a periodic Fourier series having a period of 2048 points and spanning 1024 frequency components (pairs).

## 8.5.8 Computing the Discrete Fourier Transform (DFT)

Consider the discrete time Fourier transform (DTFT) of a continuous signal $s(t)$ that has been sampled every $T$ s over an interval spanning $N$ samples (or $NT$ s in total) where a sample at time $n$ is the value $s(nT)$, for integer $n$ in the interval $0 \leq n < N$. This was found from Eqn. 8.2 with exactly $N$ sample points:

$$S(f) = \sum_{0}^{N-1} s(nT)e^{-j2\pi fnT}$$

In the DFT, the frequency-domain components are also discrete. If the time-domain interval spanning $NT$ s is defined as one period, then discrete frequency components will lie at multiples of the fundamental frequency defined by $f_0 = 1/NT$ Hz. Therefore, $S(f)$ becomes $S(mf_0)$ or $S(m/NT)$, for integer $m$ where:

$$S(mf_0) = S(m/NT) = \sum_{0}^{N-1} s(nT)e^{-j2\pi mn/N}$$

It is normal to scale this result by a factor $N$ to give the usual form of the discrete Fourier transform (DFT):

$$S(mf_0) = \frac{1}{N}\sum_{0}^{N-1} s(nT)e^{-j2\pi(mn)/N}$$

This formulation will now be explored in more detail.

### 8.5.9 The Fast Fourier Transform (FFT)

Computation of the discrete Fourier transform equation requires that for every point $m$ in the frequency domain, the summation over all $n$ must be computed. Each computation requires evaluation of the complex exponential (possibly through a lookup table), followed by a multiplication by $s(nT)$. If there are as many points in time as in frequency, then this operation requires on the order of $N^2$ multiply and sum operations (often referred to as *multiply-and-accumulate* or MAC operations). For 1000 points, the requirement to perform over one million MAC operations detracts from the potential usefulness of the DFT. It is simply too slow for practical applications.

Recall that one application of the Fourier transform is to avoid the computationally demanding convolution operation by performing one simple multiplication in the frequency domain. If transforming a signal to the frequency domain is extremely difficult, then nothing is saved.

In 1965, a new algorithm to perform this calculation more efficiently was published.[4] The *fast Fourier transform* revolutionized signal processing by reducing the above computational requirements from $N^2$ to only $N\ log_2 N$. The DFT on $N = 1024$ points now required only about 10,000 MAC operations (since $log_2 1024 = 10$). The advantage of this new computational approach increases dramatically as the number of points grows. It is common today to hear about the fast Fourier transform (FFT) as if it was different from the DFT. The main thing different about the FFT as compared to the DFT is that the FFT is faster. Furthermore, optimization of the manipulations to compute DFT quickly requires that the value $N$ must be a power of two (8, 16, 32, 64 . . .). This is not a significant limitation for most applications.

In MATLAB, the function `fft` is used universally for the discrete Fourier transform and may be applied to a vector of any length; however, unless the vector length is a power of two, a (slow) DFT algorithm will be used to compute the answer. In either case, many people will refer to *taking the FFT* as if somehow this this was a new and mysterious transform technique (and one more thing to learn). What these people mean to say is simply *computing the DFT efficiently.*

## 8.6 Illustrative Examples

The simple scenario from Figure 8.26 with only eight sample points will now be examined in detail to illustrate the important concepts behind the discrete Fourier transform. Given a sample rate of $1/T$ Hz, the complete period defined by eight samples is $8T$ s long. The frequency content of this one period is discrete and spans eight points representing the components from $-4/T$ to $4/T$ Hz. The discrete frequency-domain signal is also periodic, with a period of $8/T$ Hz. These examples use MATLAB to compute the DFT derived above.

1. **DC Component**—Figure 8.27A shows a constant value in the time domain, sampled regularly for eight samples as [1 1 1 1 1 1 1 1]. The corresponding discrete Fourier transform is shown at right in Figure 8.27 as a single unit value (impulse) at $f = 0$, described by the vector [1, 0, 0, 0, 0, 0, 0, 0].

   The result should be appreciated on two levels. Within the smaller window spanning one period around the origin as seen here, a single impulse at the

[4] James W. Cooley; John W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comp.* 19 (1965), 297–301.

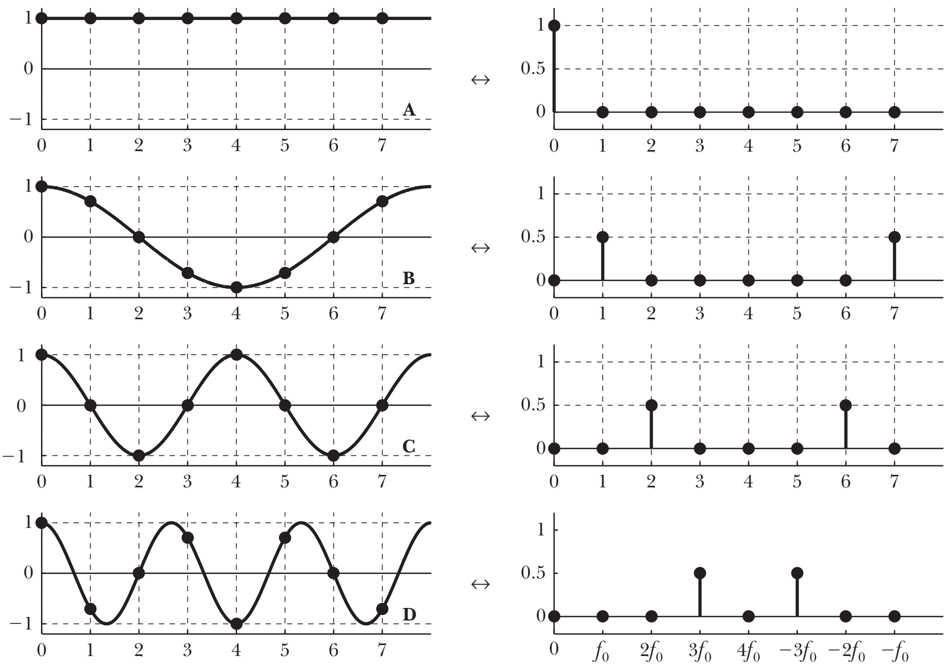


**FIGURE 8.27 8-Point FFT Examples** The 8-point discrete signal `s[n]` and the corresponding discrete Fourier transform result `fft(s)/8` shown for four examples. Lines connect the discrete sample values for clarity.

origin corresponds to a constant in the transform domain. In the overall signal for all time, an impulse train in one domain (one impulse per sample time) gives an impulse train in the other (one impulse every eight sample times).

2. **Fundamental Frequency Component**—The fundamental frequency component is the first non-zero frequency component in the signal, corresponding to a complete period of cosine in the transform domain. The EVEN pair of impulses located at $+f_0$ and $-f_0$ Hz can be defined starting the origin by the vector [0, 1, 0, 0, 0, 0, 0, 1] with impulse at components 1 and 7 as in Figure 8.27B. If these components are equal, then the resulting frequency-domain signal is EVEN and periodic, and the time-domain signal should be a single period of cosine. Eight samples beginning at the origin are sufficient to define this EVEN periodic signal for all time.

   The discrete Fourier transform relates only these eight frequency-domain samples [0, 1, 0, 0, 0, 0, 0, 1] to the eight time-domain samples describing a cosine. Again, a full appreciation of the DFT requires knowledge of what is understood to be present outside of the sampled region (the smaller windows).
3. **Next-Higher Frequency Component**—The next frequency component corresponds to two complete periods in the transform domain. The EVEN pair of impulses located at $+2\,f_0$ and $-2\,f_0$ Hz can be defined starting the origin by the vector [0, 0, 1, 0, 0, 0, 1, 0] and numbered 2 and 6 in Figure 8.27C. If these components are equal, then the time-domain signal should be two complete periods of a cosine. Eight samples beginning at the origin are sufficient to define this EVEN periodic signal for all time.
4. **Next-Higher Frequency Component**—The next frequency component corresponds to three complete periods in the transform domain. The EVEN

pair of impulses located at $+3\,f_0$ and $-3\,f_0$ Hz can be defined starting the origin by the vector [0, 0, 0, 1, 0, 1, 0, 0] and numbered 3 and 5 in Figure 8.27D. If the components are equal, the time-domain signal should be three complete periods of a cosine. Eight samples beginning at the origin are sufficient to define this EVEN periodic signal for all time.

5. **Highest Frequency Component**—The final frequency component corresponds to four complete periods in the transform domain. The EVEN pair of impulses located at $+4\,f_0$ and $-4\,f_0$ Hz can be defined starting the origin by the vector [0, 0, 0, 0, 1, 0, 0, 0]. It seems unusual to define an impulse pair with only a single component, but the appearance of this signal in the frequency-domain confirms that the resulting frequency-domain signal is EVEN and periodic. Essentially, this frequency component lies on the boundary between one period and the next, at number 4 in Figure 8.26.

   The time-domain signal should be four complete periods of a cosine. Eight samples beginning at the origin are sufficient to define this EVEN periodic signal for all time. However eight samples on four periods is at the Nyquist limit, and no higher-frequency components can be present within this sample period. Note that the cosine has half the amplitude compared to the previous examples; only a single impulse contributed to the power in this signal.

The discrete Fourier transform defines a signal in discrete form on sample points, yet calculations take place on an array of points with no reference to the sampling rate. This is not an issue insofar as the *appearance* of a cosine or of its Fourier transform components (simply labelled $f_0$ during calculations) is the same regardless of its period. While the expected spacing in the time and frequency signals are inversely related, the specific time (s) or frequency (Hz) of a given sample is only known when the sampling rate is consulted.

The above examples each use 8 sample points. Let the sampling rate be 1000 Hz. The time domain now spans 8 msec, and therefore the fundamental frequency $f_0 = 1/.008 = 125$ Hz. Components in the frequency domain are found at $Nf_0 = (0, 125, 250, 375)$ Hz.

As a practical issue, *these are the only four frequencies that can be represented exactly in this calculation.* Moreover, their definition depends on the sampling rate such that in a real system there will always be some approximation in the computed components.

**EXAMPLE 8.3**

*Find the spacing between frequency components* ($f_0$) *if the discrete Fourier transform is performed on 1024 sample points taken at 1000 Hz.*

**Solution:**

The time domain spans $1024 \times .001$ s $= 1.024$ s, and the fundamental frequency would be $f_0 = 1/1.024 = 0.977$ Hz. Frequency components are found at $nf_0$ Hz for $n =$ 0 to 511. The highest frequency component is $511\,f_0 = 499.02$ Hz.

**MATLAB Exercise 1: The FFT and the Inverse FFT** The preceding examples can be reproduced directly in MATLAB using the `fft` and `ifft` functions. These built-in functions compute the discrete Fourier transform and inverse discrete Fourier transform, respectively, on an input vector consisting of equally spaced signal samples in either domain. In the following examples, a simple 8-element vector is used as found above and in Figure 8.26.

1. A single impulse at the origin above gives a constant in the transform domain:

```
ifft([1,0,0,0,0,0,0,0])

ans =   Columns 1 through 4
 0.1250   0.1250   0.1250   0.1250

Columns 5 through 8
 0.1250   0.1250   0.1250   0.1250
```

2. The `fft` and `ifft` also differ by a scaling factor ($N$), as shown below, where the `fft` on the same $N = 8$ points as above is scaled by $N$ compared to `ifft`.

```
fft([1,0,0,0,0,0,0,0])

ans =  1  1  1  1  1  1  1  1
```

3. In general, the `fft` and `ifft` operations return complex results. The above results were purely real because the input was even and real.

   If the frequency domain signal is real and odd [0 1 0 0 0 0 0 −1], the result is an imaginary sine waveform in the time domain. MATLAB gives the purely imaginary result as the complex values below:

```
ifft([0,1,0,0,0,0,0,-1])

ans =
  Columns 1 through 3
    0                  0.0000 + 0.1768i    0.0000 + 0.2500i

  Columns 4 through 6
    0.0000 + 0.1768i   0                   0.0000 - 0.1768i

  Columns 7 through 8
    0.0000 - 0.2500i   0.0000 - 0.1768i
```

4. If a real impulse is shifted as [0 1 0 0 0 0 0 0 ], the result reflects the phase shift by a constant magnitude signal having both real EVEN (cosine) and imaginary ODD (sine) components.

   MATLAB gives the complex components, and their magnitudes are shown to be constant:

```
result = ifft([0,1,0,0,0,0,0,0])

result =
  Columns 1 through 4
    0.1250    0.0884 + 0.0884i   0.0000 + 0.1250i   -0.0884 + 0.0884i

  Columns 5 through 8
   -0.1250   -0.0884 - 0.0884i   0.0000 - 0.1250i    0.0884 - 0.0884i
```

```
magnitude = abs(result) % compute magnitude

magnitude =
  Columns 1 through 4
    0.1250 0.1250 0.1250 0.1250
  Columns 5 through 8
    0.1250 0.1250 0.1250 0.1250
```

As expected, this magnitude is the same as that of the unshifted signal from Section 8.6.

### 8.6.1 FFT and Sample Rate

When the FFT of a sampled signal is computed, an interpretation of the resulting spectrum depends on the original sample rate. For example, the square wave signal in Figure 8.28 with samples numbered from 1 to 1024 results in a familiar set of discrete frequency components (magnitudes) but what frequencies should be associated with the peaks shown in at frequency samples numbered (8, 24, 40, 56...)?

Let the sample rate of the time-domain signal $s(nT)$ be $f_s = 1/T$ Hz. Let the number of sample points be $N$ such that there are $N = 1024$ sample points in both the time- and frequency-domain signals. The total interval spanned by the time-domain samples is $NT$ s, or $N/f_s$ s. Since the output from the FFT calculation also spans $N$ points, then the spacing between frequency samples is given by:

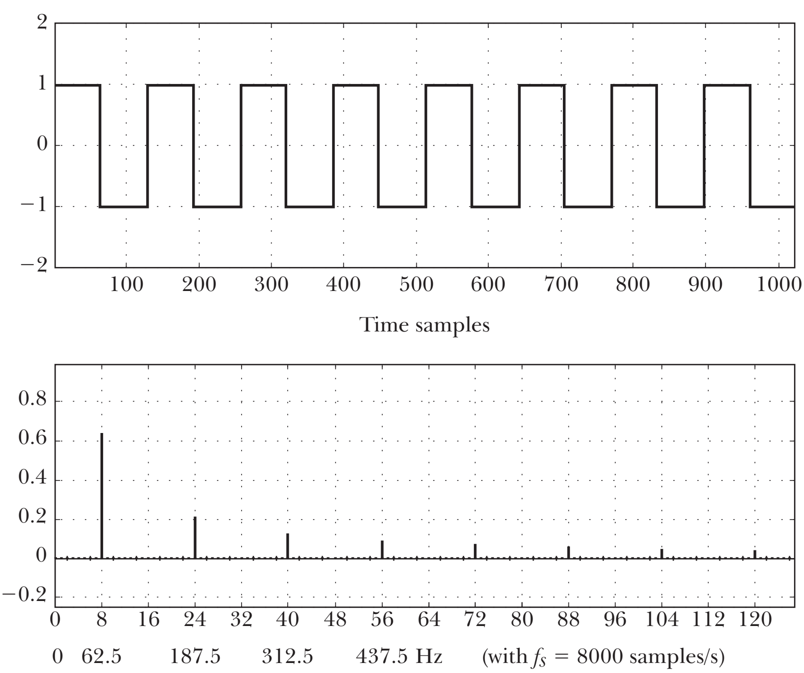



**FIGURE 8.28 1024-Point FFT Example—Sampled signal** The first 128 points of the discrete Fourier transform of the 1024-point discrete signal $s(nT)$ can be re-labelled meaningfully if the original sampling rate $f_s = 1/T$ Hz is known. For $f_s = 8000$ Hz, $\Delta f = f_s/N = 7.8$ Hz.

$$\text{Frequency-domain sample interval} = \Delta f = \frac{f_s}{N}\,\text{Hz}.$$

**Check:** Let the sample rate in Figure 8.28 be 8000 Hz. Then $f_s = 1/T = 8000$ Hz, the 1024 samples in the upper graph span an interval $NT = 1024/8000 = 0.128$ s or eight periods of a square wave with a fundamental frequency of $8/0.128 = 62.5$ Hz. In the frequency domain, the 1024 samples are spaced every $f_s/N$ Hz, or $8000/1024 = 7.8$ Hz, so the fundamental frequency component at sample 8 has frequency $8 * 7.8 = 62.5$ Hz as expected. This figure could now be re-labelled accordingly.

## 8.6.2 Practical DFT Issues

A number of issues conspire to give sometimes unexpected results when using the discrete Fourier transform. Differences from example problems can often be traced to several interrelated considerations:

1. Discrete input samples are spaced at the sampling rate;
2. There is implicit periodicity in any block of samples used for the DFT calculation;
3. The sample block generally fails to span complete periods of the input signal or its frequency components.

**Constructing the Ideal Discrete Signal** Textbook examples can generally be constructed so that a sample vector describes complete periods of an input signal, although even that is not always easy to achieve. For example, to define 128 time samples on $T$ s requires that 128 equally spaced points at multiples of $T/128$ s span the interval $0 \leq t < T$, specifically not including $T$ as that is also the first element of the next period. With $T = 1$ and $N = 128$ points:

```
N = 128; % define 128 samples
```

the required time interval is now defined as:

```
t = 0 : 1/N : (N - 1)/N;
```

where $t$ now goes in 128 steps from 0 to 127/128. Once the time variable is defined, a corresponding sinusoid can be created having an integral period; in this case, 10 Hz:

```
s = cos(2 * pi * 10 * t);

plot(t,s);
```

and the discrete Fourier transform of $s(t)$ has the expected properties as shown in Figure 8.29.

```
S = fft(s)/N;

stem(real(S(1 : 20)));  % plot the real part only
```

The first element in this frequency vector corresponds to zero frequency. Using MATLAB, where vectors are not zero-based the $\pm 10 f_0 = \pm 10$ Hz frequency components are found at elements `S(11)` and `S(119)` as:

```
S(11)
  0.5000 + 0.0000i
S(119)
  0.5000 + 0.0000i
```

In Figure 8.29, the sample block of 128 points spans one second in the time domain, and the 10 Hz cosine exactly spans the block. When the DFT is computed, this block implicitly represents one period of a signal defined for all time, as shown in the figure before and after the numbered samples. In the frequency domain (below) showing only the first 20 points, the 10 Hz component is exactly represented by a distinct component.

**A Typical Discrete Signal** In practice, it would be unlikely that $N$ input samples would span exactly a complete period of an input sinusoid, and a more realistic example uses a non-integral period. For example, if the cosine frequency is 10.5 Hz as in Figure 8.30, there is no way to represent exactly 10 Hz in the frequency domain. Moreover, the 128 time samples define a signal having a period of one second that is no longer EVEN and that has significant discontinuities at the origin and every

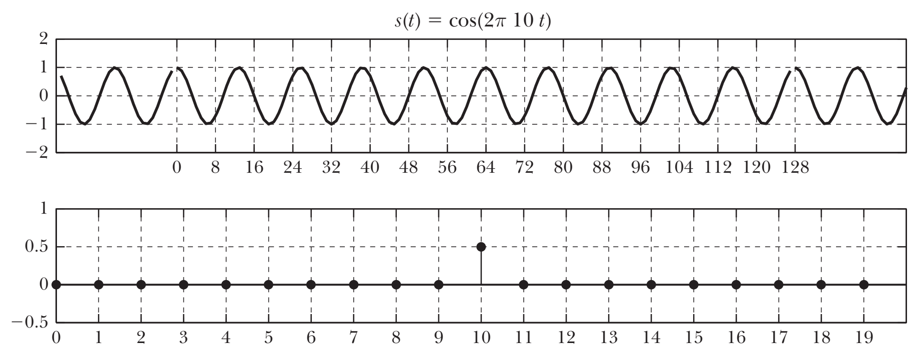



**FIGURE 8.29 128-Point FFT Example—10 Hz cosine** The 128-point discrete signal spans 10 complete cosine periods. The discrete Fourier transform shows a well-defined impulse corresponding to the 10 Hz cosine frequency (only the first 20 points are shown). Such an ideal result occurs only for specific example signals.

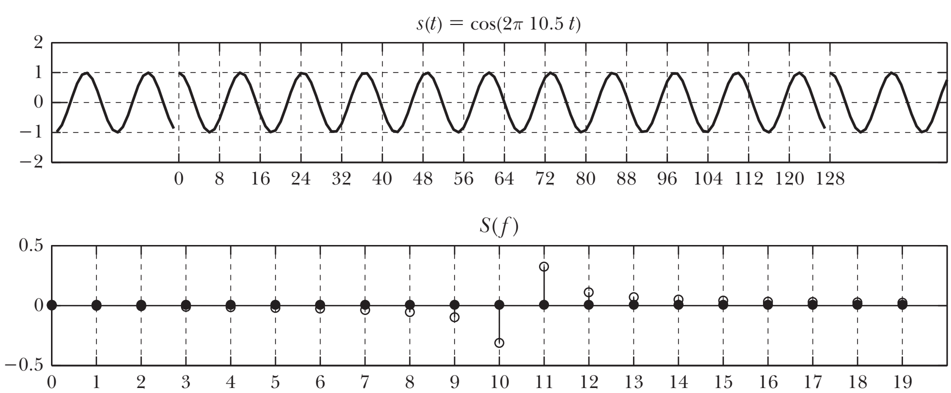



**FIGURE 8.30 128-Point FFT Example—10.5 Hz cosine** Sampling a 10.5 Hz waveform over one second results in a discontinuous periodic waveform (above). Not only is it impossible to exactly represent 10.5 Hz, but the implicit discontinuities introduce new spectral components (below).

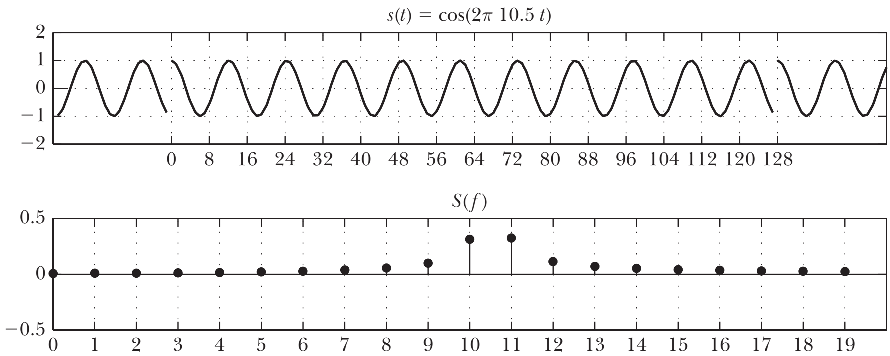



**FIGURE 8.31 128-Point FFT Example—10.5 Hz cosine (magnitude)** Sampling a 10.5 Hz waveform over one second results in a discontinuous periodic waveform (above). In the frequency domain (below) non-zero components appear around the correct frequency, with additional small contributions arising from the discontinuities.

128 samples. Combined, these effects give a Fourier transform that has multiple spectral components in and around 10.5 Hz. In this particular example, the input signal has essentially been multiplied by an ODD square ware having a period of two sample blocks. The output spectrum appears to be very different from what would be expected for a pure 10.5 Hz cosine input.

Fortunately, absolute phase information is rarely as important as the frequency content of a signal. The DFT is typically employed to determine the frequency content of an input signal by examining only the magnitude in the frequency domain. For example, if a mechanical system is vibrating at 10 Hz, its phase (sine or cosine) is irrelevant. Figure 8.31 shows that the magnitude output has significant non-zero components found around 10.5 Hz as expected and additional smaller terms emerging from the discontinuities found in the time-domain representation. The spectrum would be cleaner if these discontinuities could be reduced.

**A DFT Window** The inevitable discontinuities found when sampling real signals are commonly reduced by pre-multiplying the input sample block by a function that attenuates the time-domain components at either end, thus ensuring a smooth transition in every case. This operation is called *windowing*, and various forms of windows have been developed to accomplish this goal. In Figure 8.32, an input signal is multiplied by a *cosine window* of the form window$[t] = 1 - \cos(2\pi t)$ before the DFT is computed.

```
win = 0.5 * (1 - cos(2 * pi * t));
s = s .* win;
S = fft(s)/N;
```

Using this window, the product term goes to zero at the ends of the sample block, minimizing the effects of any discontinuities. Subsequently, the DFT performed on this windowed function no longer includes spurious components introduced by the discontinuity, as shown in the magnitude plot of Figure 8.33. This improvement comes at the expense of resolution and spectral purity. The best spectral resolution would seem to be be achieved for a high sampling rate and a long data block. High sampling rates increase the computational demands of signal processing, while long data blocks may fail to recognize rapid variations in signal behavior.

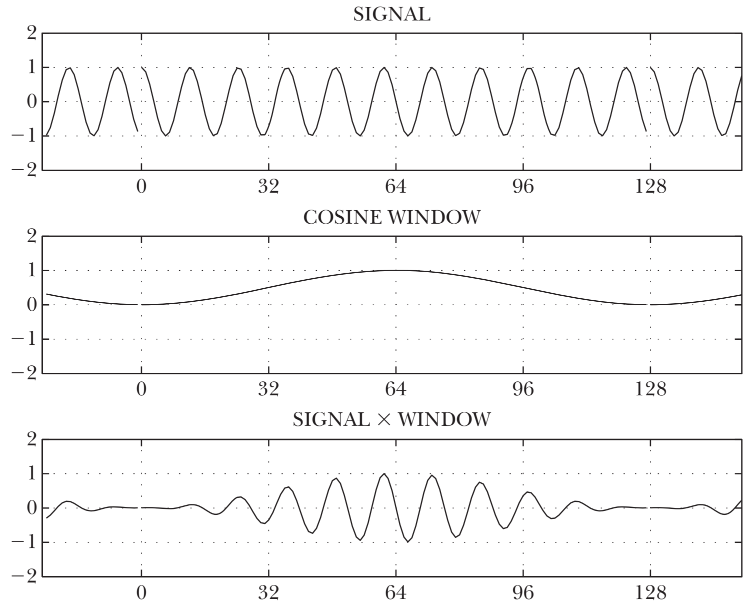



**FIGURE 8.32 128-Point FFT Example—Cosine Window** Samples of a 10.5 Hz waveform over a one-second block are multiplied by a *cosine window* that is zero at each end of the sample block. This operation attenuates the signal at the extremes to limit the impact of discontinuities.

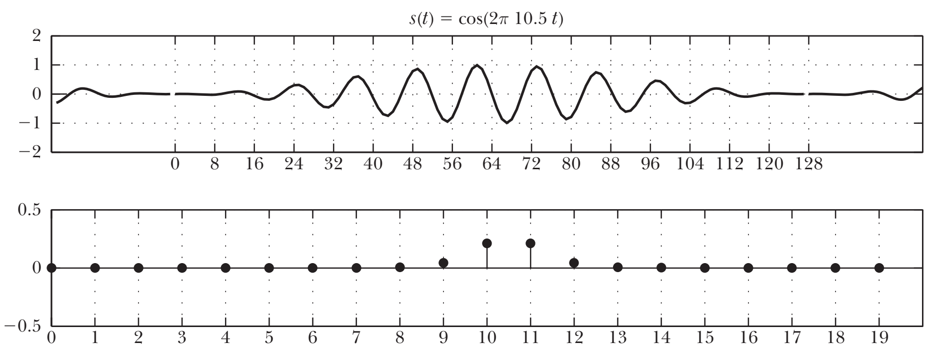



**FIGURE 8.33 128-Point FFT Example—10.5 Hz Cosine (with Cosine Window)** Samples of the windowed 10.5 Hz waveform lead to a magnitude DFT with significant components near 10.5 Hz as expected. The resulting spectrum has fewer spectral artifacts resulting from discontinuities as compared to Figure 8.31.

## EXAMPLE 8.4 (The Effect of Windowing)

*In Figure 8.32 an input sample array was multiplied by function* $\mathrm{s(t)} = \frac{1}{2}(1 - \cos(2\pi \mathrm{t}))$. *What is the corresponding effect on the frequency domain?*

**Solution:**

The frequency-domain signal is convolved with the Fourier transform of $s(t)$. This effect is shown clearly in Figure 8.34 where the components at 3 Hz and 10 Hz both exhibit the characteristic form associated with the $(1 - \cos(t))$ window function. Refer to Appendix A.7.

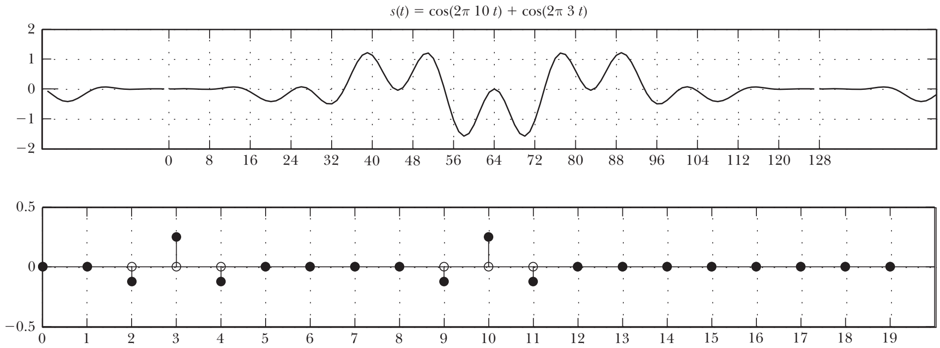



**FIGURE 8.34 128-Point FFT Example—10 Hz Cosine +3 Hz Cosine (with Cosine Window)** This signal $s(t) = \cos(6\pi t) + \cos(20\pi t)$ is multiplied by a cosine window. Consequently, the two components at 3 Hz and 10 Hz appear *convolved* with the Fourier transform of the window function.

**Windowing in MATLAB**—The windowing technique described above is a fundamental operation whenever spectral analysis is performed using the FFT. In particular, the 128-point cosine window can be generated directly using the function `hann(128)`. A variety of other (mostly similar) window shapes are also predefined in MATLAB.

# 8.7 Discrete Time Filtering with MATLAB

To conclude the discussion of discrete signals using MATLAB, a lowpass filter is to be implemented as the discrete convolution with a rectangle. The results will be compared in the time and frequency domains using the `fft()` command. Let the time-domain filter be a unit height rectangle 8 ms wide, and let the test signal be a 20 Hz cosine with unit amplitude. The signals will be defined in MATLAB on a vector with 2048 points assuming a sample rate of 1000 samples per second.

## 8.7.1 A Discrete Rectangle

Assume that the system uses 1000 samples per second and define a rectangle 8 samples wide.

```
t = 1:2048;                              % define time domain
filter(t)=0;
filter(1:8) = 1;                         % define a rectangle filter
f_filter = fft(filter);                  % frequency domain
figure(1);
stem(1:32,filter(1:32),'fill','b');      % plot the rectangle
figure(2);
plot(t,abs(f_filter));                   % plot filter magnitude
```

Figure 8.35 shows the first 32 points of the rectangle. The total area of the rectangle is 8. The `stem()` command options `'fill','b'` produce filled dots at each sample point. Defined in this way on the first eight points of a vector with 2048 points, the rectangle is not an even function, and its Fourier transform would have both real and imaginary components. The corresponding Fourier transform is found using the `fft()` command and its magnitude is shown in the figure. For this rectangle of width 8 defined on 2048 points, the first zero crossing in the sinc function is found at $1/8 \times 2048 = 256$ as shown. The value of the sinc at the origin equals the area under the rectangle.

## 8.7.2 A Cosine Test Signal

The input and output signals are shown in Figure 8.36. The test input signal is a cosine with frequency 20 Hz and has a period $1/20\text{s} = 50$ ms corresponding to 50 samples; it will not matter that this period does not evenly divide 2048. The following code segment continues from above. The convolution results in a vector 4095 points long; in any case, 128 points from the central section of each signal are chosen for plotting.

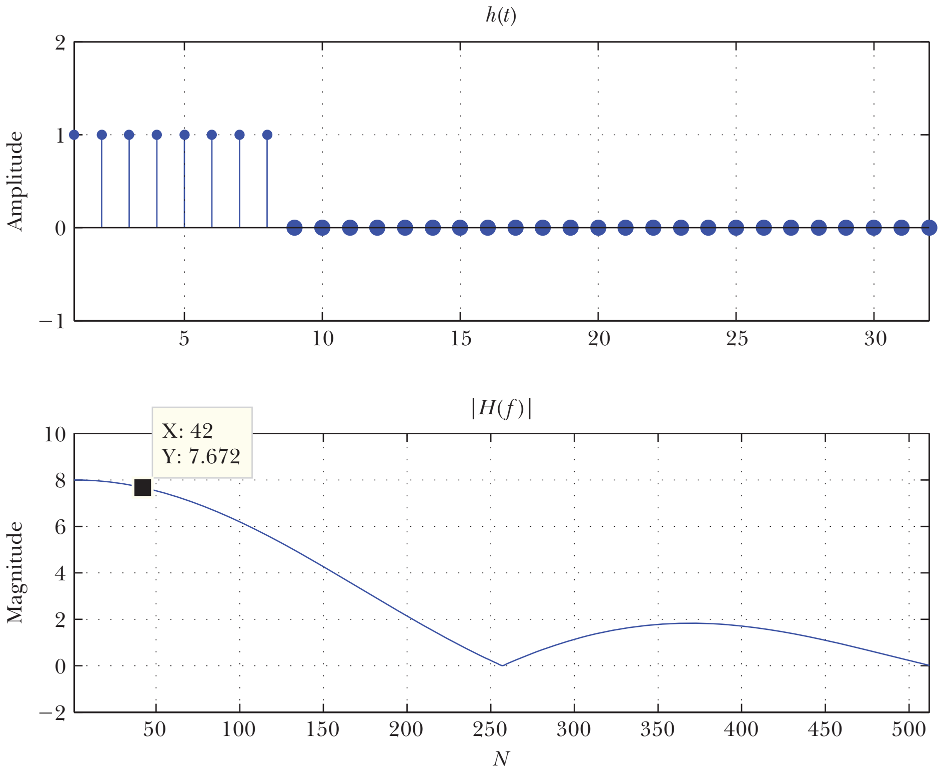



**FIGURE 8.35** A rectangle $h(t)$ is defined to be 8 samples wide on 2048 samples. The FFT yields a sinc function $|H(f)|$ with a first zero crossing at $f = 2048/8 = 256$.

```
input  =  cos(2 * pi * t/50);        % define cosine period 50 samples
output  =  conv(input, filter);   % apply the filter
figure(3);
plot(128 : 256, input(128 : 256));  % plot the input
figure(4);
plot(128 : 256, output(128 : 256)); % plot the input
```

The result may now be checked against the `fft` output. Using the Data Cursor in the MATLAB figure window, observe that the output signal in Figure 8.36 has amplitude 7.657 as shown highlighted. In the frequency domain, the sinc function would multiply impulses located at the cosine frequency, corresponding to 20 ms $\times$ 2048 = 40.96. The closest integer point is $N = 41$, and since the DC component is at $N = 0$, the point of interest lies at $N = 42$ where the amplitude is 7.672 as expected.

## 8.7.3 Check Calculation

The Fourier transform of the rectangle can be written directly by observing that a time-domain rectangle of width 8 ms corresponds to a sinc function with zero crossings spaced every $1/0.008 = 125$ Hz. If the rectangle has area = 8, then

$$S(f) = 8\,\text{sinc}(f/125)$$

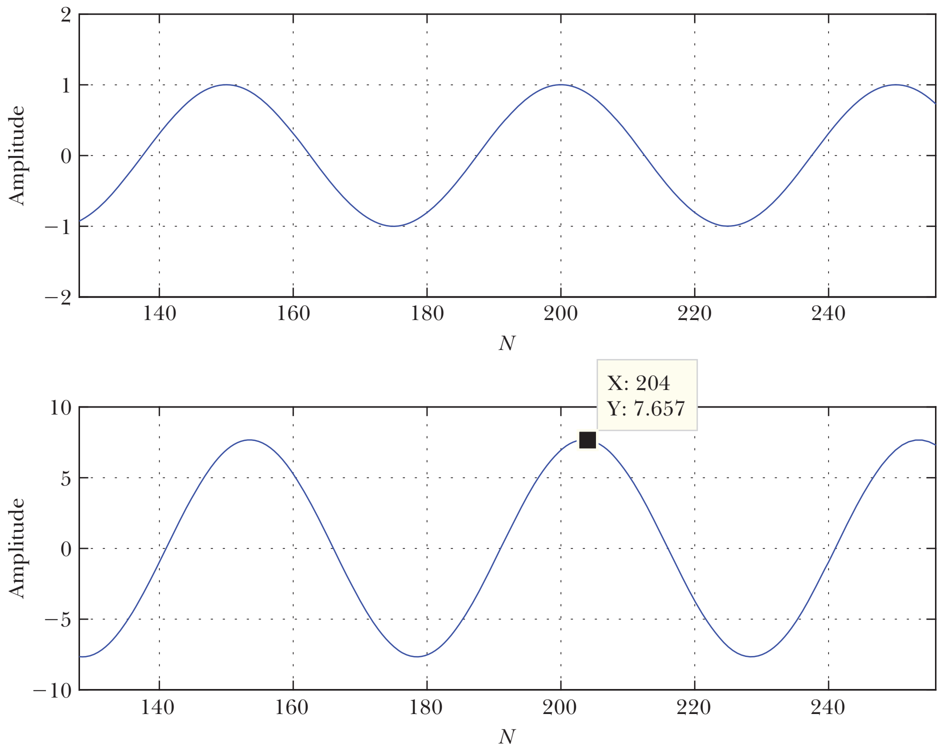



**FIGURE 8.36** A 20 Hz cosine is sampled at 1000 samples per second and convolved with the rectangle in Figure 8.35. The output signal is a cosine with that same period and an amplitude determined by the filter shape in the frequency domain.

The magnitude of this transfer function describes the effect of the system on the amplitude of an input signal component at any frequency. A unit amplitude input signal at $f = 20$ Hz leads to an output signal with an amplitude that is consistent with both answers found above.

```
abs(8 * sinc(20/125))
ans = 7.6673
```

In this example, MATLAB has been used to perform a time-domain convolution producing a lowpass filter, and the result is confirmed in the discrete Fourier transform output. Calculations on discrete signals always require reference to the underlying sample rate, and as shown here the output from the `fft()` command corresponds directly to the expected results within the limitations imposed by the discrete operations.

## 8.8 Conclusions

Discrete signals come about naturally whenever continuous signals are to be processed in a digital computer, since sample values must be converted to discrete vector components for storage. If a suitable sampling rate is chosen for a bandlimited signal, a discrete signal contains all the information necessary to exactly reproduce the original continuous signal. Filtering techniques can be implemented on discrete samples through simple delays and additions. The Fourier transform performed on discrete signals is known as the discrete Fourier transform (DFT), and yields a discrete frequency-domain signal. The DFT algorithm can be optimized, where it is then known as the fast Fourier transform (FFT).

# End-of-Chapter Exercises

**8.1** Consider the sampling problem of photographing a clock tower, as described on Page 315. What is the highest-frequency component in the clock hands? What if photos were taken once every 65 minutes? Once every 120 minutes? What minimum sampling rate should be chosen for the photographs?

**8.2** Consider a signal:

$$s(t) = \cos(200\pi t) - 10\sin(1000\pi t) + 2\cos(1500\pi t) - 3\sin(2000\pi t)$$

What is the minimum sampling rate that should be used with this signal?

**8.3** A cosine $g(t) = \cos(16\pi t)$ is sampled at 10 Hz. The samples are then sent to you for analysis. You begin by putting them through an ideal lowpass filter with 5 Hz cutoff frequency. What do you observe in the time domain and in the frequency domain? *Hint: Make a sketch.*

**8.4** Use MATLAB to plot the signal $s(t) = \cos(22\pi t)$ using a time variable $t$ defined to have 100 elements spanning the interval [0, 10] seconds. Show how your result compares to Figure 8.11 and explain both results in terms of the sampling theorem.

**8.5** A sinusoidal signal has a period of $N$-samples at a sample rate of 8000 samples per second. What is the signal frequency (Hz)?

**8.6** A 2400 Hz cosine is sampled at 20,000 samples per second. What is the period in samples?

**8.7** A 1024 point FFT has a frequency component at $N = 125$. Assuming a sample rate of 8000 samples per second, what is the frequency (Hz)?

**8.8** A 128 Hz cosine is sampled at 1024 samples per second and stored in a vector with 2048 elements. When the FFT is computed, what elements (n) of the output vector contain the cosine components?

**8.9** The DFT of a sampled signal is computed. In what element (n) of the output vector is the DC component found?

**8.10** The sampled signal in Figure 8.37 is a sinc function shifted to lie at $n = 400$. The zero crossings closest to this location are at $n = 368$ and $n = 432$.

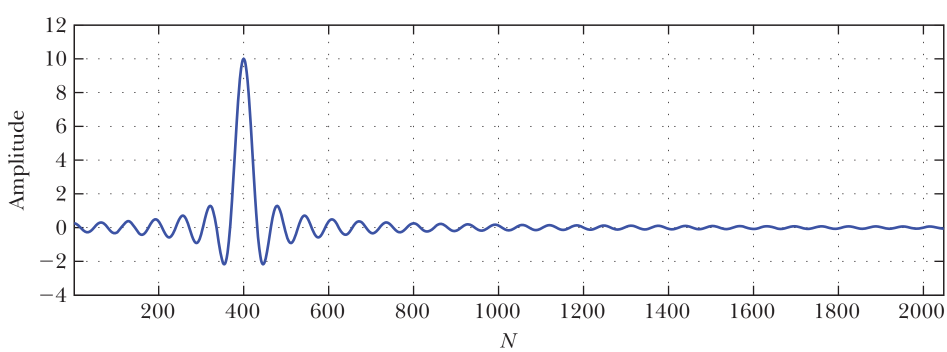



**FIGURE 8.37** Figure for Questions 8.10 to 8.12.

**(a)** Give a MATLAB expression for this signal in terms of `sinc()`.
**(b)** Sketch the expected magnitude of the FFT of this signal.

**8.11** An impulse response $h(t)$ is sampled at 1000 samples per second to give the signal in Figure 8.37, which is described as a sinc function shifted to lie at $n = 400$. The zero crossings closest to this location are at $n = 368$ and $n = 432$.

**(a)** Give an expression for $h(t)$.
**(b)** Find $H(f)$ and $|H(f)|$.
**(c)** If a signal is convolved with $h(t)$, describe in detail the resulting filter operation.
**(d)** Give a MATLAB expression for this signal in terms of `sinc()`.
**(e)** Compute the Fourier transform of this sampled signal and compare to the $H(f)$ computed above.

**8.12** Repeat Question 8.11 with a sampling rate of 44,000 samples per second.

**8.13** A signal is stored using MATLAB in a vector of 1024 points, and the FFT is computed to give the output vector `output`. Subsequently, the command `x = fftshift(output)` is executed. Write an equivalent expression for the `fftshift()` command in this example.

**8.14** A signal is stored using MATLAB in a vector of 2048 points at a sample rate of 8000 samples per second and the FFT is computed.

**(a)** What is the maximum frequency component (Hz) in the transform?
**(b)** What is the spacing (Hz) between frequency components?

**8.15** A signal is stored using MATLAB in a vector of 512 points at a sample rate of 20000 samples per second.

**(a)** What is the maximum frequency component (Hz) in the transform?
**(b)** What is the spacing (Hz) between frequency components?

**8.16** Consider the sampling system block diagram of Figure 8.38.

**(a)** Under ideal conditions, how are the signals at **A** and **C** related?
**(b)** Describe the purpose of the lowpass filter $H(f)$.
**(c)** Give an expression for the value of $f_m$ as a function of $f_s$.
**(d)** What important block is absent in this system design?

**8.17** Consider the sampling system block diagram of Figure 8.38.

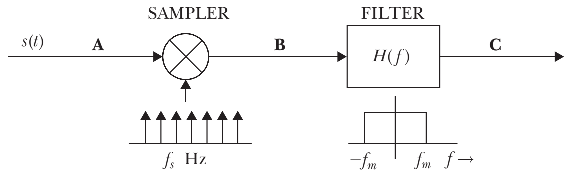



**FIGURE 8.38** Figure for Questions 8.16 to 8.18.

**(a)** Redraw the diagram and include an anti-alias filter block.
**(b)** What type of filter will be used?
**(c)** Where is the filter located?
**(d)** What cutoff frequency will be used?

**8.18** The sampling system of Figure 8.38 is set up to handle input frequencies up to 150 Hz (choosing $f_s = 300$ Hz and $f_m = 150$ Hz) and is tested at the maximum expected input frequency using the signal $s(t) = \cos(200\pi t)$. The designers did not read this chapter and saw no need for an anti-alias filter.

Once installed and operational, an unexpected noise source $n(t) = 0.5\cos(400\pi t)$ gets added to the input signal and the system samples the sum $w(t) = s(t) + n(t)$.

**(a)** Give an expression for the Fourier transform of the signal at point **B** in the system for the input $w(t)$ and accurately sketch this spectrum. Identify the aliasing components.
**(b)** Give an expression for the Fourier transform of the signal at point **C** in the system for the input $w(t)$ and accurately sketch this spectrum.
**(c)** Redraw the above system with an appropriate anti-alias filter.

**8.19** Consider the signal $s(t)$ that is a 1000 Hz square wave.

**(a)** What is the maximum frequency component in $s(t)$? What sampling rate should be chosen for this signal?
**(b)** This signal $s(t)$ is sampled by a well-designed sampling system with a sampling rate of 8000 Hz and an anti-alias cutoff frequency of 4 kHz. Accurately sketch the recovered signal in the time and frequency domains.
**(c)** What sampling rate and anti-alias filter should be chosen so that at least 90 percent of the power in $s(t)$ is retained?

**8.20** A system consists of two mixers in series as shown in Figure 8.39. The first mixer multiplies an input signal by a cosine at frequency $f_1 = 2000$

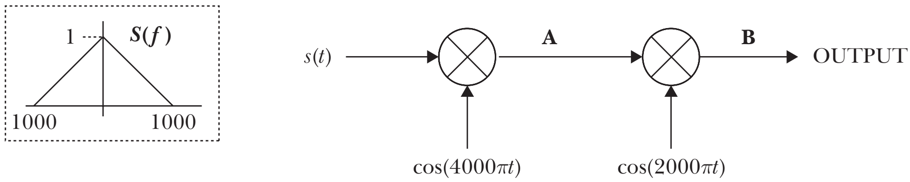



**FIGURE 8.39** Figure for Question 8.20.

Hz. The second mixer multiplies an input signal by a cosine at frequency $f_2 = 1000$ Hz.

**(a)** Use a sketch and describe this system in the frequency domain.
**(b)** An input signal with spectrum $S(f)$ as shown enters the system.
  **i.** Sketch the spectrum of the output signal.
  **ii.** What is the maximum frequency present in the output signal?
  **iii.** What minimum sampling rate should be chosen for the output signal?

**8.21** A system is designed to show the frequency components in an input signal using an FFT calculation on consecutive blocks of 2048 data points. The analog-to-digital converter sampling rate is 20 kHz.

**(a)** Describe the anti-alias filter that should be used in this system.
**(b)** What is the time span in seconds of a block of 2048 samples?
**(c)** What is the spacing (Hz) of the 2048 frequency-domain components?
**(d)** What is the total range (Hz) of frequency components (positive and negative)?
**(e)** What specific frequency-domain values are the closest to 1000 Hz?

**8.22** Suppose that in Figure 8.28 a sampling rate of 20,000 Hz was used to collect the time-domain signal. What are the frequencies associated with the frequency components shown?

**8.23** The following `fft` operation is performed in MATLAB. Assume that the output vector is a time-domain signal $s(t)$ and provide an interpretation of the discrete input and output signal vectors.

**(a)** Sketch the frequency-domain signal $S(f)$ around the origin.
**(b)** Give an expression for the continuous $s(t)$ as a sum of sinusoids.
**(c)** Sketch the continuous time-domain signal $s(t)$ corresponding to the output samples in $s$.

```
s = fft([1 1 0 1 0 1 0 1 ])
s = 5 1 1 1 −3 1 1 1
```

**8.24** Use MATLAB to find the Fourier transform of the discrete signal $H(f)$ defined on 2048 points.

**(a)** Define $H(f)$ to be an even rectangle $H(f)$ that extends 256 points on either side of zero.
**(b)** Plot the corresponding time-domain impulse response $h(t)$ in the interval [1 : 128].
**(c)** Confirm that $h(t)$ has no imaginary components.

**8.25** Use MATLAB to model a time-domain filter $h(t)$ that is a rectangle 16 msec wide, using 2048 points and a sampling rate of 1000 samples per second.

**(a)** Show the magnitude of the transfer function $|H(f)|$.
**(b)** Show the output for an input signal that is a 20 Hz cosine.

**8.26** Use MATLAB to model a time-domain filter $h(t)$ that is a rectangle 32 ms wide, using 2048 points and a sampling rate of 1000 samples per second.

**(a)** Show the magnitude of the transfer function $|H(f)|$.
**(b)** Show the output for an input signal that is a 20 Hz cosine.

CHAPTER 9

# The *z*-Transform

**LEARNING OBJECTIVES**

By the end of this chapter, the reader will be able to:

- Describe the difference between Fourier, Laplace, and *z*-transforms
- List the properties of the *z*-transform
- Identify the Fourier transform in the *z*-transform
- Explain the relationship between *z*-transform pairs
- Identify overall system behavior from a pole-zero diagram
- Apply the *z*-transform to analyze sampled systems
- Use MATLAB to analyze a system defined on *z*-transform parameters
- Apply the *z*-transform to frequency-domain filtering
- Describe the difference between FIR and IIR filters
- Use MATLAB to analyze the performance of FIR and IIR filters
- Design simple filters with specific properties using the *z*-transform

## 9.1 Introduction

In Chapter 8 the use of discrete time signals was described as an important consequence of using digital computers in signal processing. The analysis of sampled signals revealed a number of special properties of the Fourier transform with discrete time signals (DTFT). From these observations, it emerges that the application of *digital signal processing* (DSP) may benefit from a different transform that simplifies both the mathematics and the computational requirements by recognizing the special nature of discrete signals. The *z*-transform incorporates aspects of both the Fourier transform and the Laplace transform when dealing with discrete signals.

## 9.2 The *z*-Transform

When dealing with discrete signals, the *z*-transform is generally preferred over the more general Fourier or Laplace transform techniques owing to the unique properties of discrete signals in both the time and frequency domains. The *z*-transform can be regarded as a generalization of the discrete time Fourier transform (DTFT) much as the Laplace transform is a generalization of the continuous Fourier transform. Consequently, the *z*-transform can also be described as a discrete time version of the Laplace transform. Like the Laplace transform, the *z*-transform incorporates the Fourier transform and adds another dimension to create a two-dimensional domain to be called the *z*-domain. The comparison to the Laplace transform is strengthened when dealing with *z*-transform poles and zeros, system stability, and regions of convergence.

## 9.2.1 Fourier Transform, Laplace Transform, and z-transform

As the *z*-transform shares characteristics of both the Fourier transform and the Laplace transform, it is useful to introduce this new transform by direct comparison with familiar results. Figure 9.1 compares and contrasts the nature of the three transforms and specifically identifies the Fourier transform of a discrete signal in each of the other transforms.

Consider the Fourier transform of the sampled signal $s(nT)$ shown in Figure 9.1. This resulting discrete time Fourier transform (DTFT) leads to a periodic frequency-domain signal where the period $F_s = 1/T$ is the sample rate of the time-domain signal. It can be assumed that the sample rate $F_s$ was well chosen and that the input signal has no frequency components greater than or equal to one half $F_s$; in that case, the periodic copies do not overlap. In any case, it should be emphasized that only one period of the frequency-domain signal is necessary to completely describe its appearance for all frequencies; this period could be defined near the origin in the interval $(-F_s/2, +F_s/2)$ Hz as indicated by dashed lines.

In Chapter 7, it was shown that Fourier transform components $|S(f)|$ as a function of $\omega$ can be found along the vertical $(\sigma = 0)$ axis in the *s*-domain of the Laplace transform. In the Laplace transform of Figure 9.1 the DTFT spectrum $|S(f)|$ can be readily identified. The line $(\sigma = 0)$ is of particular importance as it divides the *s*-plane into *lefthand* and *righthand* parts in which system stability depends on restricting the presence of Laplace transform poles to the lefthand plane. Calculation of the Laplace transform necessarily includes a *region of convergence* on the *s*-plane, which for causal systems lies to the right of a vertical line delimiting a region that does not include any poles.

Finally, the *z*-transform embodies both the periodic nature of the DTFT spectrum and the pole-zero approach to system analysis possible with the Laplace transform. Essentially, the *z*-transform maps the Laplace $(j\omega, \sigma)$ axes onto a new surface where polar coordinates as $(r, \omega)$ lead to circles centered on the origin of the *z*-plane, in which $\omega$ describes an angle in the interval $(0, 2\pi)$ around the circumference of a circle and $\sigma$ defines its radius $(r)$. In this model, the $(\sigma = 0)$ line of the Laplace transform is specifically described by a unit circle $(r = 1)$ in the *z*-transform. It follows that points inside and outside this circle correspond to the *lefthand* and *righthand* *s*-plane, respectively. As shown in Figure 9.1, the unit circle is well suited to the periodic nature of the DTFT spectrum, as a single copy of the DTFT output can be found by tracing around its circumference, multiple times if necessary, to identify components at frequencies beyond $\pm F_s$ Hz. Like the Laplace transform, a *region of convergence* will be associated with the *z*-plane, which for causal systems will now lie outside a circle delimiting a region that does not include any poles.

This new perspective on familiar transforms is especially useful as the properties of the *z*-transform will resemble those of the transforms already studied. As with the Laplace transform, the inverse *z*-transform will be accomplished as much as possible by using tables of the *z*-transforms of common signals.

In this chapter, the discrete signal $x(nT)$ sampled at intervals of $T$ is written $x[n]$ by assuming a sample spacing $T = 1$. A general discrete signal $x[n]$ may be related to specific signal parameters by substituting $T = 1/f_s$ to represent actual samples $x(nT)$ taken at a rate of $f_s$ Hz.

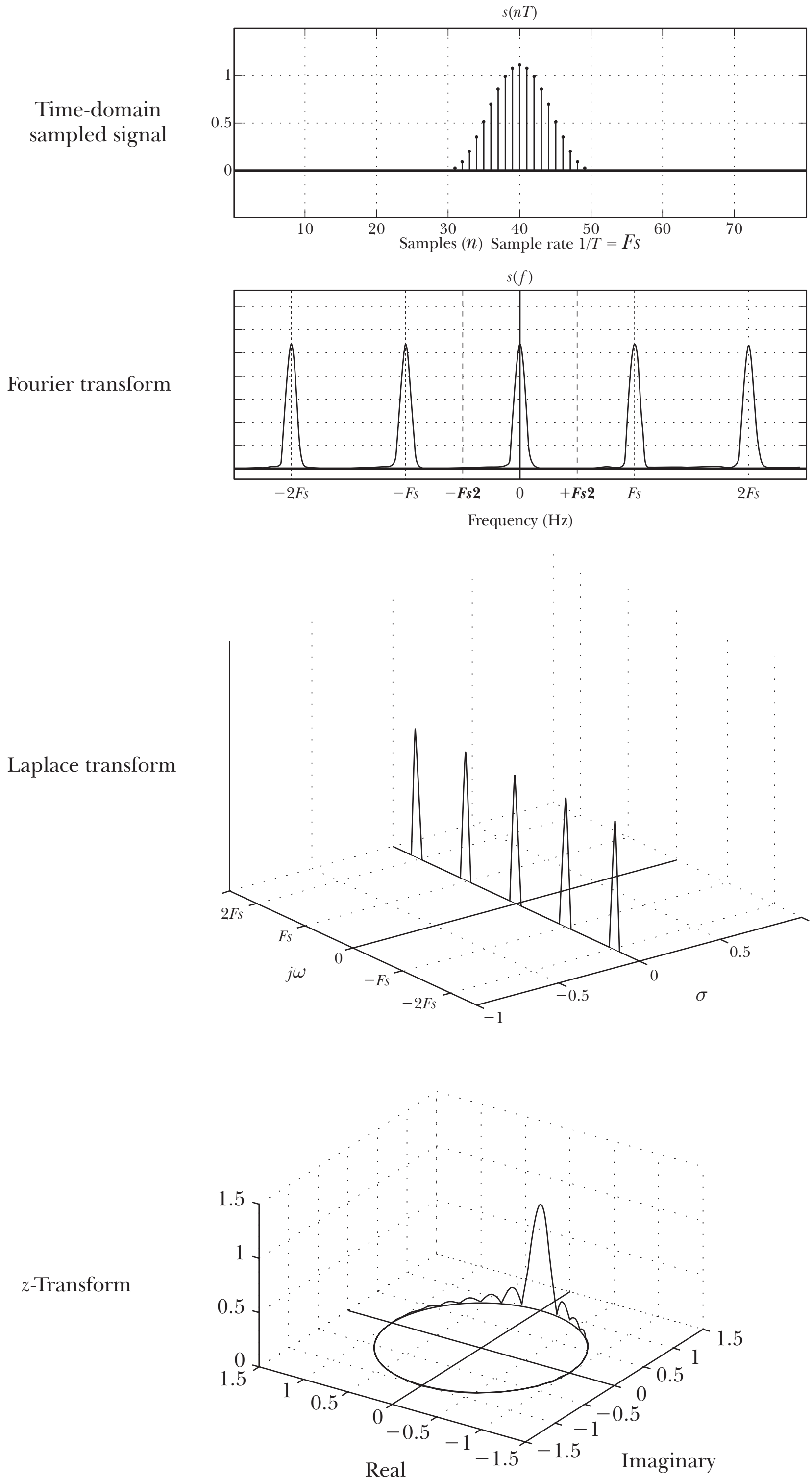



**FIGURE 9.1 Three Transforms** The (periodic) Fourier transform $S(f)$ of a discrete signal $s(nT)$ can be found in the Laplace transform along the $(\sigma = 0)$ line, while the same result can be found in the *z*-transform as a single period around the unit circle.

## 9.2.2 Definition of the z-Transform

The z-transform of the discrete signal $x[n]$ is defined as:

**DEFINITION 9.1 Two-Sided z-Transform**

*If* x[n] *is a discrete signal and* z *is a complex value, then*

$$X(z) = \sum_{n=-\infty}^{+\infty} x[n]z^{-n} \tag{9.1}$$

*is called the* two-sided *or* bilateral z-transform.

In general, the function $X(z)$ is a continuous function of $z$. The use of the z-transform with *causal* linear time invariant systems often involves signals of the form $x[n]u[n]$, in which case the z-transform integral becomes a one-sided form starting at $n = 0$ as:

**DEFINITION 9.2 One-Sided z-Transform**

*If* x[n] *is a discrete signal and* z *is a complex value, then*

$$X(z) = \sum_{n=0}^{+\infty} x[n]z^{-n} \tag{9.2}$$

*is called the* one-sided *or* unilateral z-transform.

and this definition is to be assumed in this chapter unless otherwise stated. The relationship between a discrete time-domain signal $x[n]$ and its z-transform $X(z)$ will be expressed as:

$$x[n] \overset{z}{\longleftrightarrow} X(z) \tag{9.3}$$

where the script $\mathcal{Z}$ distinguishes the z-transform.[1] Where there is no chance of confusion, the simple notation $x[n] \longleftrightarrow X(z)$ can be used. Note that this is not an equality; the double arrow indicates the unique relationship of a z-transform pair. The form

$$\mathcal{Z}\,[x[n]] = X(z)$$

is also seen.

## 9.2.3 The z-Plane and the Fourier Transform

Any complex value $z$ can be expressed as real and imaginary components of the form $z = x + jy$ and the ordered pair $(x, y)$ plotted on a Argand plane using real and imaginary axes. The same value $z$ can be expressed in polar form as $z = r\,e^{j\omega}$ where $x = r\cos(\omega)$ and $y = r\sin(\omega)$ as shown in Figure 9.2.

[1] The letters $\mathcal{F}$, $\mathcal{L}$, and $\mathcal{Z}$, will specify the Fourier, Laplace, and z-transforms, respectively.

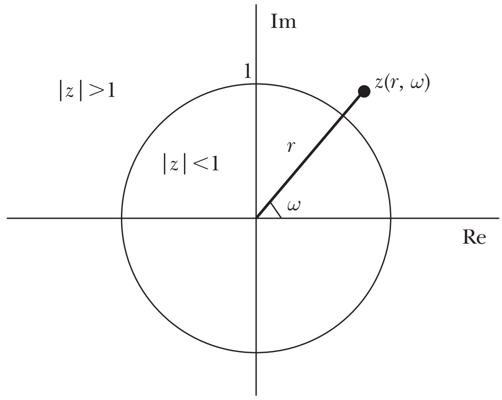



**FIGURE 9.2 The z-Plane** A complex value $z$ may be expressed in polar coordinates as $(r, \omega)$ and plotted on the Argand plane with real and imaginary axes. This is the z-plane defined for the z-transform. For the special case of $r = 1$, values of $z$ trace the unit circle shown here for reference.

For fixed $r$, values of $z$ given by $(r, \omega)$ describe circles of radius $r$ on the z-plane. For the special case of $r = 1$, the $z$ values define a unit circle as shown in Figure 9.2. In this case, the z-transform from Eqn. 9.1 becomes:

$$X(z) = \sum_{n=-\infty}^{+\infty} x[n]e^{-j\omega n} \tag{9.4}$$

which may immediately be recognized as the discrete time Fourier transform of $x[n]$. Consequently, the unit circle plays a central role in any study of the z-transform, as was noted in Figure 9.1.

## 9.3 Calculating the z-Transform

For simple signals of limited duration, the z-transform can be computed by inspection since the summation is easily accomplished. In Figure 9.3, four discrete signals are shown with their z-transforms. As with the Laplace transform it is important to associate a region of convergence with z-transforms. In this figure, the signal $a[n]$ is an impulse at the origin; the z-transform is given by a summation with only one non-zero term:

$$A(z) = \sum_{n=0}^{7} a[n]z^{-n} = 2\, z^0 = 2$$

This result is consistent with the Fourier transform of an impulse $2\delta(t) \leftrightarrow 2$, and the constant value result is independent of $z$ so the region of convergence is the entire z-plane. The signal $b[n]$ is a single impulse at $n = 3$ or a shifted version of the impulse $a[n-3]$, giving:

$$B(z) = \sum_{n=0}^{7} b[n]\, z^{-n} = 2\, z^{-3} = z^{-3}\, A(z)$$

This result is consistent with the Fourier transform of a shifted impulse, where $2\delta(t-3) \leftrightarrow 2e^{-j6\pi ft}$ and $B(z)$ is identical to the Fourier transform $B(f)$ when $z = e^{2\pi ft}$

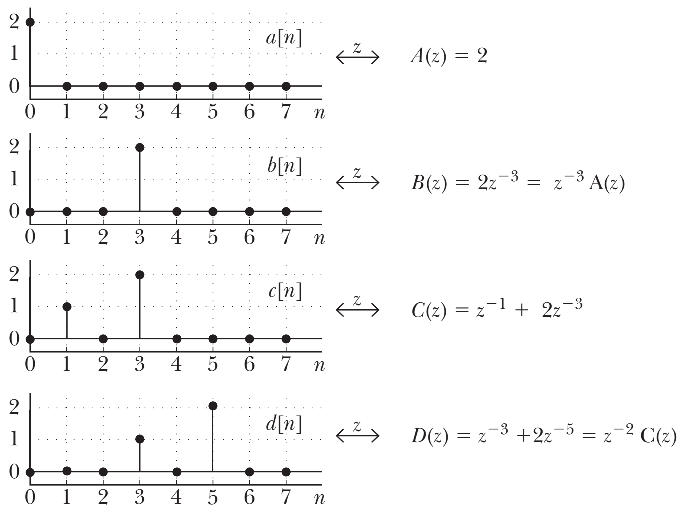



**FIGURE 9.3** ***z*-Transforms** Four discrete signals are shown along with their *z*-transforms. The signal $a[n]$ is an impulse leading to a constant value. The signal $b[n]$ is a single value at $n = 3$, while signal $c[n]$ consists of two values. This signal $d[n] = c[n-2]$ is a shifted version of $c[n]$ showing the shifting property of the *z*-transform.

(defining the unit circle). The term in $z^{-3}$ would not exist for $z = 0$, so the region of convergence is the entire *z*-plane for $z \neq 0$. The signal $c[n]$ has two non-zero components and the summation involves two terms, as:

$$C(z) = \sum_{n=0}^{7} c[n]z^{-n} = 1z^{-1} + 2z^{-3}$$

The region of convergence is the entire *z*-plane for $z \neq 0$. The signal $d[n]$ is the shifted signal $c[n-2]$:

$$D(z) = \sum_{n=0}^{7} d[n]z^{-n} = 1\,z^{-3} + 2z^{-5} = z^{-2}C(z)$$

In general, for a signal shifted to the right by $N$ samples $x[n-N]$, the *z*-transform is given by:

**THEOREM 9.1** *(Shifting)*

*If*

$$x[n] \xleftrightarrow{z} X(z)$$

*then*

$$x[n-N] \xleftrightarrow{z} z^{-N}X(z)$$

This result is a fundamental and important theorem of the *z*-transform. It means that the term $z^{-1}$ corresponds to a *delay* by 1 sample. If a discrete signal is delayed by $N$ samples, its *z*-transform is multiplied by $z^{-N}$. This result can also be confirmed in signal $b[n]$ above, which is a right-shifted version of $a[n]$.

Another interesting property of the $z$-transform can be shown by considering the product $x[n] = a^n\, s[n]$ where $a$ is a constant. The powers of $a$ represent an exponentially *decreasing* series only if $a < 1$. In any case, the general rule for the $z$-transform of the product $a^n\, s[n]$ can be determined by inspection using one of the sample signals from Figure 9.3; consider the signal $c[n]$ with $z$-transform $C(z) = z^{-1} + 2z^{-3}$ and find the $z$ transform of $a^n c[n]$:

$$\sum_{n=0}^{7} a^n\, c[n] z^{-n} = a^1 z^{-1} + 2a^3\, z^{-3} = C\left(\frac{z}{a}\right)$$

where it is clear that each term in $z^{-n}$ becomes a term in $a^n\, z^{-n} = \left[\frac{z}{a}\right]^{-n}$ and this would happen for any number of terms $n$. In other words:

**THEOREM 9.2** *(Scaling)*

*If*

$$x[n] \overset{z}{\longleftrightarrow} X(z)$$

*then*

$$a^n x[n] \overset{z}{\longleftrightarrow} X\left(\frac{z}{a}\right)$$

which may be called the *scaling* property as the $z$-transform argument is scaled by the constant value $a$.

In general, calculating the $z$-transform involves computing a (much longer) summation while ensuring that the result converges to give a finite sum. In many cases, given the definition of a signal $s[n]$, a closed form solution of $S(z)$ is desirable. For arbitrary signals, the mathematics of *infinite series* will prove to be useful. For example, the well-known geometric series:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad \text{for} \quad |x| < 1 \tag{9.5}$$

is of special interest for $z$-transforms, which will be evident as the following useful $z$-transforms are evaluated.

## 9.3.1 Unit Step $u[n]$

Consider the $z$-transform of the step function $u[n]$ given by:

$$U(z) = \sum_{n=-\infty}^{\infty} u[n] \quad z^{-n} = \sum_{n=0}^{\infty} z^{-n}$$

or the infinite series

$$U(z) = 1 + \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \frac{1}{z^4} + \ \ldots \ + \frac{1}{z^n}$$

which by inspection converges only for $|z| > 1$. This defines the $z$-plane region of convergence to be the area outside a circle of radius $r = 1$. If the summation is rewritten with positive exponents, then Eqn. 9.5 may be applied directly to give:

$$U(z) = \sum_{n=0}^{\infty} \left(\frac{1}{z}\right)^n = \frac{1}{1-\frac{1}{z}} = \frac{z}{z-1} \tag{9.6}$$

The result is summarized as:

$$u[n] \xleftrightarrow{z} \frac{z}{z-1} \quad \text{ROC} = \quad |z| > 1 \tag{9.7}$$

In this case a closed form solution was possible. In general, solutions to the *z*-transform may be written as $A(z)/B(z)$ such that (as with the Laplace transform) the transform result is *zero* when the numerator is zero, and there will be a *pole* on the *z*-plane whenever the denominator is zero.

From Eqn. 9.7, $U(z)$ will have a zero at the origin $(z = 0)$ and a pole at $(z = 1)$. Figure 9.4B shows this result on a *z*-plane pole-zero diagram where the region of convergence is shown shaded. Figure 9.4A shows the pole location as a Fourier transform impulse component on the *z*-plane. The Fourier transform frequency domain is defined around the unit circle $(r = 1)$, and the pole position $(r, \omega) = (1, 0)$ corresponds to its origin $(\omega = 0)$. The presence of a single impulse at the origin is consistent with the Fourier transform of a constant. In this case, the unit step $u[n]$ is a constant discrete signal for $n \geq 0$ as shown in Figure 9.4C, and the corresponding Fourier transform is a periodic series of impulses spaced along the frequency axis at

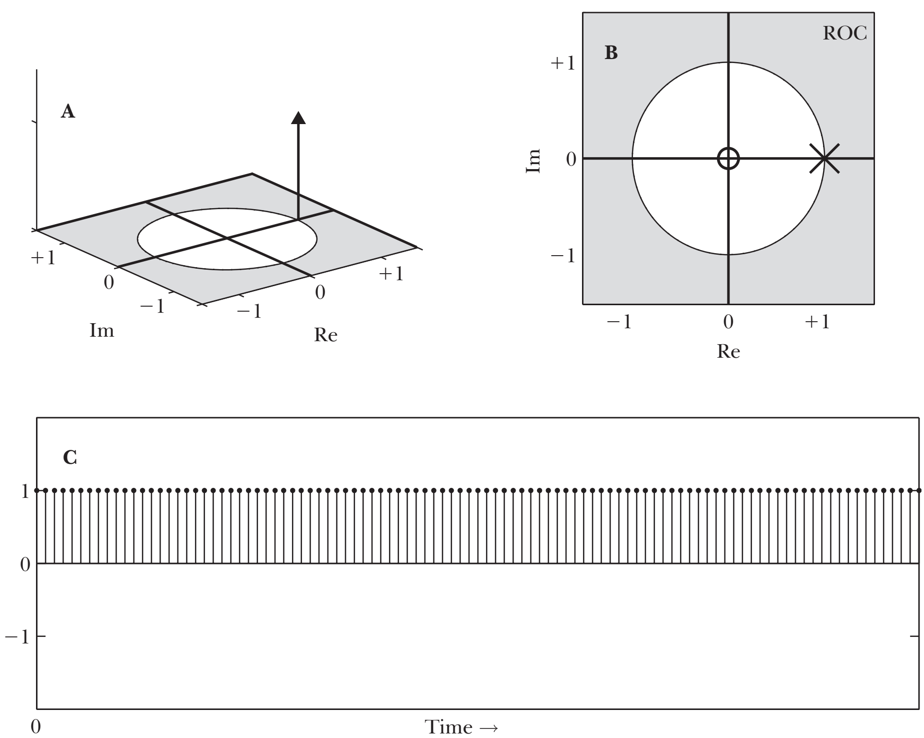



**FIGURE 9.4 The *z*-Plane—unit step $u[n]$** The *z*-transform of the discrete unit step $u[n]$ has a single pole at $z = 1$ and a zero at $z = 0$, as seen in **B**. This pole lies on the unit circle. **A** is a three-dimensional view of the familiar Fourier transform component of a single impulse at the origin. **C** shows the corresponding time-domain waveform, which is a (sampled) constant one for $n > 0$.

the sample rate. As $\omega$ increases in the $z$-domain, the pole is repeatedly encountered to match the expected DTFT result.

### 9.3.2 Exponential $a^n\, u[n]$

Consider the $z$-transform of the exponential function $x[n] = a^n u[n]$ for real-valued $a$ given by:

$$U(z) = \sum_{n=0}^{\infty} [a^n\, u[n]]\ z^{-n} = \sum_{n=0}^{\infty} a^n z^{-n}$$

or the infinite series

$$U(z) = 1 + \frac{a}{z} + \frac{a^2}{z^2} + \frac{a^3}{z^3} + \frac{a^4}{z^4} + \ \ldots\ + \frac{a^n}{z^n}$$

which converges only when the fractions are each less than 1 or when $|z| > |a|$. This defines the $z$-plane region of convergence to be the area outside a circle of radius $r = |a|$. Observe that for negative values of $a/z$ the terms in the summation would alternate sign as odd or even powers of $n$ are encountered. If the summation is rewritten with positive exponents, then Eqn. 9.5 may be applied directly to give:

$$U(z) = \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n = \frac{1}{1 - \frac{a}{z}} = \frac{z}{z - a} \tag{9.8}$$

The result is summarized as:

$$a^n u[n] \overset{z}{\longleftrightarrow} \frac{z}{z - a} \qquad \text{ROC} = \quad |z| > |a| \tag{9.9}$$

From Eqn. 9.7, this $z$-transform will have a zero at the origin ($z = 0$) and a pole at ($z = a$). Figure 9.5B shows this result on a $z$-plane pole-zero diagram where the region of convergence is shown shaded. For $|a| < 1$ as shown, the pole lies within the unit circle. Figure 9.5A shows the same pole location as a Fourier transform impulse component on the $z$-plane. The pole position $(r, \omega) = (0.4, 0)$ corresponds to a single impulse located inside the unit circle; this pole would be found on the lefthand plane in the Laplace transform. The corresponding discrete time-domain decreasing exponential signal for $n \geq 0$ is shown in Figure 9.4C.

**EXAMPLE 9.1 (Scaling Theorem)**

*Use the z-transform of* u[n] *from Eqn. 9.7 and the scaling property of Eqn. 9.2 to obtain the z-transform of* $a^n u[n]$.

**Solution:**
Since

$$u[n] \overset{z}{\longleftrightarrow} \frac{z}{z - 1}$$

from Eqn. 9.7, then by the scaling property:

$$a^n\ u[n] \overset{z}{\longleftrightarrow} U\left(\frac{z}{a}\right) = \frac{\frac{z}{a}}{\frac{z}{a} - 1} = \frac{z}{z - a}$$

as found in Eqn. 9.9. The result is confirmed.

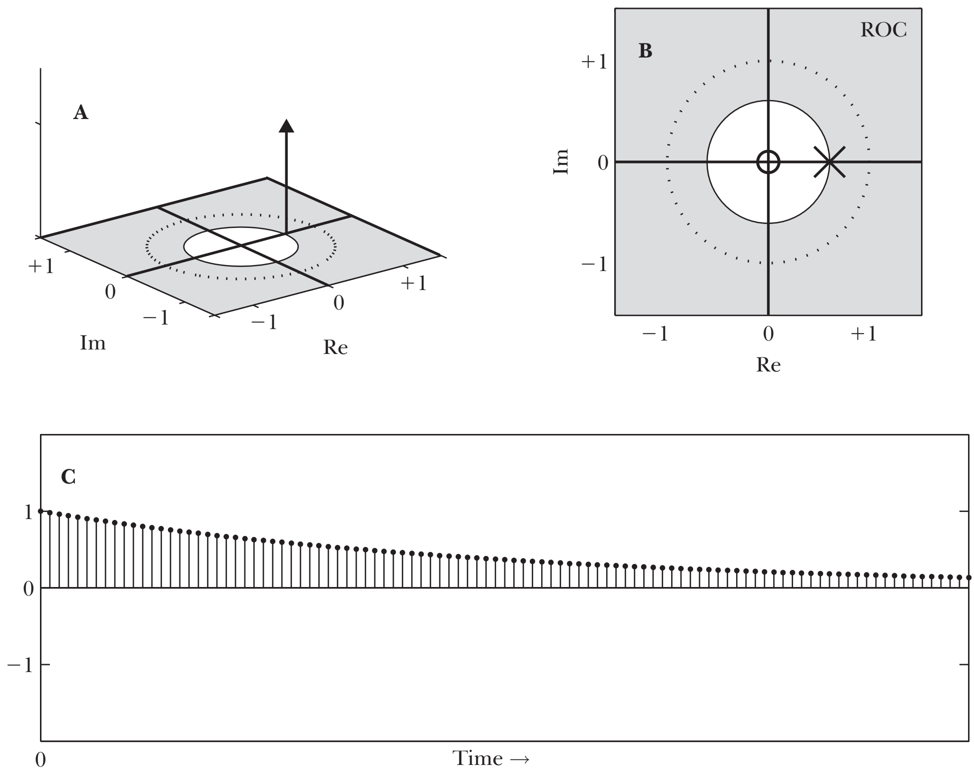



**FIGURE 9.5** ***z*-Plane Poles—($\omega = 0$)** **A** is a three-dimensional view of a single pole located at $(r, \omega) = (a, 0)$, where $a < 1$. The unit circle ($r = 1$) is shown as a dashed line for reference. **B** shows the same component on a *z*-plane pole-zero diagram, where X marks the pole position. **C** shows the corresponding time-domain waveform where the amplitude falls exponentially over time.

## 9.3.3 Sinusoid cos($n\omega_0$) $u[n]$ and sin($n\omega_0$) $u[n]$

Consider the *z*-transform of the cosine $x[n] = \cos(n\omega_0)\, u[n]$ for $\omega_0 < \pi$ to preserve a minimum sampling rate:

$$U(z) = \sum_{n=0}^{\infty} [\cos(n\omega_0)\, u[n]]\, z^{-n} = \sum_{n=0}^{\infty} \cos(n\omega_0) z^{-n}$$

which may be written as the sum of two complex exponentials:

$$U(z) = \sum_{n=0}^{\infty} \left[\frac{e^{+j\omega_0 n} + e^{-j\omega_0 n}}{2}\right] z^{-n}$$

or

$$U(z) = \frac{1}{2} \sum_{n=0}^{\infty} \left[e^{+j\omega_0 n}\right] z^{-n} + \frac{1}{2} \sum_{n=0}^{\infty} \left[e^{-j\omega_0 n}\right] z^{-n}$$

then

$$U(z) = \frac{1}{2} \sum_{n=0}^{\infty} \left[\frac{e^{+j\omega_0}}{z}\right]^n + \frac{1}{2} \sum_{n=0}^{\infty} \left[\frac{e^{-j\omega_0}}{z}\right]^n$$

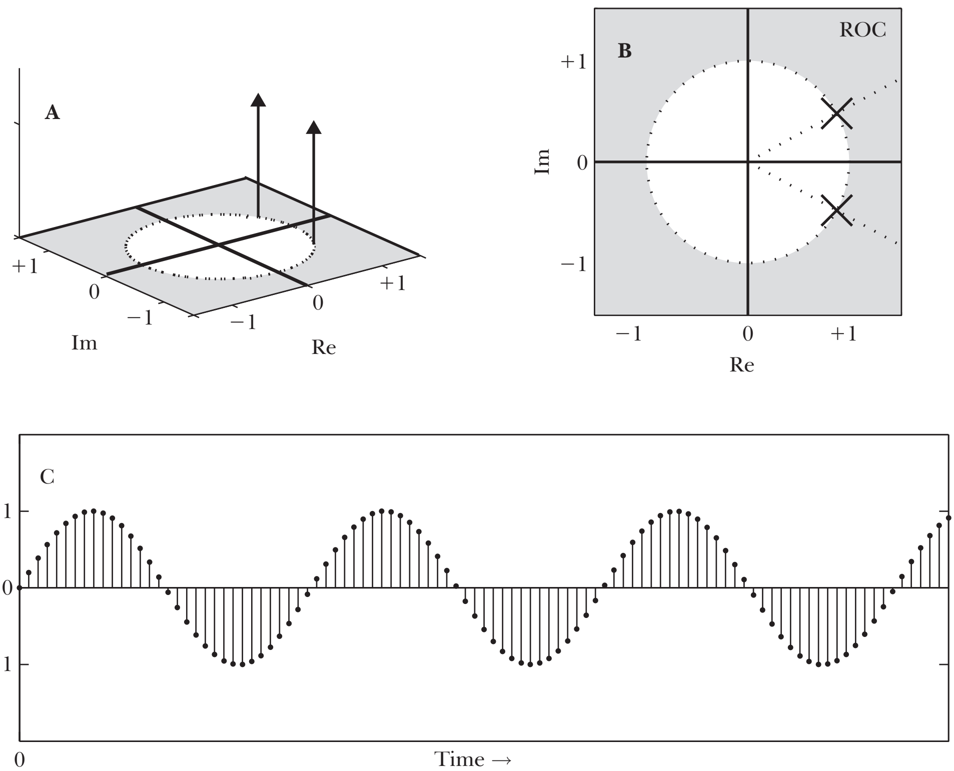



**FIGURE 9.6** ***z*-Transform Sinusoid A** is a three-dimensional view of the familiar Fourier transform *impulse pair* from a time-domain sinusoid. The region of convergence is shown shaded. **B** puts the same components on a *z*-plane pole-zero diagram, where X marks the pole positions. Zeros are not shown for clarity. These poles lie on the unit circle, shown as a dashed line for reference. **C** shows the corresponding discrete time waveform. Compare to Eqns. 9.10 and 9.11.

from Eqn. 9.5, this may be resolved directly to give two terms:

$$U(z) = \frac{1}{2}\left[\frac{1}{1-\frac{e^{+j\omega_0}}{z}} + \frac{1}{1-\frac{e^{-j\omega_0}}{z}}\right] = \frac{1}{2}\left[\frac{z}{z-e^{+j\omega_0}} + \frac{z}{z-e^{-j\omega_0}}\right]$$

then adding the two fractions:

$$U(z) = \frac{1}{2}\left[\frac{z(z-e^{-j\omega_0}) + z(z-e^{+j\omega_0})}{(z-e^{+j\omega_0})\times(z-e^{-j\omega_0})}\right] = \frac{1}{2}\left[\frac{z(2z-[e^{-j\omega_0}+e^{+j\omega_0}]}{z^2-z[e^{+j\omega_0}+e^{-j\omega_0}]+1}\right]$$

and finally recognizing the complex exponential terms giving $\cos(\omega_0)$, the $z$-transform is:

$$\cos(n\omega_0)u[n] \overset{z}{\longleftrightarrow} \frac{z^2 - z\cos(\omega_0)}{z^2 - 2z\cos(\omega_0) + 1} \quad \text{ROC} = \quad |z| > 1 \tag{9.10}$$

This $z$-transform will have two zeros, at $z = 0$ and $z = \cos(\omega_0)$, where the position of the second zero along the real axis will vary over the interval $[-1, +1]$ with the constant value $\omega_0$. There will also be two complex roots corresponding to the quadratic roots of the denominator; these roots lie on the unit circle as shown in Figure 9.6.

It is left as an exercise to derive the corresponding z-transform for $\sin(n\omega_0)$:

$$\sin(n\omega_0)\, u[n] \stackrel{z}{\longleftrightarrow} \frac{z \sin(\omega_0)}{z^2 - 2z\cos(\omega_0) + 1} \qquad \text{ROC} = \quad |z| > 1 \tag{9.11}$$

This z-transform will have a single zero at $z = 0$. Both the sine and cosine have the same denominator and therefore the same paired poles on the unit circle as shown in Figure 9.6. Similarly, sinusoidal components at the same frequency but inside or outside the unit circle would lie on radial lines with the same angle $\omega$ radians as shown highlighted in the figure.

## 9.3.4 Differentiation

The behavior of the z-transform under differentiation provides another useful tool to simplify calculations. For a signal $x[n]$ the z-transform is given by:

$$X(z) = \sum_{n=0}^{\infty} x[n]z^{-n}$$

and its derivative with respect to $z$ is:

$$\frac{d}{dz}\, X(z) = \frac{d}{dz}\sum_{n=0}^{\infty} x[n]\, z^{-n} = \sum_{n=0}^{\infty} -n\, x[n]\, z^{-n-1} = -\frac{1}{z}\sum_{n=0}^{\infty} [nx[n]]\, z^{-n}$$

giving:

**THEOREM 9.3** *(Derivative)*

*If*

$$x[n] \stackrel{z}{\longleftrightarrow} X(z)$$

*then*

$$n\, x[n] \longleftrightarrow -\, z\frac{d}{dz}X(z)$$

## 9.3.5 The Effect of Sampling Rate

Because the z-transform works on discrete (sampled) signals, the specific angular position $\omega$ of the paired poles in this example depends on the sample rate $f_s$ in use, or the time between successive discrete signal points. On the z-plane, the angular value $\omega$ can vary over the interval $[-\pi, +\pi]$ radians, and it has generally been assumed in this discussion that $T = 1$ in the signal $s(nT)$ such that $f_s = 1$ and the radial line at $\omega = \pi$ radians corresponds to a component frequency $f_s/2$ or 0.5 Hz. For a given sample rate $f_s$, the angular value of $\omega$ in radians is related to the actual cosine frequency $f_0$ or $\omega_0$ as:

$$\omega = \frac{2\pi f_0}{f_s} = \frac{\omega_0}{f_s}\, radians$$

This result means that for a higher sample rate, the poles in Figure 9.6B would be found closer together for the same frequency cosine input signal.

**Check:** For sample rate, $T = 1$ s or $f_s = 1$ Hz, the above result reduces to $\omega = 2\pi f_0$ as expected.

**EXAMPLE 9.2 (Pole Location)**

*Where are the z-transform poles located for the signal* $\mathrm{s(t) = cos(200\pi t)}$ *sampled at* $\mathrm{f_s = 1000}$ *Hz?*

**Solution:**

The two poles are located on the unit circle ($r = 1$) at points $(r, \omega) = (1, +\omega)$ and $(1, -\omega)$. As shown in Figure 9.7, $f = 100$ Hz and $f_s = 1000$ Hz:

$$\omega = \frac{2\pi f}{f_s} = \frac{200\pi}{1000} = \frac{\pi}{5} \text{ rad} = 36 \text{ deg}$$

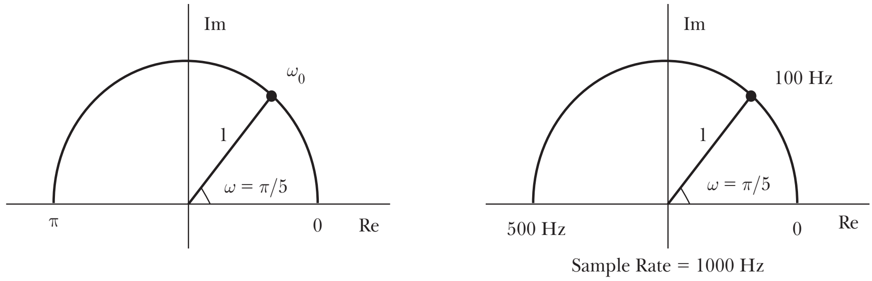



**FIGURE 9.7** ***z*-Domain Unit Circle** The sampling rate $f_s$ determines the specific frequencies found around the unit circle on the *z*-plane. The same (negative) frequency components appear for the lower half plane.

# 9.4 A Discrete Time Laplace Transform

The *z*-transform definition can written with $z = re^{j\omega}$ to give:

$$X(z) = \sum_{n=0}^{+\infty} x[n](re^{j\omega})^{-n} = \sum_{n=0}^{+\infty} [x[n]r^{-n}]e^{-j\omega n} \tag{9.12}$$

which resembles the Laplace transform with $e^{-\sigma n}$ replaced by $r^n$. In the exponential form $r^{-n}$, the value of $r$ serves the role of $\sigma$ from the Laplace transform. For the special case of $r = 1$ (as with $\sigma = 0$), Eqn. 9.12 becomes the Fourier transform of $x[n]$, or:

$$X(z) = \sum_{n=0}^{+\infty} x[n]e^{-j\omega n} \tag{9.13}$$

which is the discrete time Fourier transform of $x[n]\ u[n]$. As $\omega$ varies, values of $X(z)$ are found around a unit circle as observed in Figure 9.1. Continuing as above, the Laplace transform of the discrete time signal $x[n]$ is obtained directly when $z = e^{st}$, and the *z*-transform becomes:

$$X(z) = \sum_{n=0}^{+\infty} x[n]e^{-sn} = \sum_{n=0}^{+\infty} [x[n]e^{-\sigma n}]e^{-j\omega n} \tag{9.14}$$

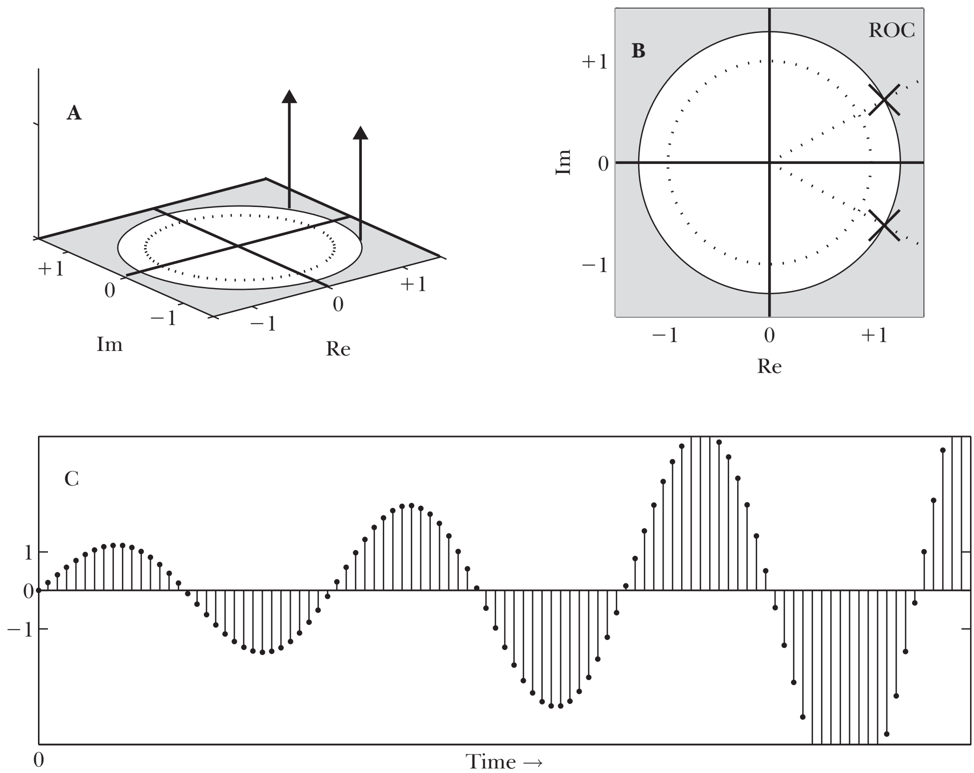



**FIGURE 9.8** **The *z*-Plane—($r > 1$)** **A** is a three-dimensional view of the familiar Fourier transform components of a sinusoid displayed when $r > 1$. The unit circle ($r = 1$) is shown as a dashed line for reference. **B** shows the same components on a *z*-plane pole-zero diagram, where X marks the pole positions. **C** shows the corresponding time-domain waveform where the amplitude grows exponentially over time. Compare to Figure 9.6.

Figures 9.6 through 9.9 illustrate the *z*-transform as paired poles located inside, outside, or directly on the unit circle for values of $(r, \omega)$ with a fixed frequency $\omega_0$. Any zeros associated with the respective *z*-transforms are not shown.

In Figure 9.8A, a *z*-transform is sketched as a pair of poles at position $(r, \omega) = (+0.5, \omega_0)$ and $(-0.5, \omega_0)$ on the *z*-domain. This sketch evokes the Fourier transform of a sinusoid with a decreasing exponential term, as $r < 1$. The unit circle is shown as a dashed line for reference. In this case, the region of convergence is outside a circle of radius $r = 0.5$ so as not to include any poles; this region is shown shaded on the figure. Because these poles lie within the unit circle, a response function corresponding to this *z*-transform would be expected to be stable. Because the unit circle is within the region of convergence, the Fourier transform exists for this signal. The pole-zero diagram of Figure 9.8B shows a pair of poles directly on the radial lines corresponding to frequencies of $+\omega_0$ and $-\omega_0$, each the same distance $r = 0.5$ from the origin. Figure 9.8C shows the corresponding discrete time-domain waveform as a sinusoid with decreasing amplitude. This situation corresponds to components lying to the left of $\sigma = 0$ axis of the Laplace transform.

In Figure 9.9C, a *z*-transform is sketched as a pair of poles at position $(r, \omega) = (+1.5, \omega_0)$ and $(-1.5, \omega_0)$ on the *z*-domain. This sketch evokes the Fourier transform of a sinusoid with an increasing exponential term, as $r > 1$. The unit circle is

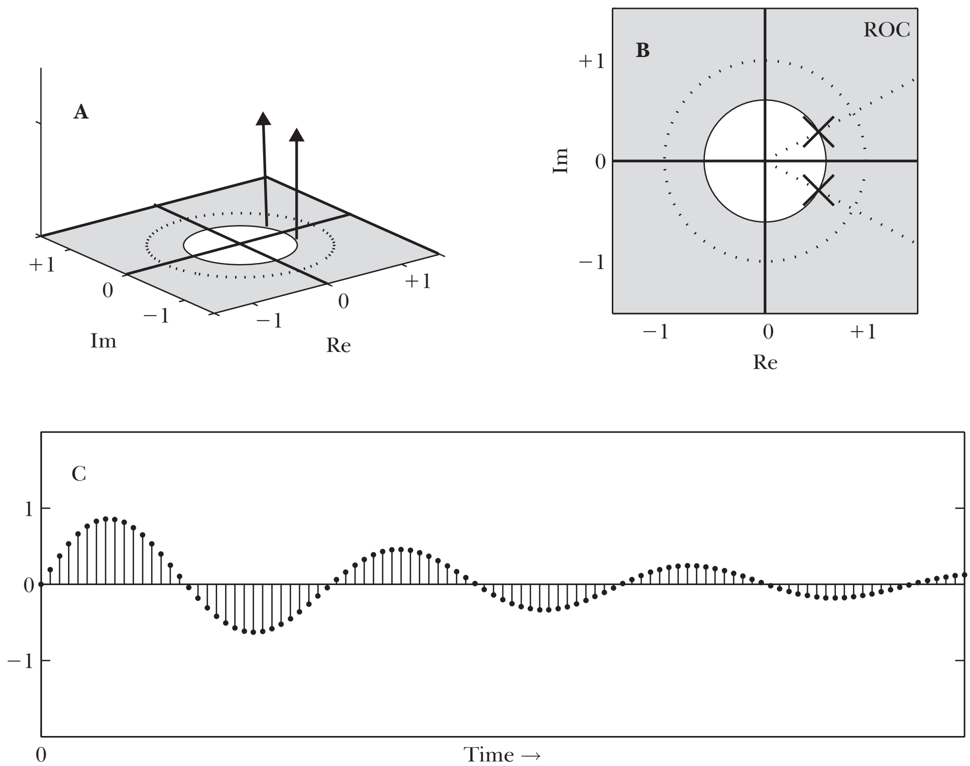



**FIGURE 9.9** **The *z*-Plane—($r < 1$) A** is a three-dimensional view of the familiar Fourier transform components of a sinusoid displayed when $r < 1$. The unit circle ($r = 1$) is shown as a dashed line for reference. **B** shows the same components on a *z*-plane pole-zero diagram, where X marks the pole positions. **C** shows the corresponding time-domain waveform where the amplitude falls exponentially over time. Compare to Figure 9.6.

shown as a dashed line for reference. In this case, the region of convergence is outside a circle of radius $r = 1.5$ so as not to include any poles; this region is shown shaded on the figure. Because these poles lie outside the unit circle, a response function corresponding to this *z*-transform would not be expected to be stable. Because the unit circle is not within the region of convergence, the Fourier transform does not exist for this signal. The pole-zero diagram of Figure 9.9B shows a pair of poles directly on the radial lines corresponding to frequencies of $+\omega_0$ and $-\omega_0$, each the same distance $r = 1.5$ from the origin. Figure 9.9C shows the corresponding discrete time-domain waveform as a sinusoid with increasing amplitude. This situation corresponds to components lying to the right of the $\sigma = 0$ axis of the Laplace transform. The corresponding *z*-transform would have a zero in the denominator when $z = 0.5$.

## 9.5 Properties of the *z*-Transform

The properties of the *z*-transform parallel those of the Fourier transform and the Laplace transform.

linearity:

$$A\, s_1[n] + B\, s_2[n] \xleftrightarrow{z} A\, S_1(z) + B\, S_2(z)$$

amplifying:

$$ks[n] \overset{\mathcal{Z}}{\longleftrightarrow} kS(z)$$

scaling:

$$k^n s[n] \overset{\mathcal{Z}}{\longleftrightarrow} S\left(\frac{z}{k}\right)$$

shifting:

$$s[n-k] \overset{\mathcal{Z}}{\longleftrightarrow} z^k\, S(z)$$

convolution:

$$a[n]*b[n] \overset{\mathcal{Z}}{\longleftrightarrow} A(z) \times B(z)$$

derivative:

$$nx[n] \overset{\mathcal{Z}}{\longleftrightarrow} -z\frac{d}{dz}X(z)$$

# 9.6 *z*-Transform Pairs

A few important *z*-transform pairs are summarized here for reference.

$$\delta[n] \overset{z}{\longleftrightarrow} 1 \quad \text{ROC} = \quad \text{all } z \tag{9.15}$$

$$u[n] \overset{z}{\longleftrightarrow} \frac{z}{z-1} \quad \text{ROC} = \quad |z| > 1 \tag{9.16}$$

$$n\, u[n] \overset{z}{\longleftrightarrow} \frac{z}{(z-1)^2} \quad \text{ROC} = \quad |z| > 1 \tag{9.17}$$

$$a^n\, u[n] \overset{z}{\longleftrightarrow} \frac{z}{z-a} \quad \text{ROC} = \quad |z| > |a| \tag{9.18}$$

$$\cos(n\omega_0)\, u[n] \overset{z}{\longleftrightarrow} \frac{z^2 - z\cos(\omega_0)}{z^2 - 2z\cos(\omega_0) + 1} \quad \text{ROC} = \quad |z| > 1 \tag{9.19}$$

$$\sin(n\omega_0)\, u[n] \overset{z}{\longleftrightarrow} \frac{z\sin(\omega_0)}{z^2 - 2z\cos(\omega_0) + 1} \quad \text{ROC} = \quad |z| > 1 \tag{9.20}$$

# 9.7 Transfer Function of a Discrete Linear System

As in the continuous case, if $h[n]$ is the response function of a linear system, any input signal is convolved with $h[t]$ to produce the output signal $g[n] = s[n]*h[n]$. Moreover, in the *z*-domain, the same operation corresponds to a multiplication: $G(z) = S(z)H(z)$. The *z*-transform $H(z)$ of the response function $h[n]$ is identified as the *transfer function* of the system. Either the (impulse) response function or the transfer function can be used to uniquely characterize a specific discrete linear system.

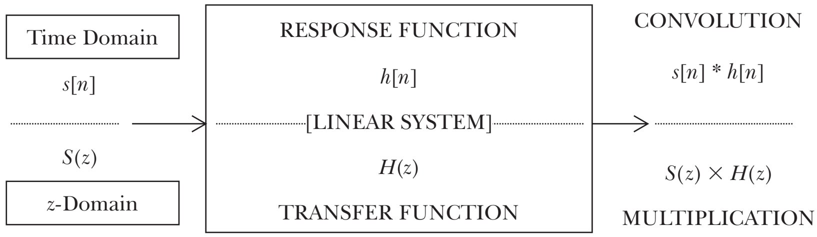



**FIGURE 9.10** A discrete linear system can be categorized in either the time domain or the *z*-domain. The time-domain impulse response $h[n]$ is convolved with an input signal $s[n]$. In the *z*-domain, input signals are multiplied by $H(z)$.

# 9.8 MATLAB Analysis with the *z*-Transform

The transfer function $H(z)$ can serve to define a *system* in MATLAB after which the system behavior and properties can be explored using the functions `bode`, `pzmap`, `impulse`, `step`, and `lsim`. In Chapter 7, these same functions were used with continuous time systems and the Laplace transform *s*-domain. These functions and their use are summarized in Appendix D.

The MATLAB function `tf()` is used to define a system transfer function directly from terms found in the numerator and denominator of $H(z)$. With the *z*-transform, this function also requires a *sampling interval* (T) for the discrete values $s(nT)$ leading to the transform function. While the same `tf()` function is used to define Laplace *s*-domain transfer functions, the inclusion of a sampling rate parameter distinguishes the *z*-domain transfer function for discrete systems.

## 9.8.1 First-Order Lowpass Filter

Consider the continuous time system in Figure 9.11, which is an RC circuit with an exponential impulse response. The cutoff frequency of this first-order lowpass filter is $\omega_c = 1/RC$.

$$h(t) = e^{-\frac{1}{RC}t}$$

This circuit behavior can be simulated in a digital system using the *z*-transform. The *z*-transform of a discrete general exponential signal $s[n] = a^n$ has already been identified and can serve as the corresponding discrete causal response function $h(t) = a^n u[n]$. From Eqn. 9.9:

$$h[n] = a^n\, u[n] \xleftrightarrow{z} \frac{z}{z-a} \quad \text{ROC} = \quad |z| > a$$

which gives the *z*-domain transfer function $H(z)$ of this first-order lowpass filter:

$$H(z) = \frac{z}{z-a} = \frac{1z+0}{1z-a} \tag{9.21}$$

In MATLAB, the numerator and denominator terms can be represented by the vectors $[1, 0]$ and $[1, -a]$, respectively. Once suitable values have been established for $a$ and for the sampling interval $T$, the system can be defined as:

```
H = tf([1 0], [1 -a], T); % H = transfer function
```

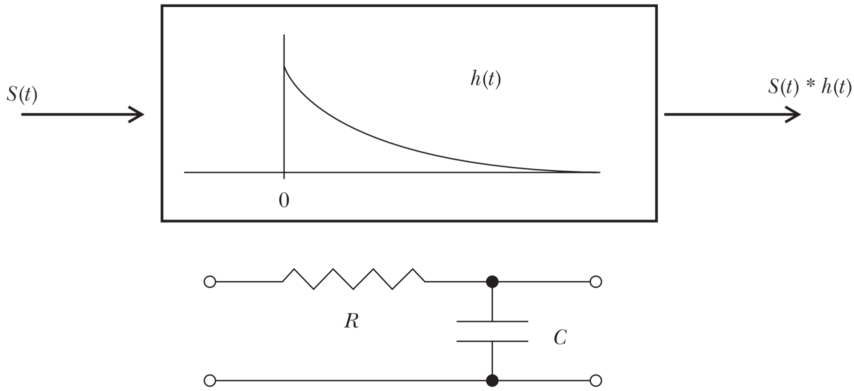



**FIGURE 9.11** An RC lowpass filter to be implemented using *digital signal processing*. The input signal is sampled and the exponential response function $h(t)$ is applied using the $z$-transform.

Alternatively, the same system (H) can be defined directly from the above equation for $H(z)$, using:

```
z = tf('z', T);   % define 'z' and sample rate
H = z/(z - a);    % H = transfer function
```

It is now necessary to relate the choice of $a$ to the circuit behavior. The cutoff frequency $\omega_c$ for this circuit depends on the choice of $(R,C)$ and the exponential impulse response $h(t)$ is defined as $h(t) = e^{(-1/RC)t}$, to give a cutoff $\omega_c = 1/RC$ rad/s. For the discrete case, the cutoff frequency $w_c = 2\pi f_c$ depends on the choice of $a$ and the sampling interval $T = 1/f_s$. It is necessary to find a value for $a$ in the $z$-transform that relates these two expressions for the discrete system by solving for:

$$e^{(-\omega_c)nT} = a^n$$

The required value of $a$ is found directly by recognizing that $a = e^{(-\omega_c T)}$.

Let the cutoff frequency be 10 Hz and choose a sampling rate $f_s = 1000$ Hz. Then $T = 1/1000$ and the required value of $a$ is given by:

$$a = e^{-\omega_c T} = e^{\frac{-20\pi}{1000}} = e^{-0.0628} = 0.939$$

Continuing with $a = 0.939$ and $T = 1/1000$, the MATLAB system (H) may now be defined:

```
H = tf([1 0], [1 -0.939], 1/1000)   % define the system
Transfer function:
    z
---------
z - 0.939
Sampling time : 0.001
```

MATLAB returns the $z$-domain transfer function and the supplied sample interval. Once the system is defined in this way, it can be studied directly through the H variable. The pole-zero diagram, Bode plot, and impulse response will now be examined.

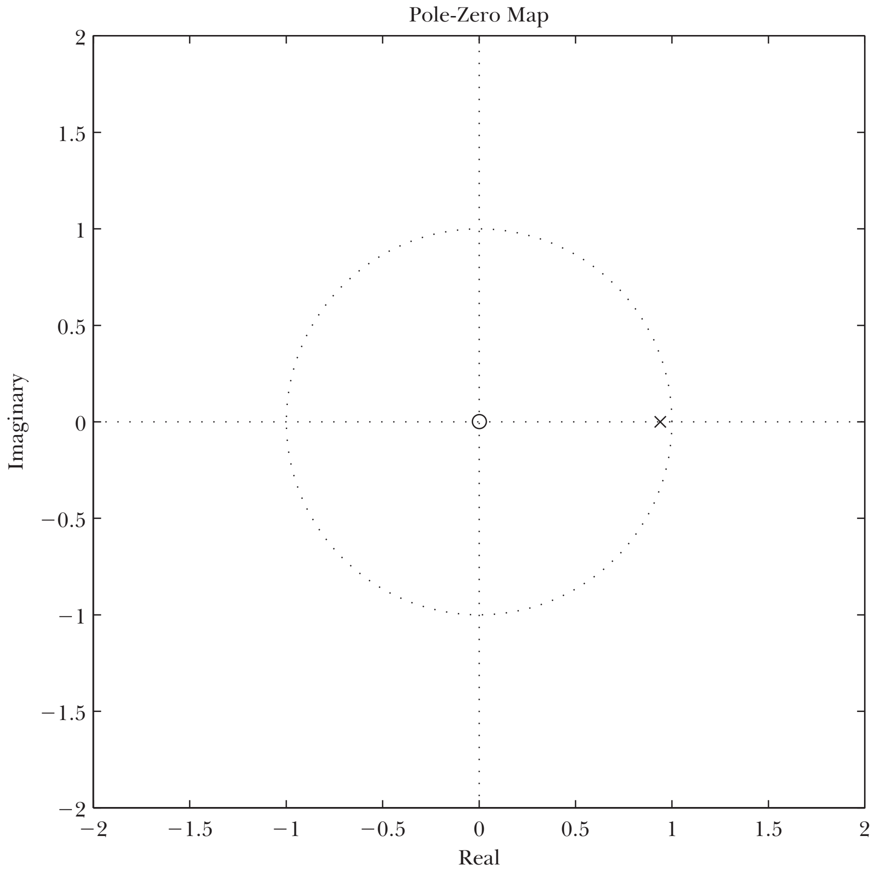



**FIGURE 9.12 First Order LPF—Pole-Zero Diagram** The pole-zero diagram uses this z-plane form whenever the MATLAB `tf()` function specifies a sampled system. The plot shows a single pole inside the unit circle, and a single zero at the origin.

## 9.8.2 Pole-Zero Diagram

```
pzmap(H); % see the pole-zero diagram
axis([ - 2 2  -2 2];
```

The pole at $z = 0.939$ and the pole at $z = 0$ are visible in the pole-zero diagram of Figure 9.12. The pole lies within the unit circle as expected. Because a sample interval was specified when `H` was defined, `pzmap(H)` produces a pole-zero diagram featuring the unit circle.

## 9.8.3 Bode Plot

```
bode(H); % see the Bode plot
grid on;
```

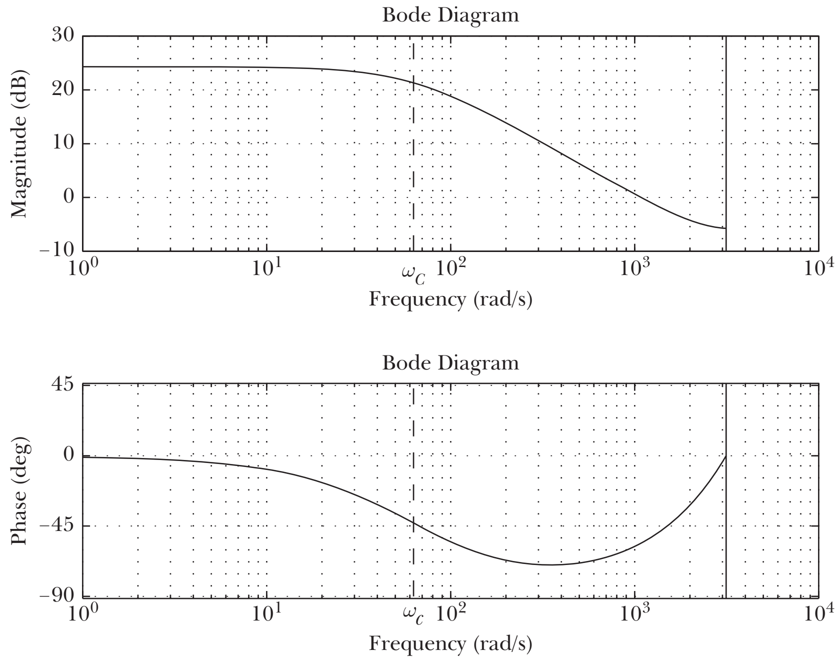



**FIGURE 9.13 First-Order LPF—Bode Plot** The Bode plot necessarily cuts off at one half the sampling frequency, or $\pi T$ rad/s, as seen in Figure 9.14. The magnitude plot shows the characteristic first-order lowpass filter transfer function with a slope of −20 dB/dec.

The Bode plot for this discrete system is shown in Figure 9.13, where both the phase and magnitude plots terminate abruptly at $\omega = 1000\pi = 3141$ rad/s. The maximum frequency found in a discrete system is necessarily less than the Nyquist frequency or one half the sample rate, and because a sample rate of 1000 Hz was chosen this upper limit is 500 Hz or $\omega_{max} = 2\pi(500)$ rad/s. This effect is further illustrated by the general example of Figure 9.14, where the Bode plot is shown to be extracted from the positive unit circle on the $z$-plane and the limiting frequency is given by $\pi$T rad/s.

The magnitude plot in Figure 9.13 shows the distinctive appearance of a first-order lowpass filter, with a downward slope of 20 dB/decade as expected. The cutoff frequency shown highlighted is $w_c = 62.8$ rad/s $= 10$ Hz as was chosen during the design of this filter. At the cutoff frequency, the magnitude is down −3 dB, and the phase change is $-\pi/4$ rad = −45 deg as expected. Both the magnitude and phase graphs deviate from the continuous result as the frequency approaches $\omega_{max}$; in fact, both these plots would be periodic if the journey around the unit circle were continued such that more accurate results were found at the lower frequencies. For higher sampling rates, the same cutoff frequency would give larger $a$, and the pole would move closer to the unit circle; the most accurate results are found using higher sampling rates.

## 9.8.4 Impulse Response

```
impulse(H); % see the impulse response
axis([0 0.03  −2 2 ])
grid on;
```

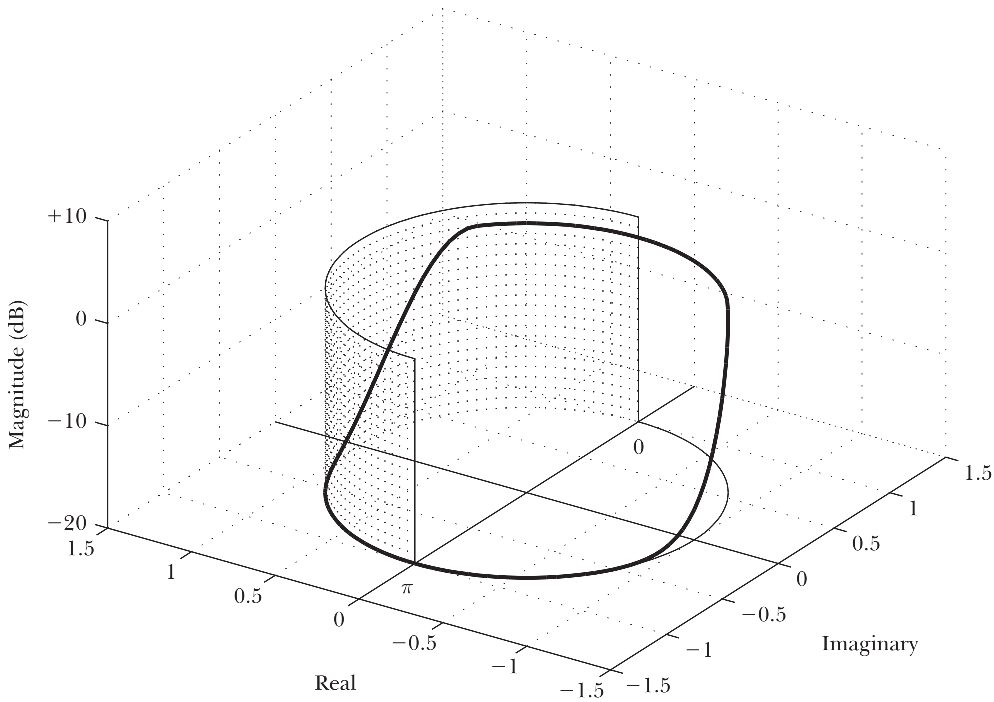



**FIGURE 9.14** ***z*-Domain Bode Plot** The Bode plot displays a strip from the *z*-plane as shown in this typical *lowpass filter*. The positive side of the unit circle spans frequencies from $\omega = 0$ to $\pi$ where the Bode plot terminates, assuming a sampling interval of $T = 1$ s. Periodic repetitions of the transfer function would be encountered if $\omega$ increased past this point.

The impulse response of the system is confirmed in Figure 9.15, where the discrete nature of the system is evident in the stepped curve. A discrete system has successfully been created to model an analog lowpass filter.

In this section, the *z*-transform has been used to emulate the behavior of a continuous time lowpass filter. A more systematic approach using the *z*-transform can be used to design and to analyze more capable digital filters. This important application of the *z*-transform is explored in the next section.

## 9.8.5 Calculating Frequency Response

Given a *z*-transform $H(z)$, the corresponding Fourier transform is found around the unit circle where $z = e^{j\omega}$, provided that the unit circle lies within the region of covergence.

Consider the *z*-transform $H(z)$. The magnitude of the (complex) $H(z)$ is given by $|H(z)| = [H(z)H^*(z)]^{\frac{1}{2}}$.

For

$$H(z) = \frac{z}{z - a}$$

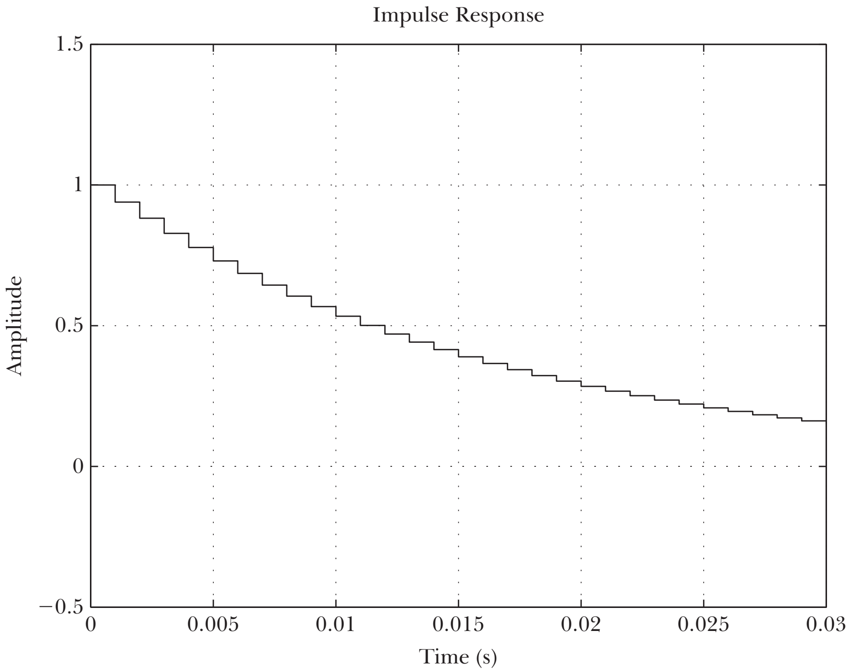



**FIGURE 9.15 First-Order LPF—Impulse Response** The discrete (stepped) impulse response shown here as determined from the MATLAB transfer function confirms the exponential form of $h[n]$ that was initially used to compute $H(z)$.

then

$$|H(z)| = [H(z)H^*(z)]^{\frac{1}{2}} = \left[\frac{z}{(z-a)}\,\frac{z^*}{(z^*-a)}\right]^{\frac{1}{2}}$$

$$= \left[\frac{zz^*}{zz^* - az - az^* + a^2}\right]^{\frac{1}{2}} = \left[\frac{zz^*}{zz^* - a(z+z^*) + a^2}\right]^{\frac{1}{2}}$$

For the specific case of points on the unit circle, let $z = e^{j\omega}$, then:

$$|H(e^{j\omega})| = \left[\frac{e^{j\omega}e^{-j\omega}}{e^{j\omega}e^{-j\omega} - a[e^{j\omega} + e^{-j\omega}] + a^2}\right]^{\frac{1}{2}}$$

The product $e^{+j\omega}e^{-j\omega} = 1$, and the complex exponential form of $\cos(\omega)$ is recognized in the denominator, leaving:

$$|H(e^{j\omega})| = \left[\frac{1}{1 - 2a\cos(\omega) + a^2}\right]^{\frac{1}{2}}$$

Therefore, the frequency response is given by:

$$|H(\omega)| = \frac{1}{\sqrt{1 + a^2 - 2a\cos(\omega)}} \tag{9.22}$$

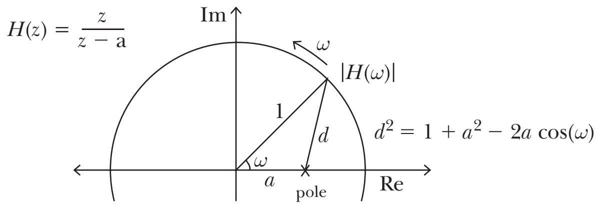



**FIGURE 9.16 First-Order LPF—Transfer Function** The single pole at $z = a$ lies a distance $d$ from the points along the unit circle. As $\omega$ increases from 0 to $\pi$ radians, $d$ grows larger, and the effect of the pole diminishes as $1/d$. The transfer function $|H(\omega)|$ describes a first-order lowpass filter.

### 9.8.6 Pole Position Determines Frequency Response

Poles on the $z$-plane define points of infinite value whose influence falls off inversely with distance from the pole. The frequency response can be found directly by examining the effect of pole position on points around the unit circle.

Let a single pole be found at $z = a$ as shown in Figure 9.16 as associated with a first-order lowpass filter ($a < 1$). The corresponding Fourier transform lies along the unit circle where the transfer function $|H(\omega)|$ is located a distance $d$ from the pole. The presence of the pole determines the appearance of the transfer function, which varies as $1/d$ for a given value of $\omega$. It is necessary to determine $d$ as a function of $\omega$.

By inspection, the value of $d$ will grow larger as $\omega$ increases, going from a minimum of $d = 1 - a$ when $\omega = 0$ to a maximum of $d = 1 + a$ when $\omega = \pi$. The exact value of $d$ is computed using the cosine rule of trigonometry, where, for a general radius $r$:

$$d^2 = r^2 + a^2 - 2ra\cos(\omega)$$

in this case, $r = 1$, giving:

$$d^2 = 1 + a^2 - 2a\cos(\omega)$$

as shown in the figure. The transfer function is then given by:

$$|H(\omega)| = \frac{1}{d} = \frac{1}{\sqrt{1 + a^2 - 2a\cos(\omega)}}$$

This result matches Eqn. 9.22 obtained using the $z$-transform.

## 9.9 Digital Filtering—FIR Filter

In Chapter 8, the design of digital filters was examined based on delay blocks as in Figure 9.17.[2] It has been shown that each term $nT$ in the discrete response function may be interpreted as an integer power of $e^{-j2\pi f}$ in the Fourier transform. The $z$-transform provides an elegant way to describe and to systematically design filters using that approach. By applying the $z$-transform, each delay block in a system can be replaced by $z^{-1}$, and each delay term $a_n$ in the discrete response function directly corresponds to a term in $z^{-n}$. Several filter configurations will be presented to elaborate this important application of the $z$-transform.

[2] See, for example, Figures 8.17 and 8.21.

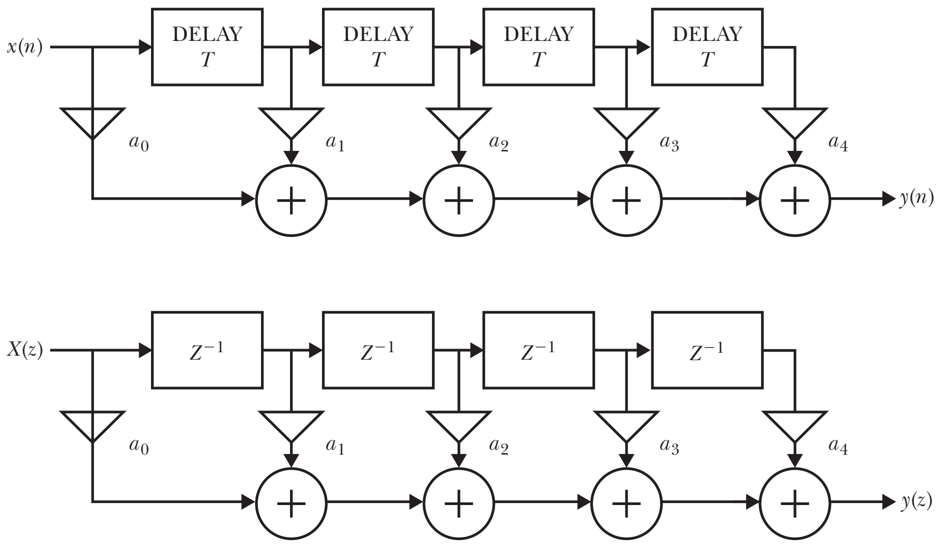



**FIGURE 9.17 An FIR Filter** A delay block system is conveniently relabelled to show its $z$-transform by treating each delay block as $z^{-1}$.

A discrete impulse $\delta[n]$ can be defined as:

$$\delta[n] = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise} \end{cases} \tag{9.23}$$

Consider the delay block system of Figure 9.17 with input $x[n]$ and output $y[n]$. For an input $\delta[n]$, the output gives the *impulse response* of the system. The corresponding $z$-transform of the circuit gives the *transfer function* $H(z)$ directly as:

$$H(z) = \frac{Y(z)}{X(z)} = a_0 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + a_4 z^{-4} + \ldots + a_n z^{-n}$$

where the terms $a_n$ correspond to each of the terms describing the system blocks. This figure represents a *finite impulse response* (FIR) filter, as an impulse input leads to an output that continues in time (samples) only as long as there are delay blocks in the system. The sum of the values $|a_n|$ limits the maximum value that an input can produce and such a system is called *bounded input, bounded output* (BIBO).

## 9.9.1 A One-Pole FIR Filter

By choosing only the terms $(a_0, a_1)$ from the model of Figure 9.17, a simple *one-pole* filter can be designed. The system impulse response becomes

$$H(z) = a_0 + a_1 z^{-1} = a_0 + \frac{a_1}{z} = \frac{a_0 z + a_1}{z}$$

which, by inspection, has a pole at $z = 0$ and a zero at $z = -a_1/a_0$.

With a pole at the origin, the region of convergence spans the $z$-plane except at $z = 0$. Because the region of convergence includes the unit circle, this LTI system is stable, and the Fourier transform components can be found around the unit circle by setting $z = e^{-j\omega}$. This allows determination of the frequency and phase response of this system as a function of $(a_0, a_1)$.

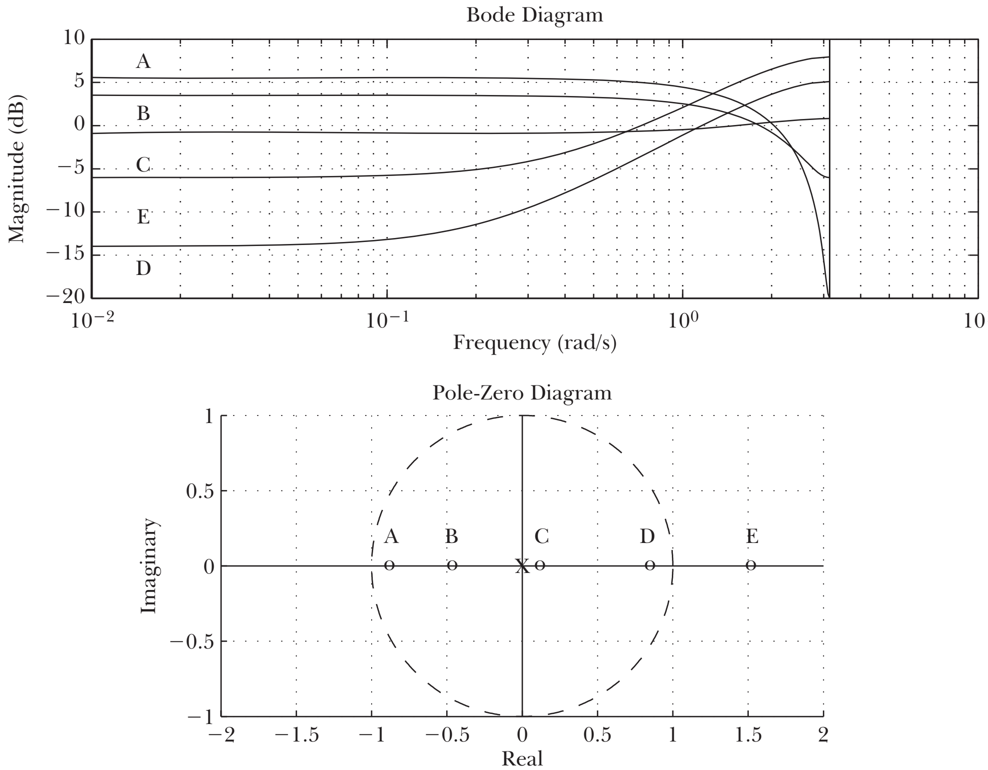



**FIGURE 9.18 An FIR Filter—One Pole** A single-pole FIR filter has a single pole at the origin and a single zero that may move along the real axis. This figure shows the effect on the transfer function of moving the zero to one of five positions (A, B, C, D, E). Compare to Figure 9.24 for the IIR filter.

To explore the effect of varying the position of the zero (which always lies on the real axis), a short MATLAB sequence can be easily modified to place the single zero as desired. In Figure 9.18, five different positions are shown as $(A, B, C, D, E) = (-0.9, -0.5, -0.1, +0.8, +1.5)$.

```
H = tf([1 0.5 ], [1 0 ], 1); % zero at z =  - 0.5
bode(H);
grid on;
```

From the figure, zeros near the origin (C) have very little effect on an input signal, while zeros close to the unit circle (A) have the greatest effect. Any zero directly on the unit circle would force the frequency response to zero at that point. In general, negative zero positions (A, B) produce a lowpass filter, while positive positions (C, D, E) achieve a highpass filter. By varying the ratio $a_1/a_0$ in $H(z)$ the filter characteristics may be adjusted as desired. For any FIR filter, the pole(s) will lie on the origin, while the zeros serve to determine the form of the transfer function.

## 9.9.2 A Two-Pole FIR Filter

By choosing only the terms $(a_0, a_1, a_2)$ from the model of Figure 9.17, a simple *two-pole* filter can be designed. The system impulse response becomes:

$$H(z) = a_0 + a_1 z^{-1} + a_2 z^{-2} = a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} = \frac{a_0 z^2 + a_1 z + a_2}{z^2}$$

which, by inspection, has a double pole at $z = 0$ and two zeros located at the roots of the numerator. In general, an $N$-pole FIR filter will have $N$ poles at the origin, while the numerator will be polynomial of degree $N$ leading to $N$ zeros.

From the quadratic numerator, zeros are found directly as:

$$z = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0 a_2}}{2a_0}$$

which gives either two real zeros or a pair of complex zeros located above and below the real axis.

**Check:** When $a_2 = 0$, the above expression should reduce to the one pole case. Let $a_2 = 0$, then the single zero is at $z = -a_1/a_0$ as expected.

## 9.9.3 Higher-Order FIR Filters

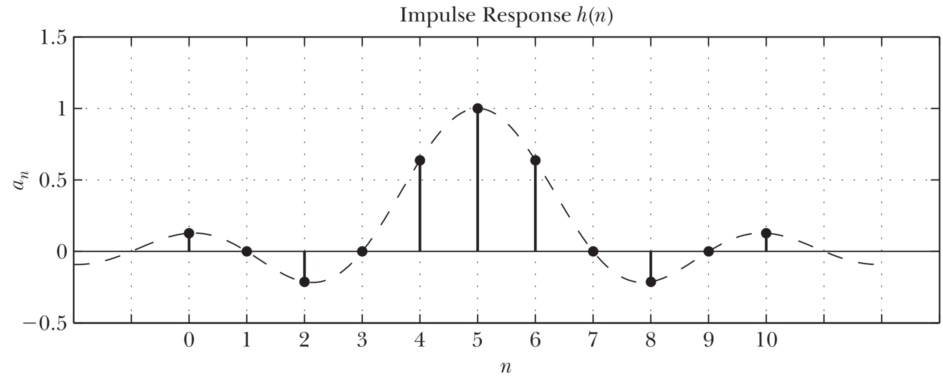
By choosing additional terms $a_n$ from the model of Figure 9.17, higher-order filters may be defined. Given a target frequency response, the principles of the DFT may be used to define the necessary values $a_n$ in the discrete impulse response. For example, the lowpass filter having a very sharp cutoff frequency (a frequency-domain rectangle, also know as a *brick wall filter*) would correspond to a sinc function in the time domain. To be causal, the sinc must be shifted and truncated appropriately, as shown in Figure 9.19, where 11 terms in $a_n$ are defined to create an FIR filter. Let the sampling rate be 8000 Hz.

This filter may be defined directly in MATLAB as:

```
n =0:10;                     % use 11 values of an
num = sinc((n-5)/2);         % H(z) numerator
den = [1 zeros(1,10)];       % H(z) denominator
H = tf(num, den, 1/8000);    % Fs = 8000~Hz
bode(H);                     % see the transfer function
pzmap(H);                    % see the pole-zero diagram
step(H);                     % see the step response
```

**Frequency Response** The Bode plot for the resulting frequency response is shown in Figure 9.19, where a nearly rectangular lowpass filter has been achieved. Beyond the cutoff frequency, the magnitude plot falls off rapidly. The three frequencies on the graph where the magnitude appears to fall off to zero correspond to the zeros of $H(z)$, which lie on the unit circle in the pole-zero diagram. The Bode plot ends at 4000 Hz = 25132 rad/s corresponding to one half the 8000 Hz sample rate.

The cutoff frequency of this particular lowpass filter depends on the width of the sinc function $h[n]$. If the first zero crossing of the sinc occurs $N$ samples from the sinc origin, then the frequency-domain rectangle has width $W = 1/N$, and the positive frequency half of the rectangle seen in the Bode plot extends to $f = 1/2N$; this is the cutoff frequency $\omega_c$. In this example $N = 2$, and for a sample rate $f_s = 8000$ Hz, the width of $N = 2$ samples is 2/8000 sec; the cutoff frequency is $f_c = 2000$ Hz or $\omega_c = 12566$ rad/s. This line has been highlighted on the Bode plot.



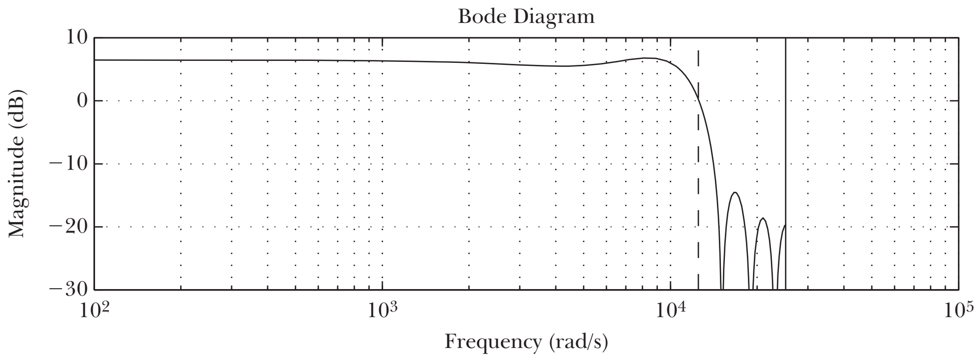




**FIGURE 9.19 A High-Order FIR Filter—Frequency Response** The *z*-transform terms $a_n$ are directly defined by a discrete impulse response. Here, a discrete sinc function is suitably truncated and shifted to give a causal impulse response. The corresponding (ideally) rectangular transfer function is a lowpass filter with a sharp cutoff in the Bode frequency plot.

**Pole-Zero Diagram** The pole-zero diagram in Figure 9.20 illustrates several basic properties of FIR filters. The only poles lie at the origin, and there are multiple zeros (ten) corresponding to the order of the filter. The region of convergence includes the entire *z*-plane except $z = 0$, and because all the poles are all within the unit circle the system is inherently stable. The three zeros lying directly on the positive half unit circle correspond to the zeros in the Bode magnitude plot of Figure 9.19. The response function $h[n]$ defined here has 11 terms, leading to 10 zeros on the pole-zero diagram, as well as 10 poles located at $z = 0$. In general for an FIR response function $h[n]$ with $N$ terms, there will be $N - 1$ poles and $N - 1$ zeros.

**Phase Response** The Bode phase plot of Figure 9.21 shows a smooth phase change with frequency with the exception of three sudden phase changes of $\pm\pi$ radians at the frequency of each of the three zeros. The overall phase change for this FIR filter is linear with frequency, although this is not immediately obvious since the frequency axis uses a logarithmic scale. It has been shown that a linear phase change reflects a simple time delay, where the phase shift $\Phi$ at a given frequency corresponding to a delay of $t_0$ seconds is given by $\Phi = -2\pi t_0 f = -\omega t_0$ radians. The slope of the *phase* vs. *frequency* graph or the change in phase with respect to $\omega$ is then $\Delta\Phi = -t_0$.

In this case, the delay is introduced by passing through the FIR filter. Since the response function $h[n]$ is an even sinc function shifted to sample $n = 5$ as $sinc(n - 5)$

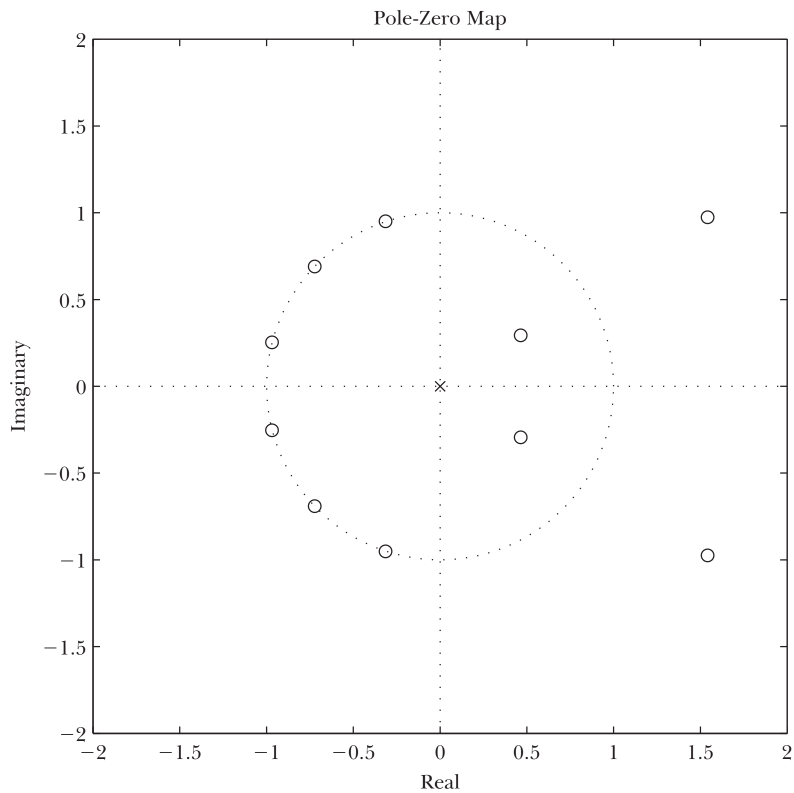



**FIGURE 9.20 A High-Order FIR Filter—Poles and Zeros** The only poles in an FIR filter are found at the origin. Multiple poles (all at $z = 0$) and multiple zeros characterize a high-order FIR filter.

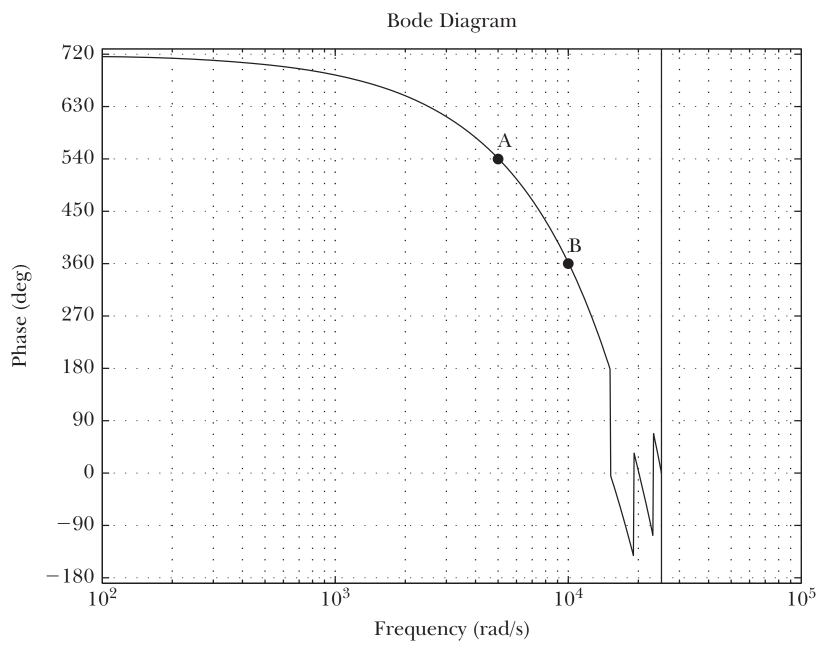



**FIGURE 9.21 A High-Order FIR Filter—Linear Phase** The overall phase change is linear for this FIR filter. Sudden phase changes of $\pi$ radians occur at the frequency of each of the three zeros.

and the sample rate is $f_s = 8000$ Hz, the actual delay is $t_0 = 5/8000 = 625\ \mu$s, giving the slope:

$$\Delta\Phi = -t_0 = \frac{5}{8000} = 6.25 \times 10^{-4} \tag{9.24}$$

The actual slope of this linear relationship may be determined directly from the Bode phase plot. For example, consider the points $A$ and $B$ shown highlighted. At point $A$, the frequency is 5000 rad/s and the phase is 540 degrees or $3\pi$ radians. At point $B$, the frequency is 10,000 rad/s and the phase is 360 degrees or $2\pi$ rad. The measured slope of the line is:

$$\Delta\Phi = \frac{3\pi - 2\pi}{5000 - 10000} = \frac{-\pi}{5000} \approx -6.28 \times 10^{-4} \tag{9.25}$$

which is consistent with Eqn. 9.24. Since the slope is constant, the same result should be found for any choice of two points on the graph. A linear phase change would be expected whenever $h[n]$ takes the form of an even function shifted as required to assure a causal impulse response. such as the sinc function used here.

**Step Response** The unit step response in Figure 9.22 illustrates the system delay $t_0$ = 625 $\mu$s as found above. The maximum value of the step response corresponds to the sum of the terms in $h[n]$ found in MATLAB using `max = sum(num)`, giving $max = 2.10$.

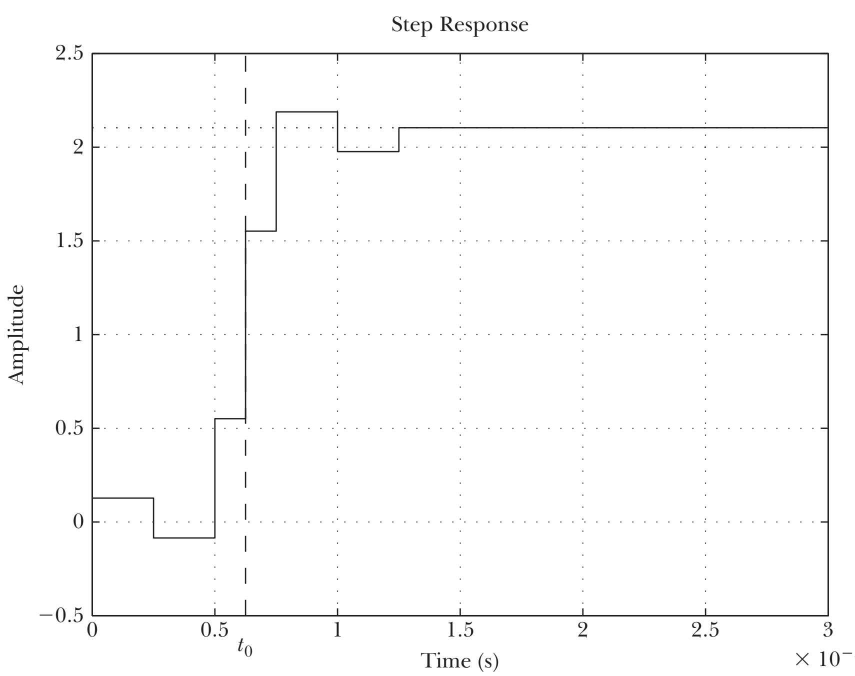



**FIGURE 9.22 A High-Order FIR Filter—Step Response** The discrete unit step response illustrates the system delay $t_0 = 0.625 \times 10^{-3}$ s. The maximum value of the step response corresponds to the sum of the terms in $h[n]$.

# 9.10 Digital Filtering—IIR Filter

The $z$-transform has been shown to be a powerful tool for the design of digital filters, as the shape of the frequency response of a filter can be tailored according to the choice of terms $a_n$ in the $z$-domain impulse response $H(z)$. The FIR filter is inherently stable, relatively simple to construct, and can offer a linear phase response. Still, a large number of terms may be required, and the delay associated with higher-degree polynomials in $z$ can prove to be troublesome.

The Infinite infinite impulse response (IIR) filter is an alternative solution to signal filtering that may not be stable yet offers some attractive advantages in construction and delay characteristics.

## 9.10.1 A One-Pole IIR Filter

Consider the system of Figure 9.23 in which the delay block output is fed back to the input.

Because of the feedback block, an impulse input $\delta[n]$ may lead to outputs that persist for a very long time. For example, if an impulse $\delta[n]$ arrives at the input of this system, it is immediately output and also fed to the delay block. After a delay, the impulse is again seen at $y[n]$ as $a\delta[n]$ and is again fed to the delay block. After another delay, the impulse remerges at $y[n]$ as $a^2\delta[n]$ and once more is fed to the delay block. After a delay, the impulse remerges at $y[n]$ as $a^3\delta[n]$ and is fed to the delay block. This

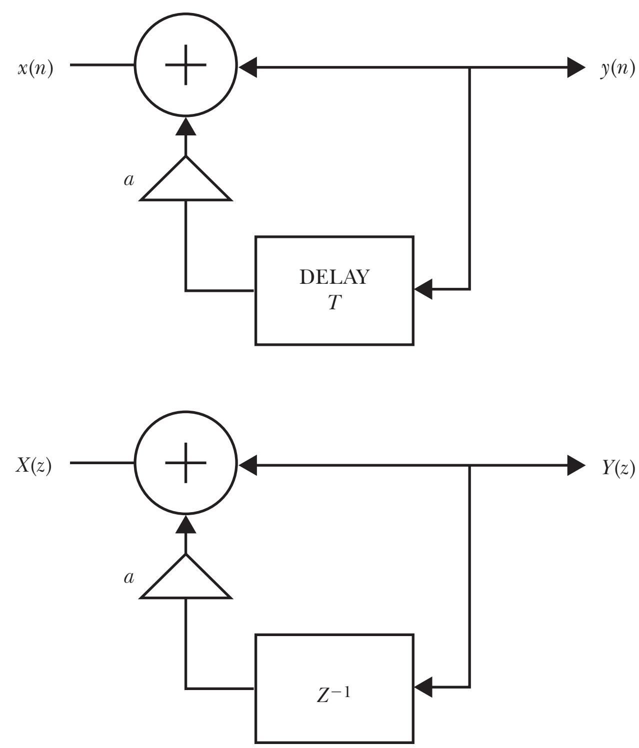



**FIGURE 9.23 An IIR Filter** A feedback circuit uses delay blocks that may be conveniently relabelled as $z^{-1}$ to show its $z$-transform.

process continues and characterizes the concept of *infinite impulse response*. By inspection, the overall impulse response $h[n]$ is:

$$h[n] = 1 + a + a^2 + a^3 + a^4 + \ldots + a^\infty = \sum_{n=0}^{\infty} a^n \tag{9.26}$$

In practice, the output terms would go to zero over time provided that $a < 1$, and for most IIR systems the response is not expected to go on forever. The above analysis assumes that the system began at rest and there were no non-zero initial conditions in any of the delay blocks.

The corresponding *z*-domain system has input $X(z)$ and output $Y(z)$ as shown in Figure 9.23. The system may be analyzed to find the transfer function $H(z)$ from the output $Y(z)$ and input $X(z)$ as:

$$Y(z) = X(z) + az^{-1}Y(z) \tag{9.27}$$

$$Y(z)[1 - az^{-1}] = X(z) \tag{9.28}$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{1}{1 - az^{-1}} = \frac{z}{z - a} \tag{9.29}$$

which has a pole at $z = a$ and a zero at $z = 0$. This $H(z)$ is exactly the *z*-transform of Eqn. 9.26 and the same transfer function that was derived from the first-order lowpass filter in Section 9.8. In that RC-circuit $a = 1/RC$, ensuring that $a < 1$ and the single pole is necessarily found inside the unit circle as in Figure 9.12. Compared to the single-pole FIR filter where the pole is at the origin and the zero can be moved, the single-pole IIR filter has the zero at the origin while the pole can be moved depending on the choice of $a$.

To explore the effect of varying the position of the pole (which always lies on the real axis), a short MATLAB sequence can be easily modified to place the single pole as desired. In Figure 9.24, five different positions are shown as $(A, B, C, D, E) = (-0.9, -0.5, -0.1, +0.8, +1.5)$.

```
H = tf([1 0 ], [1  -0.5 ], 1) % pole at z =  +0.5
bode(H);
grid on;
```

From the figure, poles near the origin $(C)$ have very little effect on an input signal, while poles close to the unit circle $(A)$ have the greatest effect. Any pole directly on the unit circle would force the frequency response to infinity at that point. In general, negative pole positions $(A, B)$ produce a highpass filter, while positive positions $(C, D, E)$ achieve a lowpass filter. By varying the $a$ in $H(z)$ the filter characteristics may be adjusted as desired; however, *the pole position* E *results in an unstable system*, and special care must be taken when IIR filters are designed or modified. For any IIR filter, the zero(s) will lie on the origin, while the poles serve to determine the form of the transfer function.

## 9.10.2 IIR versus FIR

Given the above IIR impulse response $h[n]$ and assuming that $a < 1$ to limit the total number of significant terms, an equivalent FIR filter can be constructed directly from the terms in Eqn. 9.26. Consider the RC-circuit simulation of Section 9.8, in which a value of $a = 0.939$ was used with a sampling rate of 1000 Hz. After 100 terms, the value

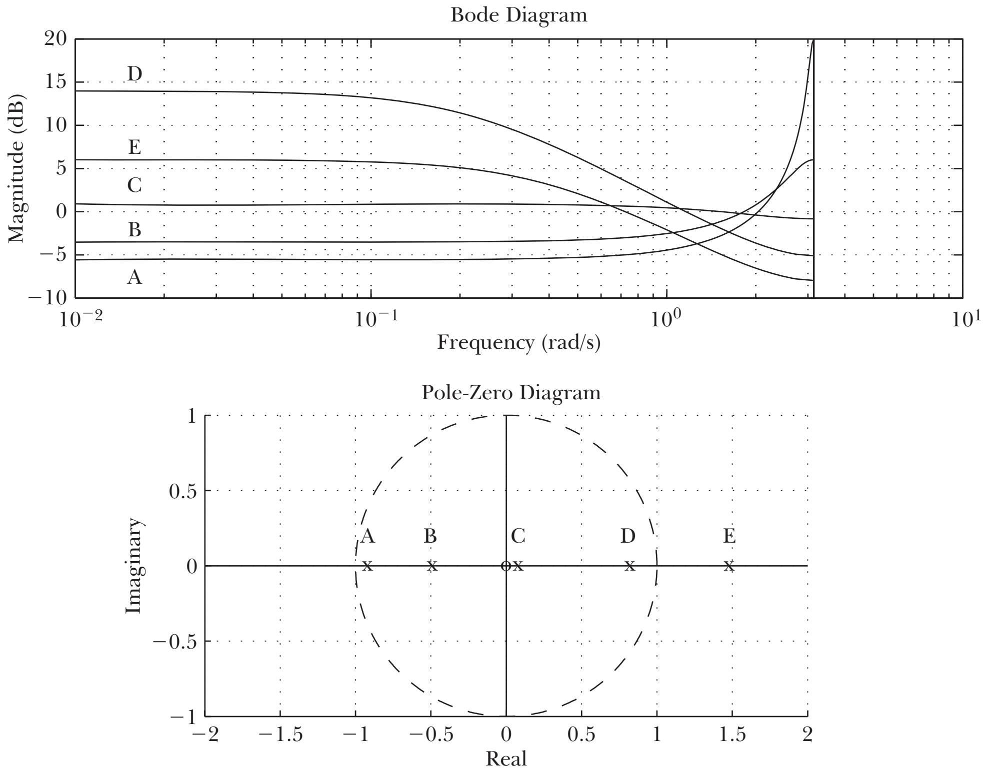



**FIGURE 9.24 An IIR Filter—One Pole** A single-pole IIR filter has a single zero at the origin and a single pole that may move along the real axis. This figure shows the effect on the transfer function of moving the pole to one of five positions (A, B, C, D, E). Position $E$ outside the unit circle represents an unstable system. Compare to Figure 9.18 for the FIR filter.

$a^{100} = (0.939)^{100} = 0.0018$, which is nearing zero such that further terms will have little contribution and the FIR filter can be defined in MATLAB and compared to the IIR filter as:

```
n = 0:100;                          % choose 100 terms of h[n]
a = 0.939;                          % specify the value of a
num = a .^ n;                       % define FIR terms
den = [1 zeros(1,99) ];             % define FIR denominator
H1 = tf(num, den, 1/1000);          % H1= FIR
H1 = tf([1 0], [1 -a], 1/1000);     % H2= IIR
bode(H1,H2);                        % show FIR & IIR, same Bode plot
pzmap(H1)
```

In this example, the IIR filter is defined from $H(z)$, while the FIR filter is defined from $h[n]$. The Bode plot in Figure 9.25 shown both the IIR and FIR transfer functions superimposed and the two curves are indistinguishable. The pole-zero diagram in Figure 9.26 features a remarkable 100 zeros found around a circle of radius $a$ and includes 100 poles all on the origin $z = 0$. The lack of a zero near $\omega = 0$ implies that frequencies near zero pass through this filter, while higher frequencies are attenuated.

In effect, the single-pole IIR filter has accomplished the same response function as a 100 (or more) pole FIR filter. For this reason, use of the IIR filter is often the

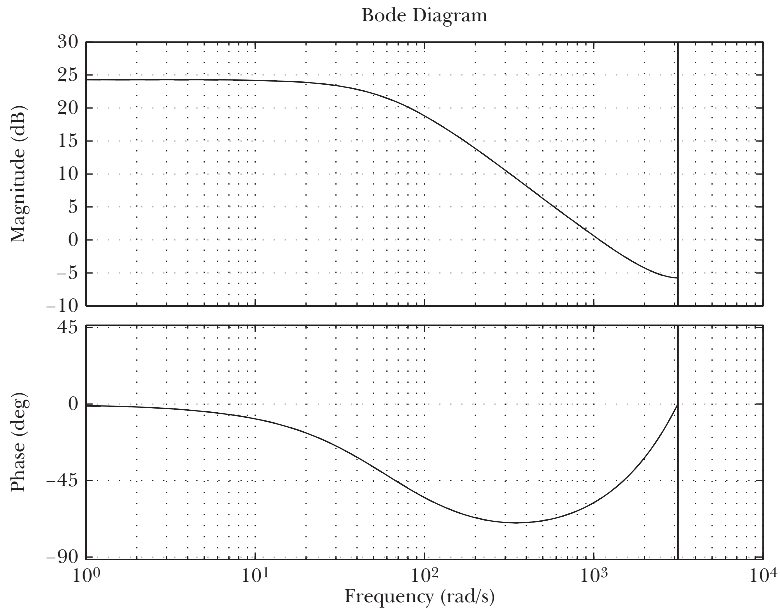



**FIGURE 9.25 An Equivalent FIR Filter—Bode Plot** The transfer function of the single-pole IIR filter of Figure 9.23 is closely matched by a 100-pole FIR filter in this Bode plot showing both transfer functions superimposed and indistinguishable ($a = 0.939$).

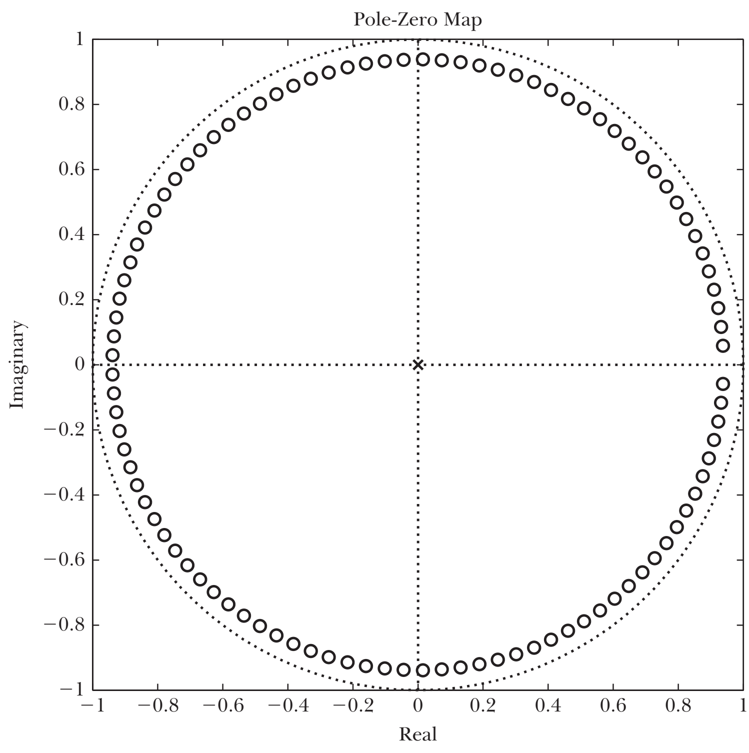



**FIGURE 9.26 An Equivalent FIR Filter—Pole-Zero** The transfer function of the single-pole IIR filter of Figure 9.23 is closely matched by this 100 pole FIR filter ($a = 0.939$). The 100 zeros describe a circle of radius $a$.

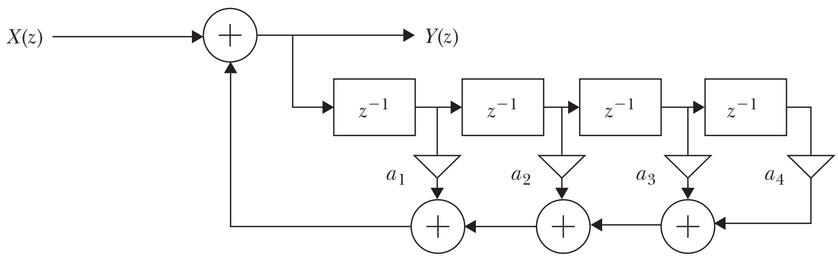



**FIGURE 9.27 A General IIR Filter** A feedback circuit is relabelled as $z^{-1}$ to show its $z$-transform.

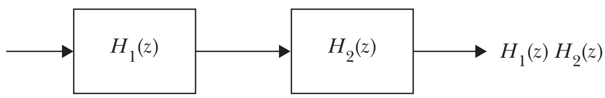



**FIGURE 9.28 Combined Filters** Cascaded filters (FIR and/or IIR) can be used for improved filter performance.

preferred approach, especially where linear phase response is not a issue. In this example, the FIR filter would not have a linear phase response (the exponential function is not even), and in general an IIR filter will not have a linear phase response.

## 9.10.3 Higher-Order IIR Filters

Higher-order IIR filters can be created following the system block diagram of Figure 9.27. For the four terms shown:

$$Y(z) = X(z) + Y(z)[a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + a_4 z^{-4}] \tag{9.30}$$

$$X(z) = Y(z)[1 - a_1 z^{-1} - a_2 z^{-2} - a_3 z^{-3} - a_4 z^{-4}] \tag{9.31}$$

$$H(z) = \frac{X(z)}{Y(z)} = \frac{z^4}{z^4 - a_1 z^3 - a_2 z^2 - a_3 z^1 - a_4} \tag{9.32}$$

This system has four zeros at the origin and four poles at the roots of the denominator. The appearance of the transfer function $H(z)$ of IIR filters resembles that of FIR filters *with the numerator and denominator switched.* This implies that an $N$-pole IIR filter has $N$ zeros at the origin and $N$ poles corresponding to the roots of the numerator (the opposite appearance of an FIR filter). This is an important difference as the poles of an IIR system (but not of an FIR system) can be located outside the unit circle where the system would no longer be stable. When an IIR filter is designed or modified, it is especially important to check for stability. In contrast, an FIR filter will have poles only at $z = 0$ where stability is assured.

## 9.10.4 Combining FIR and IIR Filters

More complex filters can be defined by cascading different systems, including both FIR and IIR filters as required. From the convolution property of the $z$-transform, the overall effect of the transfer functions $H_1(z)$ followed by $H_2(z)$ is the product $H_1(z)H_2(z)$. The order of the systems is unimportant as the product operation is commutative.

## 9.11 Conclusions

The z-transform completes the triad of transforms that underpin modern signal processing. The z-transform embodies elements of the Fourier transform and the Laplace transform when dealing with discrete signals. The z-transform is especially convenient as it can be found directly in the terms defining a delay block system. Digital filter design is an important application of this versatile transform. The system block diagrams and the characteristics of FIR and IIR filters have been studied and comparative examples of both have been examined.

# End-of-Chapter Exercises

**9.1** Prove the infinite series result shown in Eqn. 9.5 or:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad \text{for} \quad |x| < 1 \tag{9.33}$$

Confirm the validity of the approximation using only the first 10 terms if:

**(a)** $x = 0.5$

**(b)** $x = -0.5$

**(c)** $x = 3$ (violates the condition for valid $x$)

**9.2** Derive the z-transform of $\sin(n\omega_0)$ as found in Eqn. 9.11.

**9.3** Consider the discrete step function $u[n]$:

**(a)** Sketch the signal $n\ u[n]$ and describe its appearance.

**(b)** Determine the z-transform of $n\ u[n]$ in terms of $u[n]$.

**9.4** Consider the discrete step function $u[n]$:

**(a)** Sketch the signal $u[n-1]$ and describe its appearance.

**(b)** Use the derivative rule to find the z-transform of $u[n-1]$.

**9.5** Consider the z-transform of $\cos(n\omega_0)$ shown below and in Eqn. 9.10.

$$\cos(n\omega_0)\ u(n) \overset{z}{\leftrightarrow} \frac{z^2 - z\cos(\omega_0)}{z^2 - 2z\cos(\omega_0) + 1}$$

$$\text{ROC} = |z| > 1$$

**(a)** Find the roots of the denominator.

**(b)** Show that the poles lie on the unit circle.

**(c)** Accurately sketch the pole-zero diagram for $\omega_0 = \pi/4$.

**9.6** Determine the z-transform of $\cos(200\pi t)$ for a sample rate of 1500 Hz and sketch the pole-zero diagram.

**9.7** Determine the z-transform of $\sin(200\pi t)$ for a sample rate of 1500 Hz and sketch the pole-zero diagram.

**9.8** Determine the z-transform of $\cos(200\pi t)$ for a sample rate of 2000 Hz and sketch the pole-zero diagram.

**9.9** Determine the z-transform of $\sin(200\pi t)$ for a sample rate of 2000 Hz and sketch the pole-zero diagram.

**9.10** Prove the *initial value theorem* shown below, where $X(z)$ is the one-sided z-transform of $x[n]$:

$$x[0] = \lim_{z \to \infty} X(z)$$

**9.11** Consider the transfer function:

$$H(z) = 2 - \frac{1}{z} + \frac{3}{z^3}$$

**(a)** Accurately sketch the pole-zero diagram.

**(b)** Is this system stable?

**9.12** Consider the transfer function:

$$H(z) = \frac{2z(z-1)}{4z^2 + 4z + 1}$$

**(a)** Accurately sketch the pole-zero diagram.

**(b)** Is this system stable?

**9.13** Determine the impulse response for the system defined by:

$$H(z) = \frac{1)}{4z^2 - 4z + 1}$$

**9.14** Consider the transfer function:

$$H(z) = \frac{2z(z-1)}{4z^2 - 4z + 1}$$

(a) Use MATLAB to define this system as `sys`.
(b) Plot the impulse response of `sys`.
(c) Plot the step response of `sys`.
(d) Plot the pole-zero diagram of `sys`.

**9.15** Determine the DTFT corresponding to:

$$H(z) = \frac{2z(z-1)}{4z^2 - 4z + 1}$$

**9.16** Consider the pole-zero diagram $S[z]$ of Figure 9.29.

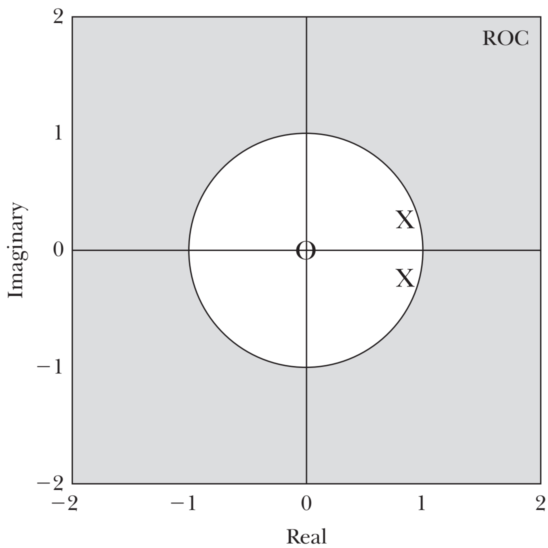



**FIGURE 9.29** Figure for Questions 9.16 to 9.20.

(a) Describe the appearance of the time-domain impulse response $s[n]$.
(b) Is this system stable? Explain.

**9.17** Consider $H[z]$ in the pole-zero diagram of Figure 9.29. For this FIR filter, how many zeros are present at the origin?

**9.18** Consider $H[z]$ in the pole-zero diagram of Figure 9.29. Describe how this diagram would change if the sampling rate is increased.

**9.19** Consider $H[z]$ in the pole-zero diagram of Figure 9.29. Describe how this diagram would change if the filter cutoff frequency is increased.

**9.20** Consider $H[z]$ in the pole-zero diagram of Figure 9.29. Sketch the corresponding pole-zero diagram for the Laplace transform *H(s)*.

**9.21** Repeat the MATLAB example of Section 9.8 for a cutoff frequency $f_c = 500$ Hz and $f_s = 20{,}000$ Hz. Identify the cutoff frequency on the Bode plot and confirm that it equals $f_c$.

**9.22** In Figure 9.24 the pole at position $E$ was described as *unstable*, yet it gives a reasonable frequency response on the Bode plot. Explain why this pole is considered unstable, and use MATLAB to find the impulse response and step response for the system if the single pole is at position $E$.

**9.23** Use MATLAB to create the Bode plot for a single-pole FIR filter with a pole at $z = 0.5$. What type of filter is this, and what is the cutoff frequency for this filter? Assume a sample rate of 1000 Hz.

**9.24** Use MATLAB to create the Bode plot for a single-pole IIR filter with a pole at $z = 0.5$. What type of filter is this, and what is the cutoff frequency for this filter? Assume a sample rate of 1000 Hz.

**9.25** Consider the pole-zero diagram of Figure 9.30, which is an IIR or FIR filter.

(a) Identify the filter type IIR or FIR.
(b) How many zeros are located at the origin?
(c) Is the system stable? Explain.
(d) The poles are located at $z = 0.25 \pm 0.66j$. Write the simplified transfer function $H(z)$.
(e) Use MATLAB to show the Bode diagram.
(f) Determine the resonant frequency of this system. Assume a sample rate of 10,000 Hz.

**9.26** Consider the pole-zero diagram of Figure 9.31, which is the transfer function of an IIR or FIR filter.

(a) Identify the filter type IIR or FIR.
(b) How many poles are located at the origin?
(c) Is the system stable? Explain.
(d) At what frequency is there a null in the frequency response? Assume a sample rate of 8000 Hz.

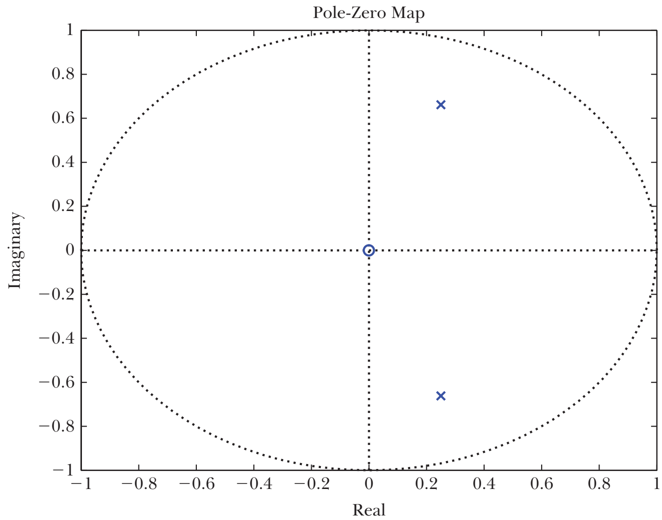



**FIGURE 9.30** Figure for Question 9.25.

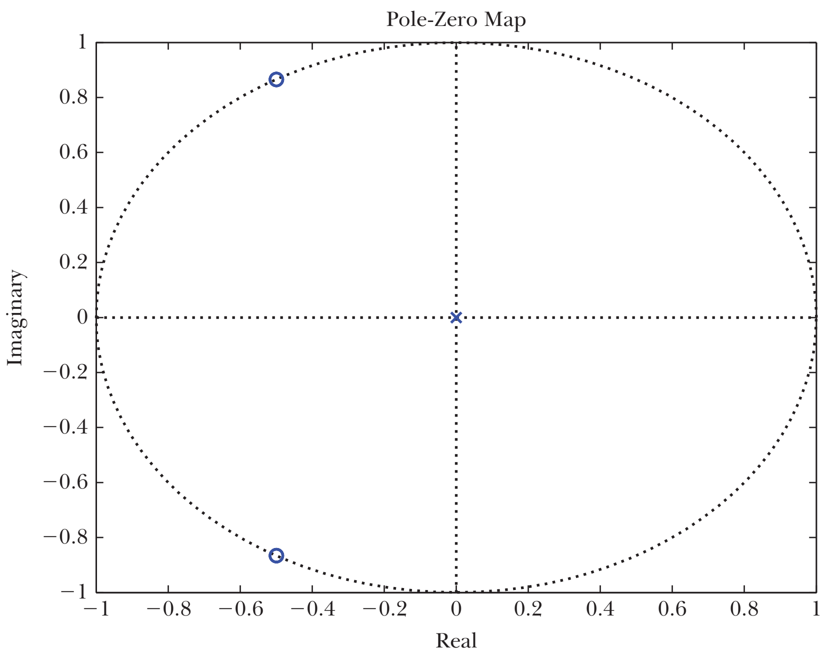



**FIGURE 9.31** Figure for Question 9.26.

CHAPTER 10

# Introduction to Communications

In this chapter, the reader will find examples and applications that demonstrate the pivotal role of signals theory in the world of communications systems. The chapter explores a selection of applications of signals theory to the transmission and reception of information, particularly by radio. Various approaches to modulation and demodulation are presented from a signals perspective. Practical problems associated with the design and operation of radio receivers also find elegant solutions through the theory of signals and systems that has been elaborated in earlier chapters.

LEARNING OBJECTIVES

By the end of this chapter, the reader will be able to:

- Describe the desirable properties of a carrier signal
- Identify bands in the radio spectrum by frequency and wavelength
- Explain the concepts of amplitude, frequency, or phase modulation
- List different forms of amplitude modulation (AM)
- Apply signals theory to relate message bandwidth to transmitted bandwidth
- Explain receiver sensitivity and selectivity
- Describe the operation of a superheterodyne receiver
- List different types of digital modulation methods
- Create digital modulation schemes from analog principles

## 10.1 Introduction

Almost every day it seems that some new communication device is enhancing our lives. In recent years, an astonishing variety of new communications services have moved from the laboratory to become commonplace items. Cellular and cordless telephones, wired and wireless computer networks, satellite radios, and a variety of radio-controlled devices come to mind immediately. All of these developments were made possible through advances in microelectronics and digital computers, yet underlying their practical implementation are sound applications of signals theory.

### 10.1.1 A Baseband Signal *m*(*t*)

The ultimate purpose of any communication system is to convey a message from one place to another. Any such message is called a *baseband signal*, which will be represented by the signal $m(t)$. Frequently, the baseband or message signal $m(t)$ is voice or music, which is bandlimited to the audio frequencies that the human ear can


**FIGURE 10.1 A Communications System** A message travels from one place to another through a communications *channel*.

detect, generally less than 20 kHz. Another message $m(t)$ may be a television picture with sound, in which case the message bandwidth may be several megahertz. In our modern world, a growing number of messages $m(t)$ are conveyed in digital form as a series of bits whose overall bandwidth varies with the speed of transmission. In every case, the baseband message is the original signal to be transmitted to another location and that ideally, should be exactly reproduced upon reception. As shown in Figure 10.1, between the sender and the recipient, the message signal passes through what is generically called a communications *channel*. Not to be confused with the channel number on a radio or television receiver, a communications channel is merely the path over which a transmitted signal must travel. A communications channel may be a direct wire connection, a radio link, a fiber optic cable, or any other means of conveying the baseband message signal from one place to another.

## 10.1.2 The Need for a Carrier Signal

Consider a voice message $m(t)$. This signal cannot travel very far in its original form as a baseband signal. Even with a powerful amplifier and loudspeakers, a voice message could not travel through the air more than a few hundred meters. Moreover, if local broadcasters used loudspeakers instead of radio frequency transmitters, there would be no way to *tune* or to select different stations and, consequently, no way that several local stations could broadcast simultaneously. Instead, something else is required to transport the message over any appreciable distance. This *carrier signal* will have the property that it can travel easily through a communications channel and carry with it the message signal.

For broadcast radio signals, the channel is free space, and the message is carried by a radio frequency sinusoid. The observation that at certain radio frequencies (RF), electromagnetic waves can propagate or travel over long distances prompts the idea that RF signals can be used to carry a message over those same distances. Used in this manner, sinusoidal signals of an appropriate frequency are called carriers, and form the basis for many communications systems. Different carrier signals will be appropriate for different channels. For example, optical carrier signals may be used for a fiber optic channel and audio frequency carriers are used in a dial-up modem to carry digital signals over a telephone channel.

## 10.1.3 A Carrier Signal *c(t)*

Define a carrier signal as $c(t) = A\cos(2\pi f_c t + \Phi)$, to where $f_c$ is a suitable frequency such that this electromagnetic wave can propagate through a given communications channel. The choice of $f_c$ ultimately depends on the channel characteristics, and carrier frequencies can vary from a few tens of hertz to several gigahertz.

Alone, the signal $c(t) = A\cos(2\pi f_c t + \Phi)$ is simply a sinusoid. It conveys no information, beyond the fixed parameters $(A, f_c, \Phi)$ that define it. If this signal was generated, amplified, and sent to an antenna, it would propagate through space, where it could be detected by a radio receiver, perhaps on the other side of the world. As such, it would be no more than a crude radio beacon, advertising its presence by its constancy and lack of variation.

Consider what would be observed at a far distant receiver tuned to the above signal. The carrier amplitude would be much attenuated after its long voyage and the absolute phase could not be established at all without some fixed reference point in time. Only the carrier frequency or any changes in phase or amplitude could be measured with any reliability by the receiving apparatus. So, while the carrier $c(t)$ alone exhibits the desired property of being able to traverse long distances, it is not especially useful by itself.

## 10.1.4 Modulation Techniques

The process of imposing a message $m(t)$ onto a sinusoidal carrier $c(t)$ is called *modulation*. To be useful, any modulation scheme must allow easy *demodulation*, the purpose of which is to recover faithfully the original message once it has been carried to its destination.

As usual, there are only three parameters $(A, f_c, \Phi)$ in the sinusoidal carrier signal $c(t)$ that can be varied as a means of modulating a message. It follows that the message $m(t)$ may be used to alter the amplitude, frequency, or phase of the carrier. This observation leads to methods generically known as *amplitude modulation* (AM), *frequency modulation* (FM), and *phase modulation* (PM), respectively. More complex modulation schemes may use various combinations of these three basic methods. The properties of the resulting modulated signal depend on the carrier signal $c(t)$, the message signal $m(t)$, and the modulation method that is used. The modulated result is the signal that is expected to pass through a communications channel, where it may be attenuated, distorted, or affected by the presence of other signals before reaching its destination. At the far end of the channel, a receiver is expected to reconstruct the original message $m(t)$ from this modulated signal as seen in Figure 10.2. Techniques for modulating and demodulating such signals are to be analyzed in this chapter, but first the question of choosing a carrier frequency will be addressed.

## 10.1.5 The Radio Spectrum

Before any message can be transmitted, a carrier frequency must be chosen. The choice of carrier frequency will affect how far and how well the message will travel, and what kind of equipment will be necessary to generate and to receive the modulated signal. Radio signals are generally classified according to either carrier frequency or wavelength. The Greek letter lambda $(\lambda)$ is generally used to indicate wavelength, where the frequency and wavelength of a radio signal are related by the speed of light $(c = 3 \times 10^8\ \text{m/s})$ where:

$$\lambda\ m = \frac{3 \times 10^8\ \text{m/s}}{f\ \text{Hz}}$$


**FIGURE 10.2 A Communication System** A baseband message $m(t)$ must generally be *modulated* to pass efficiently over a channel and then *demodulated* to restore the original message.

TABLE 10.1

**Radio Spectrum Frequency Bands**

| Frequency | Wavelength | Description |
|---|---|---|
| 3–30 kHz | 100–10 km | VLF = Very Low Frequency |
| 30–300 kHz | 10–1 km | LF = Low Frequency |
| 300–3000 kHz | 1000–100 m | MF = Medium Frequency |
| 3–30 MHz | 100–10 m | HF = High Frequency |
| 30–300 MHz | 10–1 m | VHF = Very High Frequency |
| 300–3000 MHz | 100–10 cm | UHF = Ultra High Frequency |
| 3–30 GHz | 10–1 cm | SHF = Super High Frequency |
| 30–300 GHz | 10–1 mm | EHF = Extremely High Frequency |

For radio waves, it is convenient to express frequencies in MHz, leaving a simple formula relating frequency in hertz to wavelength in meters: Any wavelength value can be converted to frequency units and vice versa using this formula. For example, the carrier frequency of an FM broadcast station at 100.5 MHz has a wavelength of about 3 meters.

The set of radio frequencies of all wavelengths describes the frequency spectrum. For convenience, the radio spectrum is usually divided into bands at wavelength multiples of 10. When these same divisions are expressed in frequency, a factor of three appears from application of the above formula. Finally, each band can be named according to its characteristic frequency or wavelength. These divisions and their usual names appear in Table 10.1. Note that since wavelength is inversely proportional to frequency, wavelengths become shorter as frequencies become higher. As frequencies go higher, each band has ten times the width in frequency of the previous band.

It is important to appreciate that any and all radio signals must lie somewhere within these bands; *every radio signal for any use and in any part of the world must fit here somewhere.* In every country, government legislation governs the use and users of the radio spectrum. In the United States, use is governed by the Federal Communications Commission (FCC), while Industry Canada (IC) enforces the sharing and use of the radio spectrum in Canada. Because the radio spectrum is both a valuable and a limited resource, this regulation is very important. For instance, no one would want a local television station or radio paging system interfering with life-saving ambulance or fire services. Where frequencies can propagate over long distances, the possibility of interference requires national or international regulation. Many radio signals can readily travel long distances and across national boundaries, so such rules must apply throughout the world. For example, aircraft flying on international routes expect to use the same communications equipment in every country they visit. Likewise, amateur radio operators in different countries must all make use of the same frequencies, or they would not be able to talk to each other. Generally, users of the radio spectrum must obtain a government-issued license, and, as seen in Figure 10.3, all must be aware of the potential for interference posed to other users. Furthermore, the rules that allocate the use of specific frequencies to different groups of users may also place limits on the maximum power transmitted, the modulation to be used, the bandwidth of a transmitted signal, or the types of radio services allowed in each band.


**FIGURE 10.3 A Radio Communication System** Transmitters and receivers complete the path that may carry the message $m(t)$ over very long distances.

Having chosen an appropriate carrier frequency, it is now possible to study the subject of modulating a message onto that carrier. In doing so, it is important to address not only the mathematics of modulation and demodulation, but also the overall characteristics of the modulated (transmitted) signal, such as the relative power of components, signal bandwidth, and harmonic content. The most effective study of these signals requires their examination in both the time and frequency domains.

Not all communications channels use radio frequency signals. For example, a telephone modem used with a computer must modulate a carrier at voice frequencies (less than 4 kHz) to transmit digital signals over a standard telephone line. Similarly, not all radio frequency carrier signals need to follow government spectrum allocations. For example, radio signals sent through a shielded channel such as a computer network or a cable television system cannot interfere directly with others, and the issue of spectrum sharing is largely irrelevant. (The issue becomes whether or not the cable is adequately shielded.) The principles of modulation and demodulation apply to all carrier signals and to all manner of communications channels.

## 10.2 Amplitude Modulation

In amplitude modulation (AM), the message $m(t)$ to be transmitted is imposed upon the amplitude of a carrier signal. There are a number of variations of amplitude modulation, each of which applies the message signal to the carrier amplitude in a slightly different way. The most common approach, although not necessarily the most efficient, will be discussed first. This is amplitude modulation as it is employed by broadcast stations throughout the world. This is the same modulated signal received and demodulated by a common pocket *AM receiver*.

### 10.2.1 Transmitted Carrier Double Sideband—(AM-TCDSB)

Consider the carrier signal $c(t) = A\cos(2\pi f_c t)$, where $f_c$ is to be chosen and assert a message $m(t)$ onto this carrier by replacing the constant amplitude term with $A = [1 + Km(t)]$, where $K$ is a positive constant less than 1 and $m(t)$ is bounded to the range $[-1, +1]$. The reasoning behind both these restrictions will be studied later. A suitable test message signal might be a simple cosine $m(t) = \cos(2\pi f_m t)$, where $f_m$ is to be chosen. This message is a constant tone, but it may represent a single-frequency component in a more complex signal such as voice or music.

The resulting modulated signal (the signal to be transmitted) is now given by $b(t) = [1 + Km(t)]\cos(2\pi f_c t)$. Note that only the amplitude of the carrier is affected and the frequency and phase are unchanged. In practice, the message frequency $f_m$ might be 1000 Hz (a voice band signal), and the carrier frequency $f_c$ might be 970 kHz (in the North American broadcast band). Recall that this is exactly the modulation method used by local broadcast stations to transmit voice and music in the band between 540 and 1600 kHz. In summary:

1. Message: $m(t) = \cos(2\pi f_m t)$
2. Carrier: $c(t) = \cos(2\pi f_c t)$
3. Modulated signal: $b(t) = [1 + Km(t)]\cos(2\pi f_c t)$

The modulated signal $b(t)$ is expected to propagate through a channel to a distant receiver and, in doing so, to carry the message $m(t)$. If this signal is to be transmitted effectively, the frequency components present in the modulated signal $b(t)$ should not differ dramatically from those of the (unmodulated) carrier frequency, *which was chosen specifically because of its ability to travel through the channel.* To be useful, a receiver must be able to extract (demodulate) the baseband signal $m(t)$ from the received signal $b(t)$. All these observations may now be explored in detail.

For simplicity, set $K = 1$ for the present time. As with any other signals analysis, the process of amplitude modulation and demodulation is best studied in both the time domain and the frequency domain. In analyzing the modulated signal, it is first necessary to construct the signal $[1 + Km(t)]$. By inspection, with $K = 1$, this specific signal $[1 + cos(2\pi f_m t)]$ is a cosine with a DC offset, as shown in Figure 10.4 in the time and frequency domains. The process of modulation and transmission will be successful if either of the signals in the figure can be obtained after reception and demodulation. The original message will have been recovered.

The modulated signal $b(t)$ is formed by multiplying (in the time domain) the signal in Figure 10.4 by the carrier $c(t) = \cos(2\pi f_c t)$ as shown in Figure 10.5. The resulting modulated signal is shown in Figure 10.5C. Note that as the amplitude of a cosine is varied in the time domain, its peak-to-peak excursions vary symmetrically. The overall form of the amplitude variations describes the *envelope* of the modulated signal; this signal has a sinusoidal envelope varying according to the message $m(t)$. This envelope is a reflection of the fact that the carrier amplitude is now conveying the cosine message $m(t)$.[1] Multiplication in time corresponds to convolution in the frequency domain. The resulting modulated signal $[1 + m(t)]c(t)$ looks like the spectrum shown in Figure 10.6.

Note that in Figure 10.6, the components of the modulated signal $m(t)$ are concentrated about the carrier frequency as expected. There are no baseband components near zero frequency. It is at the carrier frequency that a receiver must be *tuned* to capture this transmitted signal. The goal of carrying the message by

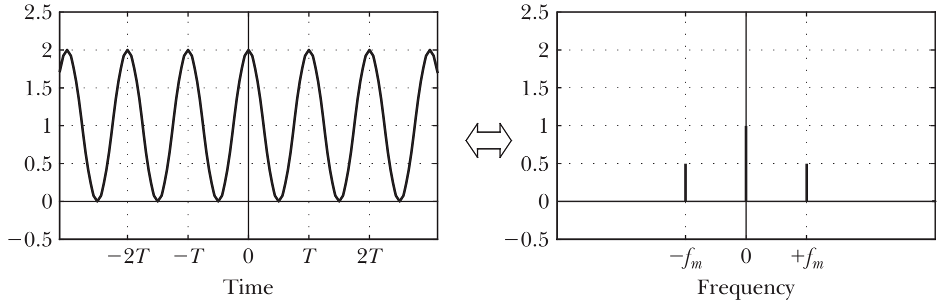



**FIGURE 10.4** The signal $[1 + m(t)] = [1 + cos(2\pi f_m t)]$ in time and frequency.

[1] The envelope is sketched here for clarity but does not normally appear superimposed on the signal.

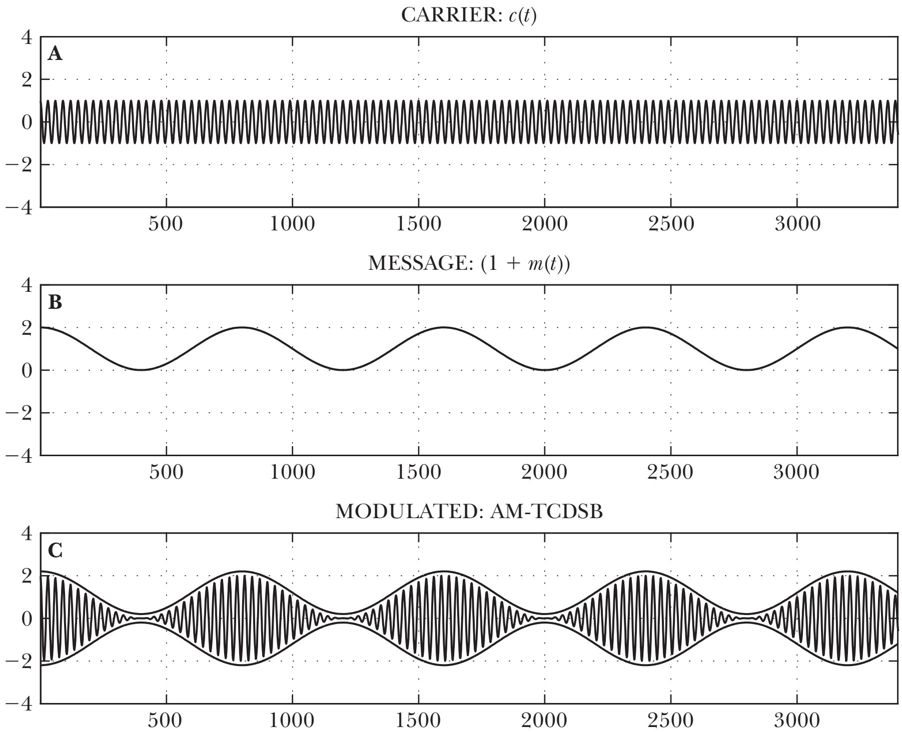



**FIGURE 10.5 Amplitude Modulation (AM-TCDSB)** The carrier signal in **A** is multiplied by the message **B** give the signal in **C**. This final signal *carries* the message content in its visible amplitude variations.

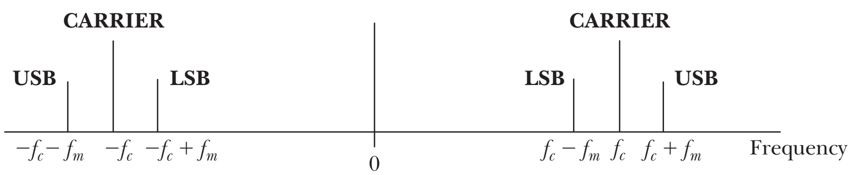



**FIGURE 10.6 Amplitude Modulation (AM-TCDSB)** A cosine message signal $m(t) = \cos(2\pi f_m t)$ is visible modulated on the carrier frequency $f_c$ as two *sidebands* labelled the upper sideband (USB) and lower sideband (LSB). This is the same signal seen in Figure 10.5C.

translating the baseband signal into something that will readily pass through the radio channel has been accomplished.

The mixer output is the sum and difference of the cosine message and cosine carrier frequencies.[2] From the close-up view of the region near the carrier frequency, three distinct components can be identified. A matched set of Fourier components lies on the negative frequency axis. The largest component lies at the carrier frequency. Symmetrically on either side of the carrier frequency lie two sidebands, called the *upper sideband* and *lower sideband*. The sidebands can be identified here as the even pair of impulse components of the cosine message straddling the carrier frequency. For a cosine message of $f_m$ Hz, each sideband lies $f_m$ Hz away from the carrier component. The total bandwidth of the modulated signal is that

[2] This application explains the origin of the term *modulation theorem*.

range of frequencies spanned by the components shown: the bandwidth of this signal is therefore $2 \times f_m$ Hz, or double that of the message. The bandwidth of a signal determines how much radio spectrum it occupies and, ultimately, how close to this station another radio station could be transmitting without their respective components overlapping. For example, if two radio stations wanted to transmit 5 kHz signals using this form of amplitude modulation without interfering with each other, their carrier frequencies would have to be at least 10 kHz apart.[3]

Because this particular method of amplitude modulation leads to these three spectral components, it is labelled *amplitude modulation, transmitted carrier, double sideband* or simply **AM-TCDSB**. Alternative modulation methods lead to variations of amplitude modulation, which differ in the presence or absence of these three components. For example, if the two sidebands are present, but no carrier, then a *suppressed carrier, double sideband* (AM-SCDSB) signal results, sometimes simply called double sideband or DSB, assuming that, unless otherwise stated, no carrier is present. It is also common to find methods in which no carrier and only one sideband is transmitted, called *single sideband* (SSB), which can then be further specified as being either only the *upper sideband* (USB), or only the *lower sideband* (LSB). These methods or variations of amplitude modulation will now be discussed.

The largest component in the above signal is found at exactly the carrier frequency $f_c$, and, taken alone, it is a pure sinusoid with no message content. The result of amplitude modulation with the message $m(t)$ was to introduce the two sidebands; it follows that any message content is found only in the sidebands. From a power perspective, the total power in this signal can be found by adding the square of each spectral component in the total (two-sided) spectrum. There are six components in all.

$$\text{TOTAL POWER} = \left[\tfrac{1}{4}\right]^2 + \left[\tfrac{1}{2}\right]^2 + \left[\tfrac{1}{4}\right]^2 + \left[\tfrac{1}{4}\right]^2 + \left[\tfrac{1}{2}\right]^2 + \left[\tfrac{1}{4}\right]^2 = \left[\tfrac{3}{4}\right]$$

Note that the power in the carrier alone is $(1/2)^2 + (1/2)^2 = 1/2$. Compared to the total power transmitted, it can be seen that 2/3 of the total power (66.7%) is found in the carrier frequency, which conveys no message at all. It may be concluded that the use of AM-TCDSB modulation is quite wasteful and very expensive for a transmitting station. For example, a typical 50,000-watt AM-TCDSB transmitter will be using 33,333 watts of that power to transmit a carrier component along with this unit amplitude sinusoidal message. Note that even if the message is removed (the music stops), the carrier is continuously transmitted. Given the fact that AM-TCDSB requires double the bandwidth of the message signal, and that two-thirds of the transmitted power is wasted in a carrier component, it might be reasonable to ask why this method is used at all. The secret lies in the demodulation, and the answer will soon become clear.

### 10.2.2 Demodulation of AM Signals

The purpose of demodulation is to extract the message component $m(t)$ from the modulated signal. This is the task that a radio receiver must perform on the signal received at its antenna. In theory, the same basic method of demodulation can be

[3] Government allocations for the carrier frequencies used by AM broadcast stations in North America are spaced evenly every 10 kHz along the band from 550 to 1600 kHz. (A radio station may use 1260 kHz, or 580 kHz, but never 1263 kHz or 578 kHz.) Given this legal restriction, the maximum allowable frequency of a baseband message transmitted using AMDSB must be less than 5 kHz. This explains in part the relatively poor quality of AM broadcast signals.

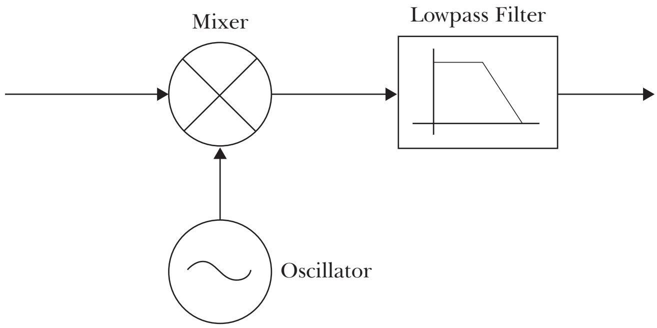



**FIGURE 10.7 Synchronous Demodulator** An AM-TCDSB signal can be demodulated by mixing with an exact copy of the carrier followed by a lowpass filter.

applied to all the various forms of amplitude modulation. Called *synchronous demodulation*, this method is based on the modulation theorem in which the time-domain multiplication by a cosine results in sum and difference components. Since multiplication by a cosine was used to modulate the signal, and thereby shift the baseband message signal to components near the carrier frequency, then multiplication by the exact same cosine should shift those message components back to their original frequencies.[4] A local oscillator in the receiver will re-create the original cosine and the received signal is mixed with this sinusoidal source in the block diagram in Figure 10.7.

The received signal in Figure 10.8A $b(t) = [1 + Km(t)]\cos(2\pi f_c t)$ is multiplied by the local oscillator signal $\cos(2\pi f_c t)$ to yield (for $K = 1$):

$$\begin{aligned}
\textit{Mixer output} &= [1 + m(t)]\, cos(2\pi f_c t)\cos(2\pi f_c t)\\
&= [1 + m(t)](1/2)[1 + \cos(4\pi f_c t)]\\
&= (1/2)[1 + m(t)] + (1/2)[1 + m(t)]\cos(4\pi f_c t)\\
&= (1/2)[1 + m(t)]\ (\textit{after lowpass filtering})
\end{aligned}$$

as seen in Figure 10.8B. In the final line, a lowpass filter removes the rightmost term in $\cos(4\pi f_c t)$, leaving the original message, as in Figure 10.8C. Such a lowpass filter need not be especially sharp, since the carrier frequency is much higher than that of the message. The original message signal $m(t)$ has now been successfully extracted from the transmitted signal. The factor of $1/2$ is of little consequence, since it can be amplified without distortion if necessary.

## 10.2.3 Graphical Analysis

The spectrum of the modulated signal was shown in Figure 10.6. In the synchronous receiver, this signal is multiplied by a locally generated sinusoid having the same characteristics as that of the original carrier. In the frequency domain, the mixing cosine is convolved with the received signal to give the results in Figure 10.9. In this figure, the carrier components of the transmitted signal are labelled **a** and **b**, while those of the cosine mixing signal are labelled **1** and **2** for clarity.

In the frequency domain, it can be seen that the original baseband signal in Figure 10.9D is back at the origin, while the higher-frequency components can be

[4] This procedure mirrors the voice scrambler example from Chapter 7.

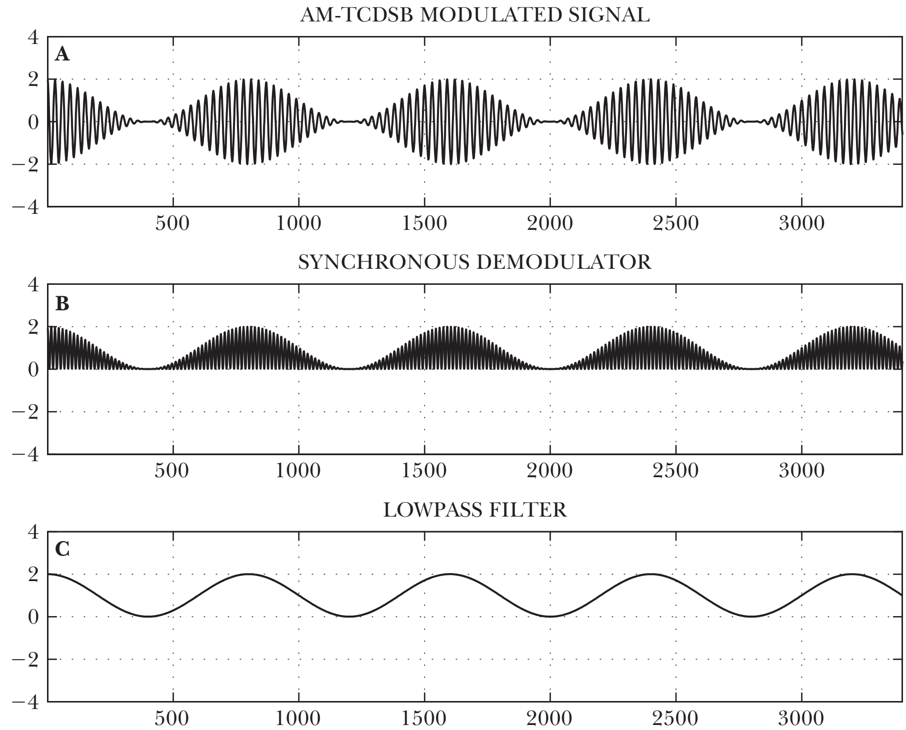



**FIGURE 10.8 Synchronous Demodulation** The AM-TCDSB signal in **A** is multiplied by an exact copy of the carrier to give the signal in **B**. After lowpass filtering, the original message is recovered in **C**.

removed if a simple lowpass filter is employed. Demodulation is successful once the original baseband signal is recovered. However, it can also be noted that the signal at the origin results from the sum of two copies of the message **1a** and **2b**. These two parts will add constructively only if the cosine in **B** exactly matches the original carrier signal *in both frequency and phase.* Otherwise:

1. **Mixing frequency too high or too low:** The components **1** and **2** in Figure 10.9B will move slightly apart or together. The corresponding components in the recovered message signal **2a** and **1b** will move away from the origin, giving two overlapped copies and a nonsense output signal after lowpass filtering.
2. **Mixing phase incorrect:** The components **1** and **2** in Figure 10.9B will rotate around the frequency axis, consistent with shifting a cosine in time. The corresponding components in the recovered message signal **2a** and **1b** will rotate about and will tend to cancel each other out as the phase varies unpredictably.

While it might be argued that a local oscillator in the receiver could generate a sinusoid that closely matched the incoming carrier frequency, this method is further complicated by the fact that the phase of the local oscillator must also match that of the carrier frequency. While this problem would appear to doom synchronous demodulation, it merely points out that *an exact copy of the original carrier in both frequency and phase* is required. In practice, for the transmitted carrier forms of amplitude modulation such as AM-TCDSB the (otherwise useless) carrier component can be tapped for use as the mixing signal; this method guarantees a mixing

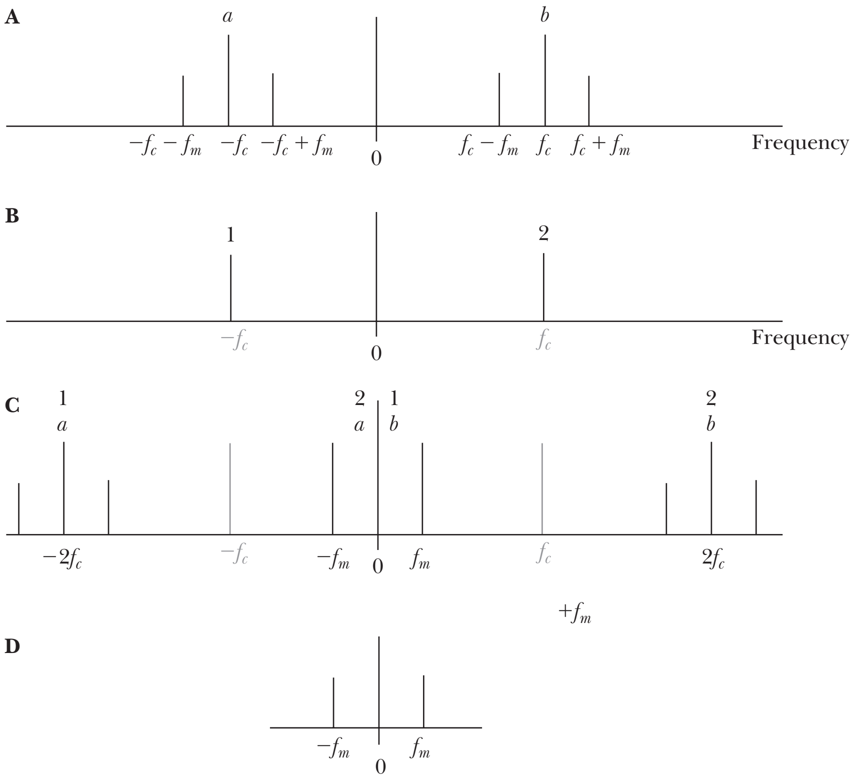



**FIGURE 10.9 Synchronous Demodulation (DSB)** The signal **A** carries a cosine message signal $m(t) = \cos(2\pi f_m t)$. The message can be recovered from this signal by mixing the received signal with a cosine at the carrier frequency **B**. The result in **C** gives the recovered message **D** after a lowpass filter.

signal that exactly matches the one that was used for modulation. Such a technique is used only in expensive receivers, and cannot be used at all if the carrier was suppressed for transmission.

For an AM-SCDSB (suppressed carrier) signal, the results of Figure 10.9 would be identical except for the missing carrier components, and demodulation fails with no reference carrier signal. Other suppressed carrier signals will be discussed in the next section, but first a unique characteristic of AM-TCDSB signals will be demonstrated.

## 10.2.4 AM Demodulation—Diode Detector

Consider an inexpensive *pocket AM radio.* Experience with such radios reveals that careful tuning is not really necessary, and their low cost seems to preclude the frequency stability and accuracy required by a synchronous demodulator. It follows that some other method of demodulation may be possible for these particular signals. The time-domain sketches of the transmitted signal and the received signal after mixing were seen in Figure 10.8. Examine these images carefully, particularly signal Figure 10.8B. Is there some simple operation that might yield this signal without performing any mixing?

To a rough approximation, the signal might simply be the positive-only part of the modulated signal. In practice, essentially the same result could be obtained by passing the received signal through a diode (no mixing necessary), as shown in Figure 10.10.

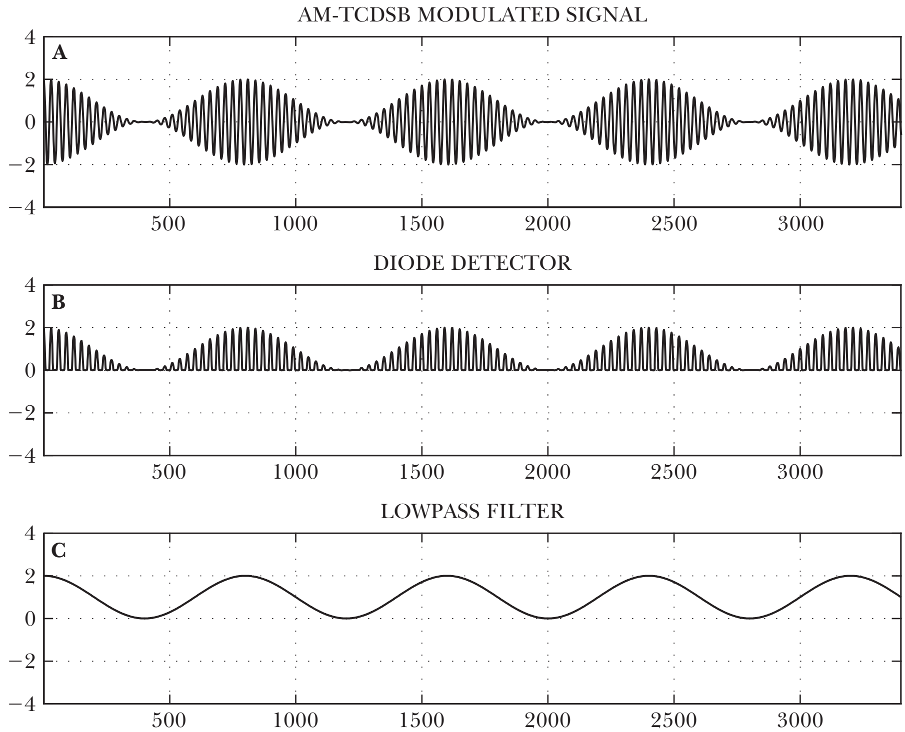



**FIGURE 10.10 Diode Detection** The AM-TCDSB signal in **A** is passed through an ideal diode to give the signal in **B**. After lowpass filtering, the original message is recovered in **C**. Compare to Figure 10.8.

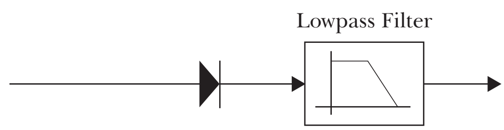



**FIGURE 10.11 Diode Detector** An AM-TCDSB signal can be demodulated using nothing more than a diode and a lowpass filter.

The method of *diode detection* is a simple and inexpensive alternative to the use of a synchronous demodulator. The simplest such receiver consists of a diode and a lowpass filter, as in Figure 10.11. This observation makes possible the construction of very low-cost receivers for AM-TCDSB signals. (This same result is not possible for any of the other forms of amplitude modulation.) It can finally be explained why AM-TCDSB is employed in broadcast stations, despite all the disadvantages that seem to weigh against its use. For one broadcast AM-TCDSB transmitter, wasting over 66 percent of its power in a carrier component and requiring a bandwidth double that of the message signal, many thousands of receivers can benefit from using a simple method of demodulation, costing only pennies.

In practice, germanium (as opposed to silicon) diodes must be used in this application, since silicon diodes exhibit an intolerable forward voltage drop. While the time-domain analysis may seem convincing enough, it is worthwhile to complete the frequency-domain analysis to determine exactly why this method works for AM-TCDSB signals. Unfortunately, it is difficult to directly analyze the diode detector

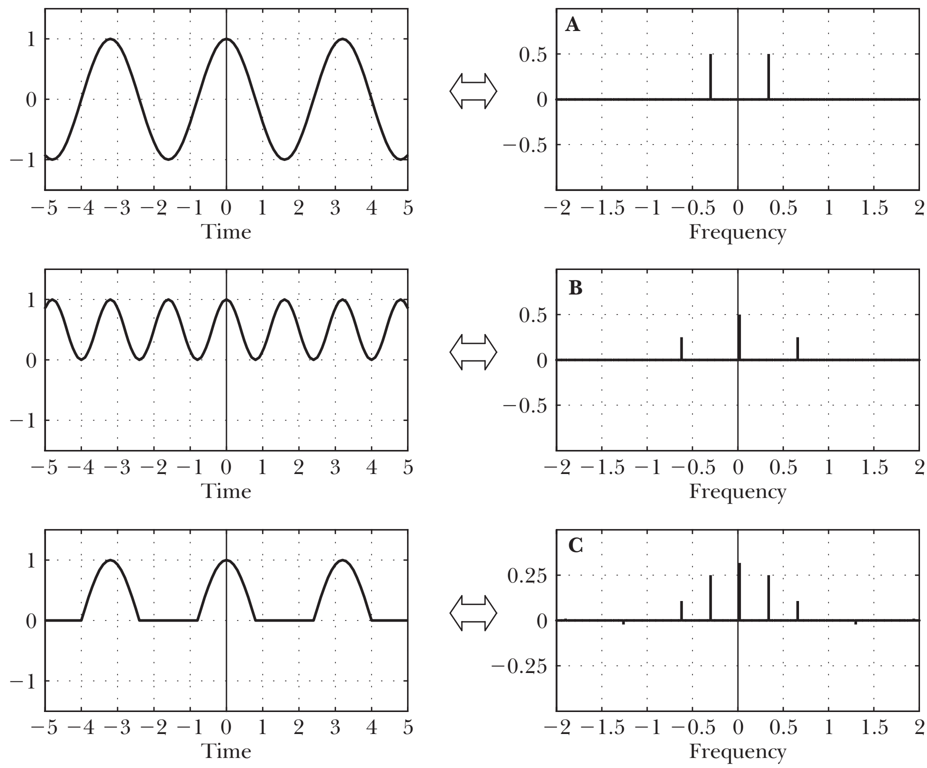



**FIGURE 10.12 Synchronous Demod vs. Diode Detector** Whether a cosine carrier **A** is squared **B** during synchronous demodulation or passed though a nonlinear component **C** in a diode detector, the presence of a DC component in both cases ensures that the AM-TCDSB signal is demodulated.

using the Fourier transform, since, as seen in Chapter 1, the diode is a nonlinear device. Recall the definition of a system containing an ideal diode:

$$s(t) \rightarrow s(t) \text{ if } s(t) > 0, \text{ 0 otherwise}$$

Mathematically, passing the modulated signal through the diode results in only positive values. By definition, the message component $b(t) = 1 + Km(t)$ is strictly positive because of the limitations previously placed on $K$ and $m(t)$. Therefore, this effect can be modelled as $b(t)$ multiplied by the now half-wave rectified carrier. The spectrum of the half-wave rectifed cosine has been computed in Chapter 2. Figure 10.12 compares the mixer output of the synchronous demodulator with the appearance of the diode detector output for an unmodulated cosine carrier.

The purpose of demodulation is to extract the message component. In synchronous demodulation, the modulated signal is multiplied by a copy of the carrier. In the frequency domain, the resulting signal consists of three copies of the message, centered at frequencies $[-2f_c, 0, +2f_c]$. The high-frequency components were filtered out to leave only the desired component at $f = 0$, as seen in Figure 10.12B.

In the case of the diode detector, the spectrum of the modulated signal is convolved with the spectrum of a half-wave rectified cosine. Figure 10.12C shows that the half-wave rectified cosine is a periodic signal with the same period as that of the carrier. This spectrum has many components of varying amplitude, spaced at integer multiples of $f_c$. Consequently, the spectrum of the received signal after passing through the diode consists of many copies of the message, with varying amplitudes, centered at integer multiples of $f_c$ along the frequency axis. This spectrum differs from that of the synchronous demodulator primarily in the presence of additional high-frequency components, and since the only component of interest lies at $f = 0$, a lowpass filter will effectively isolate the message in exactly the same way as the synchronous demodulator.

### 10.2.5 Examples of Diode Detection

Several examples serve to illustrate some practical consequences of the diode detector.

1. In the early days of radio, a *Cat's Whisker* radio was essentially a diode detector. Because solid-state diodes had not yet been invented, a slab of selenium or germanium was used. In air, an oxide layer rapidly forms on the surface of the crystal. By probing the surface with a sharp wire (the cat's whisker), a good diode effect could eventually be obtained, and demodulation was immediate. This same principle makes possible *crystal radios*, in which a germanium diode serves to detect the AMDSB signal. Of course, such radios offer no real channel selection, and the signal received would be that of the strongest nearby source of AMDSB signals (the local broadcasting station).
2. Some potentially serious problems with radio interference also come about from the principle of diode detection. For instance, a neighbor may complain that her digital piano picks up radio signals from a nearby AM broadcast transmitter. Now, the piano is not a radio receiver, but it contains many nonlinear electronic components, including a power amplifier with loudspeakers. If a radio signal enters the piano circuit because of poor shielding (even the speaker wires can form a suitable antenna), the detection of AMDSB signals is almost guaranteed. Similarly, people living near high-powered transmitters can sometimes pick up radio signals on the telephone, or other *non-radio* devices, all of which contain nonlinear components.

The possibility of radio frequency interference, or RFI, should be considered whenever any electronic equipment is designed. As these examples show, there are two possible problem scenarios:

1. Radio frequency signals that cause problems in devices not intended to be radio receivers.
2. Non-radio devices that cause problems with radio circuits; this often occurs when harmonics from a clocked digital circuit (such as a computer) extend into the radio frequency range.

If the potential for RFI is anticipated at the design stage, potential problems can often be avoided through the judicious use of shielding or filtering to eliminate sources of problem signals.

## 10.3 Suppressed Carrier Transmission

If the carrier frequency represents wasted power when transmitting with AMDSB-TC, the simple solution would seem to be to suppress the carrier and to transmit only the sidebands. Furthermore, the power that was used in transmitting a carrier can be directed to increasing the power of the transmitted sideband(s), making for

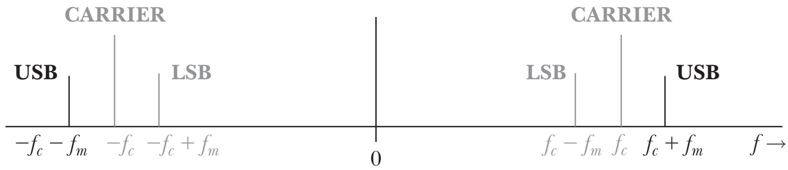



**FIGURE 10.13 Upper Sideband (USB) Transmission** A cosine message signal $m(t)$ can be recovered from this signal by recognizing that the carrier component does not contribute to the message and the other sideband components convey the same message information.

increased range and efficiency.[5] If only one sideband is transmitted, the resulting single sideband (SSB) transmission may use only the upper sideband (USB) or the lower sideband (LSB), as seen in Figure 10.13. The problem with all of these methods is that a simple diode detector no longer works and some form of synchronous demodulation is always necessary. For DSB-SC, in particular, the difficulties of generating an in-phase copy of the carrier frequency make demodulation virtually impossible. On the other hand, SSB-SC signals are quite easily received by synchronous demodulation after some careful receiver adjustments. The transmission of SSB signals is commonplace in amateur radio and by some shortwave broadcasting stations. Citizen band (CB) radios typically use USB or LSB methods in addition to an AMDSB-TC on channels 1 through 20. Using only one sideband not only eliminates the requirement to transmit the carrier component, but it also reduces the over bandwidth to one half that required when both sidebands are used.

Finally, a compromise to the above situation arises when the carrier is not completely suppressed. Such a signal with a reduced carrier component would be said to have a *vestigial carrier*. In particular, analog television transmissions use vestigial carrier single sideband to transmit the video portion of a television program.

## 10.3.1 Demodulation of Single Sideband Signals

Demodulation of single sideband signals (AMSSB-SC) requires a locally generated replica of the (suppressed) carrier frequency. A synchronous demodulator is called into play for suppressed carrier signals. The oscillator used to create this local signal must be stable and accurate. However, the effect of being slightly off frequency will be evident to a listener and retuning is a relatively simple task.

Consider a received USB signal, where the baseband message is a cosine of frequency $f_m$, as shown in Figure 10.14A. This upper sideband signal is essentially an AMDSB-TC signal with the carrier and the lower sideband removed. The only components transmitted are at $(-f_c - f_m)$ and $(f_c + f_m)$ (labelled **a** and **b** for clarity), which essentially describe a cosine at frequency $(f_c + f_m)$. Note there is no power consumed in transmitting a carrier component in single sideband transmissions.

The synchronous demodulator multiplies the above signal by a cosine at frequency $f_c$, as in Figure 10.14B, where the cosine components are labelled **1** and **2**. This time-domain multiplication, a convolution in the frequency domain, results in four sum and difference terms. Of interest are the terms **2a** at $(-f_m)$ and **1b** at $(+f_m)$. The desired message signal **D** is easily and accurately extracted with a lowpass filter.

[5] Government regulations often place limits on the total allowable transmitted power, making it worthwhile to concentrate that power in information-laden sidebands.

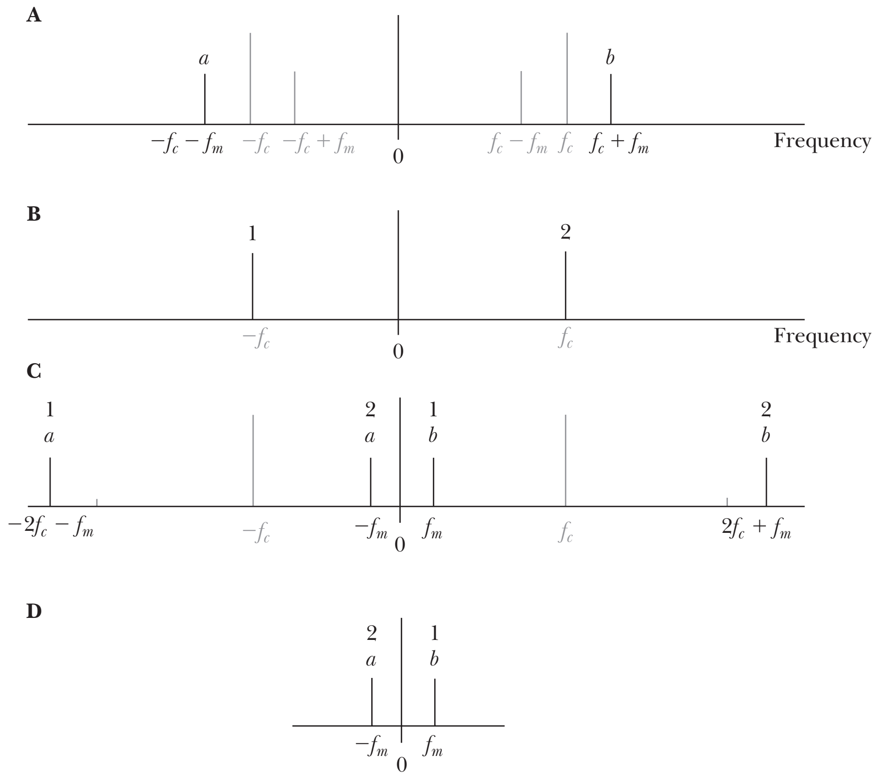



**FIGURE 10.14 Synchronous Demodulation (USB)** The signal **A** carries a cosine message signal $m(t) = \cos(2\pi f_m t)$. The message can be recovered from this signal by mixing the received signal with a cosine at the carrier frequency **B**. The result in **C** gives the recovered message **D** after a lowpass filter.

The challenge in successfully using synchronous demodulation for a USB signal is of course is generating the signal **B**, which should be the same frequency as the carrier but for which there is no reference signal at all. If stable (crystal) oscillators set to identical frequencies are used in the transmitter and receiver, this can be accomplished with reasonable success. Unlike AM-SCDSB, the phase of the mixing signal is unimportant. Often, the mixing frequency is varied by hand by slowly turing a tuning knob on a radio receiver. By varying the frequency of the signal **B** until the output frequency *sounds* correct, the correct mixing frequency can be found. So what happens if the mixing frequency is not exactly at $f_c$?

1. **Mixing frequency too high:** The components **1** and **2** in Figure 10.14B will move apart. The corresponding components in the recovered messages signal **2a** and **1b** will move closer together, giving a recovered cosine at a lower-than-normal frequency.
2. **Mixing frequency too low:** The components **1** and **2** in Figure 10.14B will move closer together apart. The corresponding components in the recovered messages signal **2a** and **1b** will move apart, giving a recovered cosine at a higher-than-normal frequency.

A similar effect can be observed for LSB signals, where synchronous demodulation can be readily applied to demodulate a received signal.

## 10.3.2 Percent Modulation and Overmodulation

When AMDSB was introduced, some care was taken to constrain the message to values in the range $[-1, +1]$, and, in particular, the constant factor $K$ was set equal to 1 in the expression:

$$\text{Modulated signal} = b(t) = [1 + Km(t)]\cos(2\pi f_c t)$$

The value $K$ multiplies the amplitude of the message signal. If K is made smaller than 1, the modulation effect is diminished, and the relative power contribution of the message component decreases. In the special case of $K = 0$, only the carrier term remains, and no modulation happens. In Figure 10.15A, the value $K = 0.5$ is used, and it can be seen that the carrier is being only lightly modulated.

**Exercise:** What are the total power and the relative power of the carrier and message components in Figure 10.15A?

As $K$ varies from 0 to 1, the power contribution from the message component increases, and the relative contribution from the carrier component falls to 66 percent when $K = 1$ for this example. Modulation and demodulation behave as expected in both synchronous demodulation and diode detection. Unfortunately, this is not the case if $K$ exceeds 1. In Figure 10.15C, $K = 1.5$, and it can be seen that the modulating term goes negative, causing the overall signal to have many unwanted components, and the demodulated signal as shown no longer reflects the original message. Clearly, this situation is to be avoided.

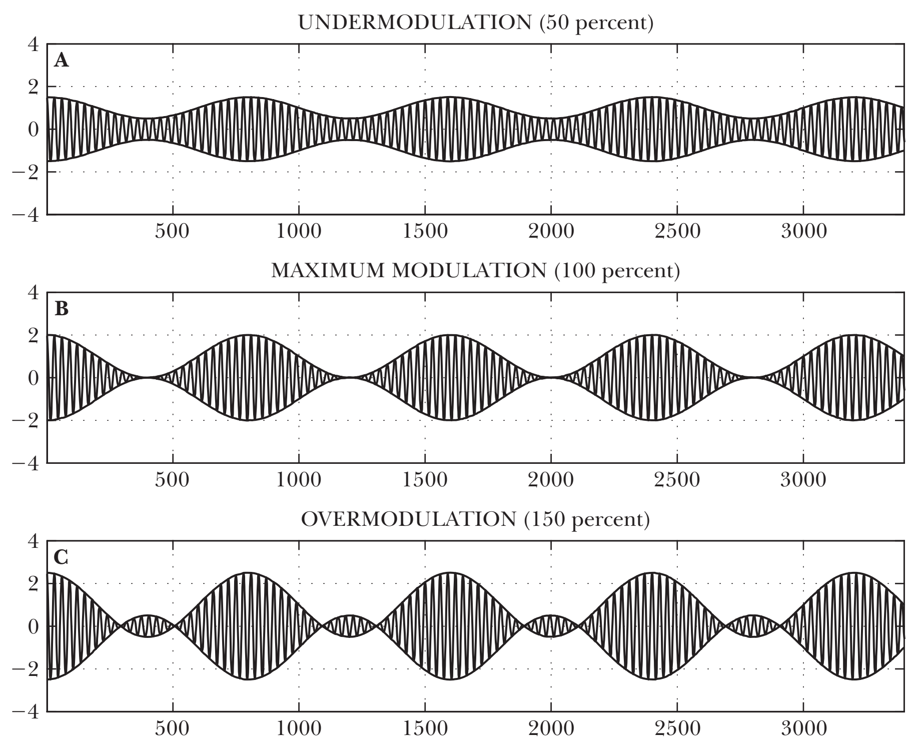



**FIGURE 10.15 Modulation Index** Showing the effect of varying $K$ in the AM-TCDSB signal $[1 + Km(t)]\cos(2\pi f_c t)$. Signal **A** for $K = 0.5$ shows *undermodulation.* Signal **B** for $K = 1$ shows full modulation as used in most examples in this chapter. Signal **C** for $K = 1.5$ shows *overmodulation.*

# 10.4 Superheterodyne Receiver

The signal received at a radio antenna is very weak (on the order of microvolts), and much amplification is necessary before that signal can be demodulated and played through loudspeakers or headphones. It is not sufficient to build a very high gain amplifier attached directly to an antenna, since the goal is to amplify only the station of interest. A radio receiver needs to have not only high gain (sensitivity), but also high selectivity. In other words, it is only necessary or desirable to amplify the one station to which the radio is tuned. In practice, this means that sharp bandpass filters and high-gain amplifiers are necessary in the *front end* of a radio receiver (that part between the antenna and the demodulator), as shown in Figure 10.16.

Unfortunately, while it is possible to build a radio frequency amplifier with the above characteristics, the receiver would be useful only for one station broadcasting on exactly the carrier frequency for which the circuit was designed. To be useful, the bandpass filter frequency and amplifier characteristics must also be readily modified to tune the radio to the carrier frequencies of different stations. This is not easily accomplished because the high-gain requirements generally necessitate several stages of amplification, and the high selectivity requires careful adjustment of the components in each stage. In the early days of radio, several stages of tuning with four or five tuning knobs were sometimes required. Since modern radios only have one tuning knob, it follows that some clever variation of this method is actually used. Virtually all radios today use the *superheterodyne* principle to avoid the need for multiple tuning knobs and the careful readjustment of several amplifier stages when tuning to different stations.

In a superheterodyne receiver, the main internal amplifiers and bandpass filters are tuned once, at the factory, to a single radio frequency called the intermediate frequency (IF). In effect, these radios are designed around a single channel receiver, which cannot be tuned to different stations. Of course, the radio can be tuned, but not by varying the amplifier characteristics. Instead, tuning is accomplished by mixing the incoming signals with a cosine, carefully chosen so that the sum or difference of the resulting spectral components falls at the IF. In essence, incoming radio signals are converted to the characteristics of the fixed-frequency

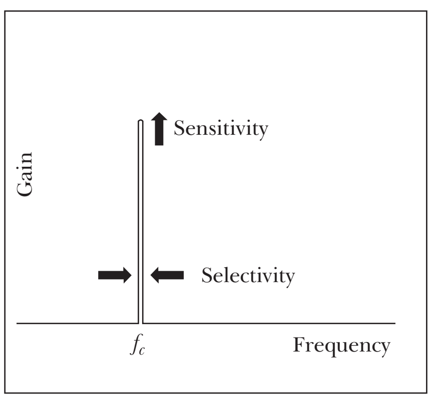



**FIGURE 10.16 Selectivity and Sensitivity** These two characteristics are both desirable for a radio receiver amplifier. It is difficult to achieve both in an electronic circuit intended to work at several different carrier frequencies $f_c$.

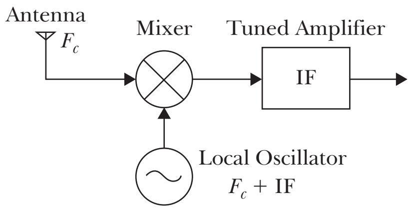



**FIGURE 10.17 Superheterodyne Receiver** A single-frequency (IF) amplifier is constructed to be both very sensitive and selective. Incoming signals are brought to the intermediate frequency by mixing with an internal frequency varied using the station selector.

| Application | Frequency |
|---|---|
| AM Broadcast Band Radios (540–1600 kHz) | 455 kHz |
| FM Broadcast Band Radios (88–108 MHz) | 10.7 MHz |
| Analog Television Receivers (CH2 = 54 MHz) | 45 MHz |

**FIGURE 10.18** Typical Intermediate Frequencies.

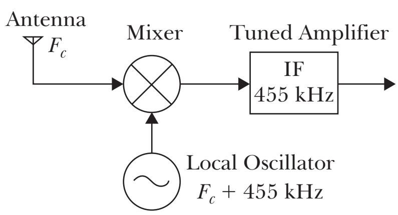



**FIGURE 10.19 Superheterodyne Receiver** An AM broadcast band radio uses a 455 kHz IF. To tune to a station broadcasting at $f_c = 1210$ kHz, the internal oscillator is set to $f_c + 455 = 1765$ kHz. At the mixer output, the resulting *difference* frequency moves the desired station into the passband of the IF stage, where it can be amplified and demodulated. The *sum* frequency lies outside the IF passband and is ignored.

stage (intermediate frequency), rather than trying to retune the amplifiers to different stations.[6]

Figure 10.17 shows a sketch of the front end of a superheterodyne receiver. To tune to a specific station, the local oscillator frequency is varied (using the tuning knob), such that its frequency is equal to that of the desired station plus the intermediate frequency as shown.

The mixer output is the sum and difference of these two signals, with the result that the desired signal enters the IF amplifier at its fixed frequency. For different applications, some common intermediate frequencies are generally used throughout the world as shown in Figure 10.18.

For example, when tuned to a station with carrier frequency $f_c$, a radio with a standard intermediate frequency of 455 kHz will internally generate a mixing signal of $f_c$ +455 kHz, as shown in Figure 10.19.

[6] This is essentially the same technique described in Chapter 1, which allowed a cable converter box to convert any incoming stations to be received by a television receiver that stays fixed on Channel 3. In this case, the fixed Channel 3 is like the intermediate frequency inside a receiver.

### 10.4.1 An Experiment with Intermediate Frequency

A relatively simple experiment can demonstrate that a radio uses the superheterodyne principle. Because the mixing frequency generated inside such receivers is itself a radio frequency carrier, then this signal might propagate out of the radio, if the circuit is not adequately shielded. If so, the radiating mixing frequency could be picked up by other nearby radios. The following experiment can be performed using two inexpensive AM pocket radios.

Let the two radios be **A** and **B** as shown in Figure 10.20. One radio will be tuned to a strong local broadcasting station in the upper half of 550–1600 kHz band. This will be the receiving radio; let this frequency be $f_B$ kHz. Now, the standard intermediate frequency for such radios is 455 kHz. Tune the other radio to a frequency $f_A$, where $f_A = f_B - 455$ kHz. For example, if Radio **B** is tuned to a local station on $f_B = 1260$ kHz, tune the Radio **A** to around $f_A = 805$ kHz. It is not important whether or not there is any station actually broadcasting near this second frequency. Now, the internal mixing frequency of Radio **A** should be $f_A + 455$ kHz. The reason for this particular choice of $f_A$ should now be clear. The frequency of this internal oscillator is precisely that frequency $f_A$ to which the other radio is tuned (in this example, $805 + 455$ kHz $= 1260$ kHz). The signal radiating from the second radio should interfere with the reception of the first.

Bring the two radios very close together, and slowly tune the Radio **A** back and forth around frequency $f_A$ (805 kHz). If all goes well, the mixing frequency inside Radio **A** will leak out of the radio to interfere with the broadcast signal. This effect confirms that both radios use a superheterodyne design with a 455 kHz IF and that the mixer in Radio **B** uses $f_c + 455$ kHz when tuned to a particular station at $f_c$ Hz.

**Supplementary Question:** Why does the interference take the form of a loud squeal?

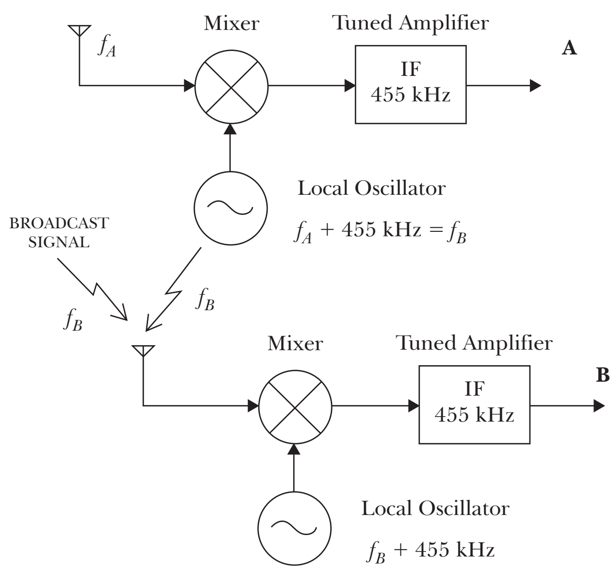



**FIGURE 10.20 Superheterodyne Receiver Experiment** Using two inexpensive AM broadcast band receivers, the presence of an internal oscillator and the use of a 455 kHz IF can be easily demonstrated. Radio **B** is tuned to a local AM broadcast station, while Radio **A** is tuned such that its internal mixing frequency matches that of station **B**. Leakage from the oscillator circuit in Radio **A** interferes with the desired signal in Radio **B**.

## 10.4.2 When Receivers Become Transmitters

In many countries of the world, radio and television receivers must be licensed. In others, use of broadcast receivers is outright illegal. In many parts of North America, police radar detectors are outlawed. In all cases, it may be possible to detect a radio receiver from the signals it unintentionally emits.

The property uncovered in the above experiments can be exploited to detect illegal or unlicensed radio receivers. At first glance, it might seem improbable that an operating receiver could be detected behind closed doors. After all, how can anyone detect illegal *reception* of a radio signal? As the experiment shows, most superheterodyne receivers will actually be transmitting (very weakly) their internal mixing frequency whenever they are switched on. In such situations, government vans equipped with sensitive receivers and direction-finding equipment can patrol the streets looking for telltale radio receiver emissions. Similarly, highway patrol officers can detect radar detectors from the roadside. By listening for the mixing frequency, not only can an active receiver be detected, but the frequency it is tuned to can also be determined, knowing the IF characteristics. Only a crystal radio or some such radio set could evade these searches.

## 10.4.3 Image Frequency

A potential problem common to all superheterodyne receivers concerns what is known as the *image frequency*. When a radio is tuned to a station with carrier frequency $f_c$, the internal oscillator is set to $f_m = f_c + IF$. The mixed components are at the difference $(IF)$ and the sum $(IF + 2f_c)$ of the two frequencies. Only the difference frequency is of interest, and this signal is amplified in the IF stage.

The problem lies in the fact that both the sum and difference of the mixing components are created. With the receiver tuned to $f_c$, let another signal arrive with frequency $f_2 = (f_c + 2IF)$. Using the same mixing frequency of $f_m = f_c + IF$, the mixed components are at the sum $(f_2 + f_m = 2f_c + 3IF)$ kHz, and the difference $(f_2 - f_m = IF)$. So, this second signal is also converted to the intermediate frequency and would be received by the radio. This unexpected result is that the signal at frequency $f_2$ is also received. There will always be two candidate input signals that would convert to the fixed IF. The desired signal and the image frequency are always separated by twice the intermediate frequency, as shown in Figure 10.21.

The solution is to filter the offending signal using a variable-cutoff lowpass filter connected to the tuning knob. The filter does not need to be sharp, as the frequencies to be blocked are separated by twice the intermediate frequency. Handheld radio scanners used to receive radio signals on the VHF and UHF bands often have poor image rejection, and interference to reception can easily occur.

## 10.4.4 Beat Frequency Oscillator

Demodulation of SSB signals within a superheterodyne radio is made easier by the presence of an IF stage. Whatever carrier frequency the radio is tuned to, the signal is converted to the intermediate frequency before demodulation. From the perspective of the demodulator, every signal is always at the same carrier frequency. For synchronous demodulation of SSB signals, the mixing frequency would be set to the intermediate frequency or be adjustable around that frequency. Now called a *beat frequency oscillator*, this mixer is seen in Figure 10.22.

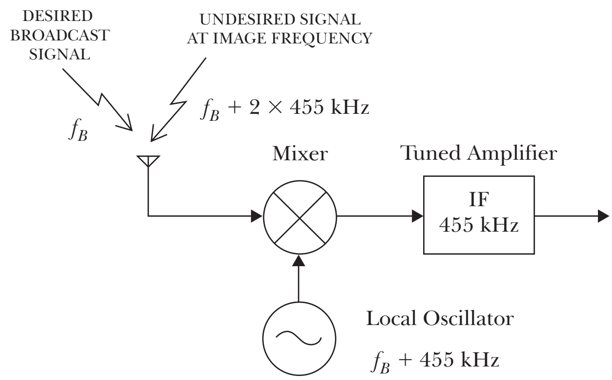



**FIGURE 10.21 Superheterodyne Image Frequency** Because the mixer produces both sum and difference frequencies, there are always two possible input frequencies that could enter the IF stage. The *image frequency* differs from the desired input signal by exactly 2× IF Hz. A lowpass filter in front of the mixer can eliminate this effect.

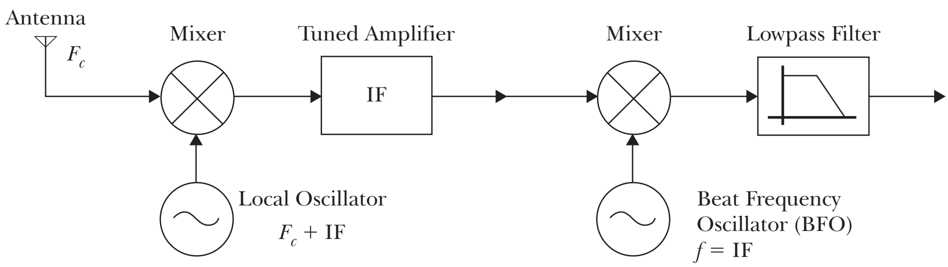



**FIGURE 10.22 Synchronous Demodulation—Superheterodyne Receiver** Because the IF output is a constant frequency regardless of the carrier frequency of the received station, the oscillator for a synchronous demodulator can be set to a constant frequency, generally adjustable around the IF.

# 10.5 Digital Communications

The transmission and reception of digital information follows the same principles as have been developed for analog applications. Modern communication systems frequently involve messages consisting of a sequence of bits such that the message signal $m(t)$ is discrete as opposed to a continuously varying analog signal source. In fact, because any analog signal could be sampled and transmitted digitally as a sequence of binary sample values, a digital communication system is universally useful for both analog and digital messages. The future of communications promises a convergence of various message types within a digital framework. Because a single high-speed digital channel can carry both analog and digital signal information, many existing analog transmission systems will become obsolete in the future. This trend is evident in many applications, from the dramatic demise of the long play vinyl records in the 1980s in favor of digital compact disks to the rapid shift from analog to digital cell phone technology in the early 2000s. Whatever the digital message being conveyed, the theory and methods already discussed in this chapter for analog messages will not change—the need to modulate a carrier signal to pass through a channel remains—although certain techniques specific to the digital application will be introduced.


**FIGURE 10.23 A Digital Communication System** For a two-way path as shown, modulators and demodulators convert a digital message $m(t)$ to a form suitable for transmission over the analog channel.

Figure 10.23 shows a digital communications system in which messages may pass in one or two directions. For example, the analog communcations path may be a broadcast satellite television channel 6 MHz wide on a carrier frequency of 2 GHz, or perhaps a pocket pager receiving messages on a 15 kHz wide channel at a frequency of 970 MHz. In every case, the digital signal must be modulated into a suitable form to pass easily over the available channel and then demodulated upon reception. The words *modulator* and *demodulator* are combined to form the common term *modem*.

In the special case of a normal voice telephone line, the channel has been designed and optimized to carry an analog voice signal and is strictly bandlimited to frequencies between 300 and 3300 Hz. A *voiceband* modem or dial-up modem is designed to convey digital information over this channel using a variety of different modulation techniques at rates exceeding 33,000 bits per second under ideal conditions.

## 10.5.1 Modulation Methods

Recall that a carrier signal $c(t) = A\cos(2\pi f_c t + \Phi)$ has parameters $(A, f_c, \Phi)$ that will vary by a message waveform (now a discrete signal). The three parameters (amplitude, frequency, phase) lead to three basic digital modulation techniques called *amplitude shift keying* (ASK), *frequency shift keying* (FSK), and *phase shift keying* (PSK). These three methods are analogous to the analog modulation types (AM, FM, PM). The principle behind digital communications is that certain variations in these three parameters will convey an underlying binary message, as seen in Figure 10.24.

In this discussion, the issue of bandwidth arises frequently as the channel to be used inevitably has a limited bandwidth, while a user generally has the desire to transmit bits as quickly as possible through that channel. In all the techniques in Figure 10.24, the bandwidth of the modulated signal increases if the rate of signal changes increases. The overriding challenge in any digital communications system is to carry digital information through a channel both quickly and accurately (without errors).

## 10.5.2 Morse Code

One of the first modern examples of digital communications may be found in the development of telegraphy in the late nineteenth century. The *Morse code* was invented by Samuel Morse in 1878 as a signalling code designed for use over telegraph wires by human operators manipulating a mechanical contact switch called a *telegraph key*. The Morse code is a true digital-coding system that also embodies the concept of data compression. The Morse code survived over one hundred years in various forms, although its commercial use was diminishing rapidly at the turn of the twenty-first century. The Morse code is perhaps most familiar in the distress message `SOS`, which is coded as three *dots* as $\cdots$ followed by three *dashes* as $- - -$ followed by three *dots* as $\cdots$ where the three dots represent the character S and the three dashes

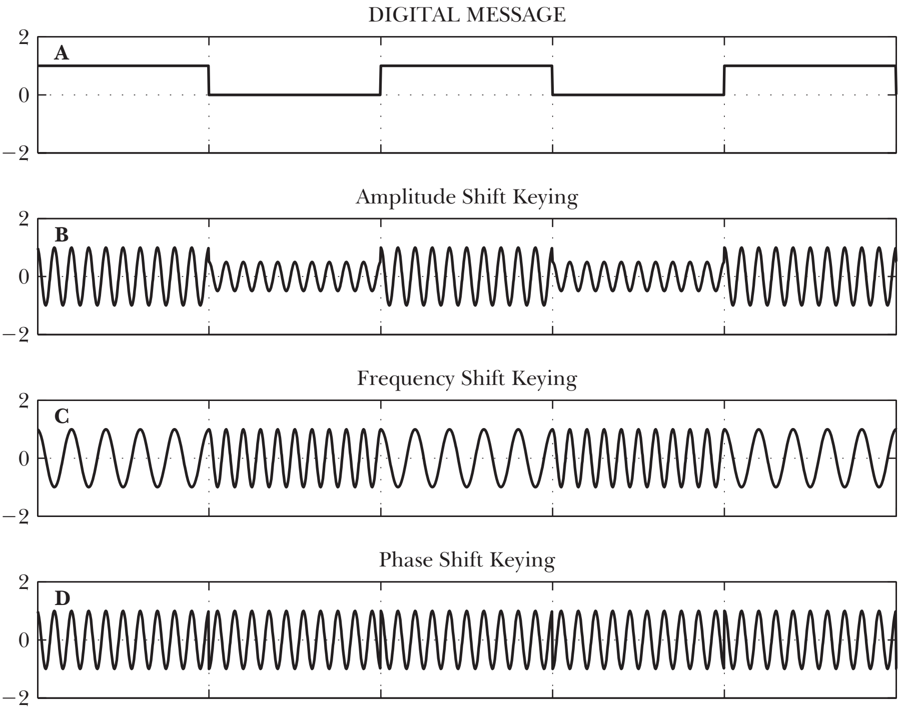



**FIGURE 10.24 Digital Modulation Techniques** The digital message in **A** may be imposed on a carrier $c(t) = A\cos(2\pi ft)$ by varying the amplitude (ASK), frequency (FSK), and/or phase (PSK) of the carrier.

represent the character O. The dot and dash are the Morse code equivalent to the ones and zeros of a binary message. Consequently, the message `SOS` could be written in binary form as the bits `111 000 111`.

It can be noted that representing a single character with only three bits is a significant accomplishment. In comparison, the ASCII code that is commonly used in digital computer and communications systems uses 7 bits for every character. Morse observed that certain characters such as E and T occur frequently in English text, while others such as Z and Q are far less common. The code he invented uses a variable number of bits for each character, so that E is a single dot ·, T is a single dash −, S and O are 3 bits each (as above), while Z is 4 bits long − − ··. The technique of using variable length codes optimized with respect to character frequency statistics is the basis for many modern data compression techniques.

A telegraph operator using the Morse code would translate text messages into Morse-coded characters for transmission and send the characters manually; each press of a telegraph key would turn the transmission circuit ON for the duration of the press. In practice, the difference between a dot and dash (or between a 1 and a 0) is in the length of time the key is pressed. Let a dash be 3 times the duration of a dot, and let 3 dot spaces separate letters. The resulting waveform corresponding to the Morse-coded message `SOS` may be modelled as a pulse train with varying pulse lengths and distance between pulses. This is the digital message signal *m(t)* to be transmitted, as in Figure 10.25.

It is significant that the ones and zeroes in this message are not simply conveyed as two signal states (i.e., on = 1 and off = 0). Communications systems typically employ a variety of ways to convey digital information, and varying pulse width is one useful approach to this task.

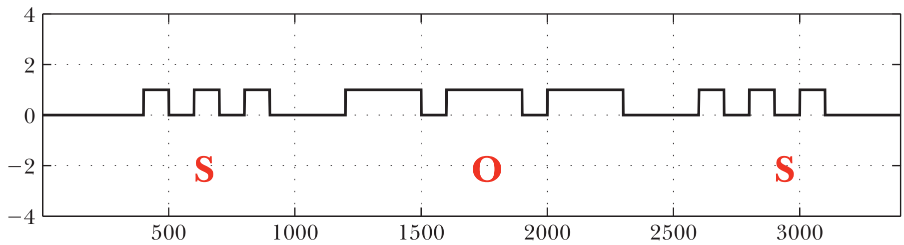



**FIGURE 10.25 Morse Code** The Morse-coded message SOS is alternatively described either as *dots* and *dashes* or as a digital signal made up of variable length pulses.

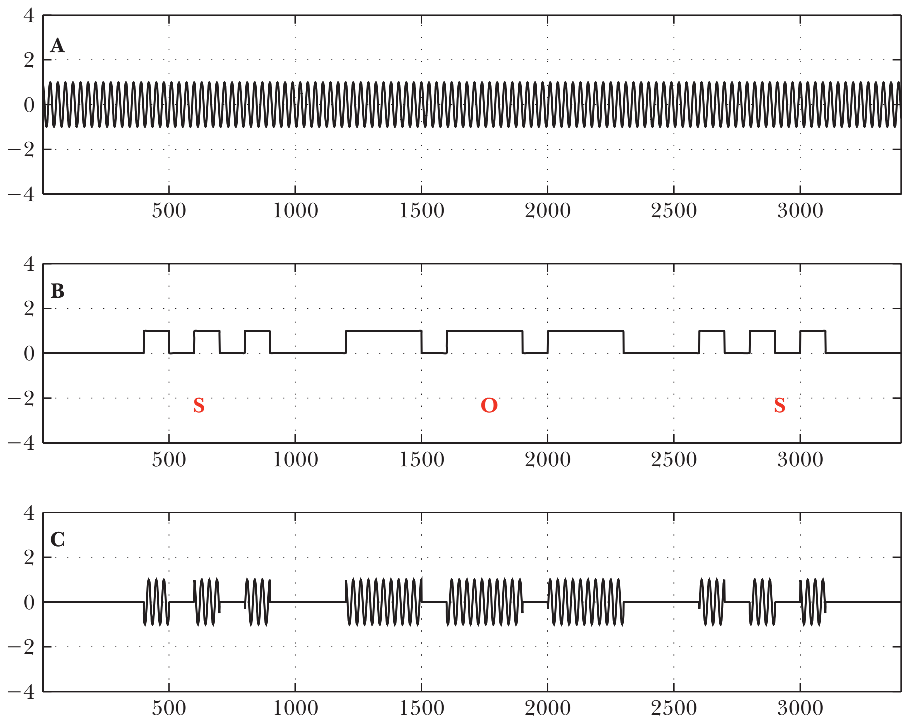



**FIGURE 10.26 Amplitude Shift Keying (ASK)—On Off Keying (OOK)** The Morse-coded message in **B** is to be transmitted by modulating the carrier **A**. The carrier is multiplied by **B**, and the resulting modulated signal in **C** is ready to transmit. Amplitude variations (0 and 100%) in the modulated carrier convey the underlying message content.

If the carrier in Figure 10.26 is an audio frequency cosine (e.g., $f = 1000$ Hz), the distinctive sounds of a telegraph signal would be heard *dit dit dit, dah dah dah, dit dit dit.* Telegraph operators are able to decode the received characters by ear and after some practice can receive and transcribe messages at speeds exceeding fifty words per minute. If the audio signal were fed into the microphone input of a radio transmitter, it would be possible to convey the Morse-coded message (now in analog form) over very long distances on a radio channel.

In practice, the message stream is generally used to turn a radio frequency carrier frequency (e.g., $f = 7000$ kHz) on and off directly. A similar pulse-length digital modulation approach is used in handheld infrared remote controls to convey control information to home entertainment devices, in this case, by turning an oscillating infrared light source on and off rather than a radio frequency carrier.

### 10.5.3 On Off Keying (OOK)

Like analog messages, the transmission of digital messages over long distance requires the use of a carrier. Given that the message is a digital waveform, which is either on or off, the simplest modulation system will turn the carrier on and off. A very simple and practical Morse code transmitter would consist of a radio frequency oscillator in which the telegraph key serves as the on/off switch. This approach is effectively a form of amplitude modulation (AM) in which the carrier amplitude varies with the message. In digital communications, the use of amplitude modulation is specified as *amplitude shift keying* (ASK), since the amplitude shifts between discrete values (in this case, either 0 or 100%). This special case of ASK is called *on off keying* (OOK). The use of the word *keying* in describing modern digital modulation schemes can be traced to the use of the telegraph key in the earliest of digital communication systems. Transmissions that involve an on-off carrier frequency are often called *continuous wave* (CW) transmissions. This includes many radar signals and pulse-coded transmissions.

### 10.5.4 Bandwidth Considerations

The example message signal `SOS` in Morse code is nonperiodic and has a shape that cannot be described by a simple mathematical expression. Moreover, if the Morse code message is transmitted by hand, there would be not be a precise timing relationship among the elements of the signal as described above. In general, this is a pulse train with variable length pulses and variable distance between the pulses. In assessing the bandwidth requirements, it is useful to consider the worst-case scenario. Recall that the bandwidth of a signal varies inversely with the pulse width. If it is assumed that the *dot duration* is the shortest pulse component seen in this signal, then the worst-case bandwidth situation arises when a sequence of uninterrupted dots is transmitted, as shown in Figure 10.27. This corresponds to a squarewave with a period equal to two dot times (one dot plus one inter-dot space in each cycle). If the dot time is T msec, then the squarewave has period 2T and a fundamental frequency of 1/2T Hz. Proceeding with a rough approximation, assume that a skilled operator sends 50 words per minute. If a typical word is five characters or about 35 dot times, there will be 50×35 = 1750 dot times per minute, giving T = 60/1750 seconds, and 1/2T = 14 Hz. This low bandwidth is of advantage when using the Morse code; the result is directly related to the low rate of transmission, referred to as the *keying speed* for digital signals.

### 10.5.5 Receiving a Morse Code Signal

If the transmitted signal consists of an on-off radio frequency carrier, then it is necessary to convert the received signal into an on-off audio frequency for a telegraph operator to hear the received dots and dashes. A 100 Hz audio tone can be created if the received on-off carrier is mixed with a sinusoid offset by 1000 Hz (e.g., mix with $x(t) = \cos(2\pi(f_c + 1000)t))$ and lowpass filter the result. If a superheterodyne receiver is used, the mixing can occur and the output of the IF stage, such that the mixing signal is $x(t) = \cos(2\pi(f_{IF} + 1000)t)$, as in Figure 10.22. This is the identical beat frequency oscillator (BFO) method used to achieve carrier recovery in

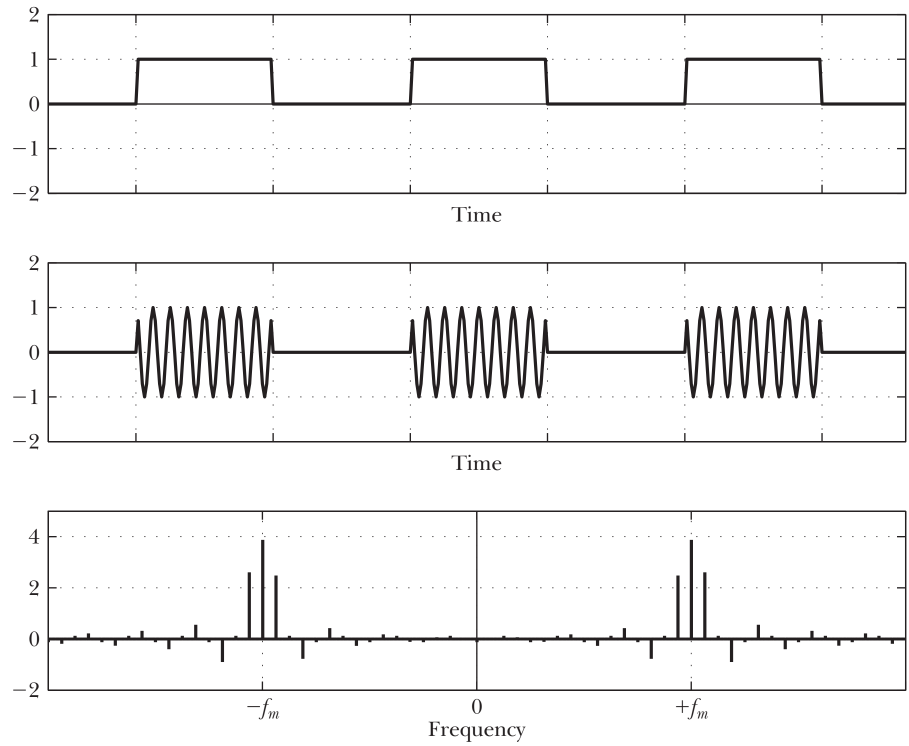



**FIGURE 10.27 Bandwidth of Periodic Digital Pulses—ASK** Amplitude shift keying (ASK) can be modelled as a squarewave multiplied by a cosine, shown here in time and frequency.

an AM-SSB receiver. Consequently, any SSB receiver will be suitable for Morse code reception, although special narrowband filters might be desirable if several Morse code signals are close together. Finally, note that if the frequency of the BFO can be adjusted (necessary for SSB), then the exact audio tone can vary from 1000 Hz depending on how the BFO frequency is varied.

# 10.6 Phase Shift Keying

As seen in Figure 10.24D, a message may also be conveyed on a carrier by varying the phase of a signal. Where the message has two states, as with single bits in a digital message, then two distinct phases would be defined (typically 0 and $\pi$ radians). Changing the phase of a cosine carrier by $\pi$ radians simply multiplies the waveform by $-1$, so the modulation can be accomplished by converting the digital pulses into a bipolar signal as in Figure 10.28.

This result must be compared to Figure 10.27, as the frequency domain appearance of the ASK and PSK signals differ only in the carrier component found at $f_m$ Hz. The bandwidth of the two signals is identical. Essentially, *the bandwidth is a function of the keying speed* and is independent of the modulation type used.

## 10.6.1 Differential Coding

One problem with PSK is that determination of the absolute phase of a signal depends on having a reference point (the origin) from which to begin. Given

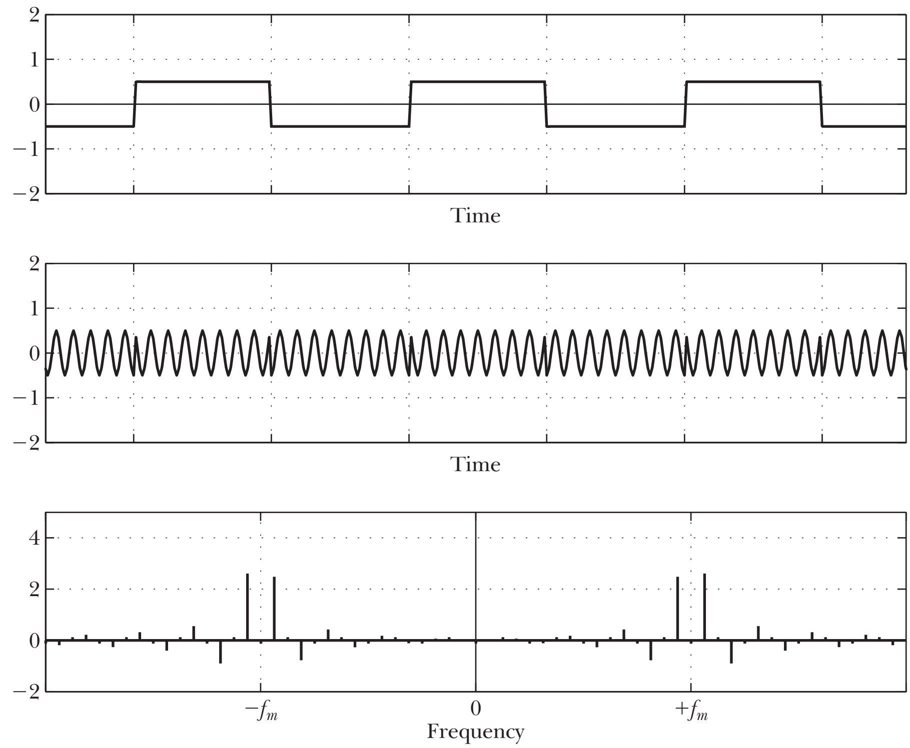



**FIGURE 10.28 Bandwidth of Periodic Digital Pulses—PSK** Phase shift keying (PSK) can be modelled as a bipolar squarewave multiplied by a cosine, shown here in time and frequency. Compare to Figure 10.27.

the PSK-modulated signal of Figure 10.28, the phase of a 1 or 0 bit cannot be determined upon reception, although it is clear that there are different phase states present in the received signal. While the absolute phase of a signal cannot be determined, *changes* in phase can be detected, and those changes are typically used to encode binary information during modulation. For example, a change of $\pi$ radians may represent the data bit 0, while no change could represent a binary 0.

The coding can be accomplished on a digital bitstream before modulation, so the waveforms in Figure 10.28 are unaffected; however, the digital input does not represent `010101010`$\cdots$ but rather `111111111`$\cdots$ where each change in phase conveys one bit of information.

## 10.6.2 Higher-Order Modulation Schemes

Given the fact that the bandwidth of a signal depends on the keying speed and not whether ASK or PSK is employed, it may be useful to incorporate *both* amplitude and phase variations in the same modulation scheme, as shown in Figure 10.29. For example, if a carrier cosine has two possible phase states and two possible amplitude states, then at each instant the signal can take on one of four possible phase/amplitude conditions. The advantage is that two bits of information can now be conveyed in the same bit period and the same bandwidth that previously carried one bit in either amplitude or phase separately. This technique is commonly employed to send a higher number of bits over a channel without increasing the bandwidth. It is just such a method that allows modems to operate over a telephone line at speeds varying from 1200 to 33,000 bits per second using the same limited telephone bandwidth.

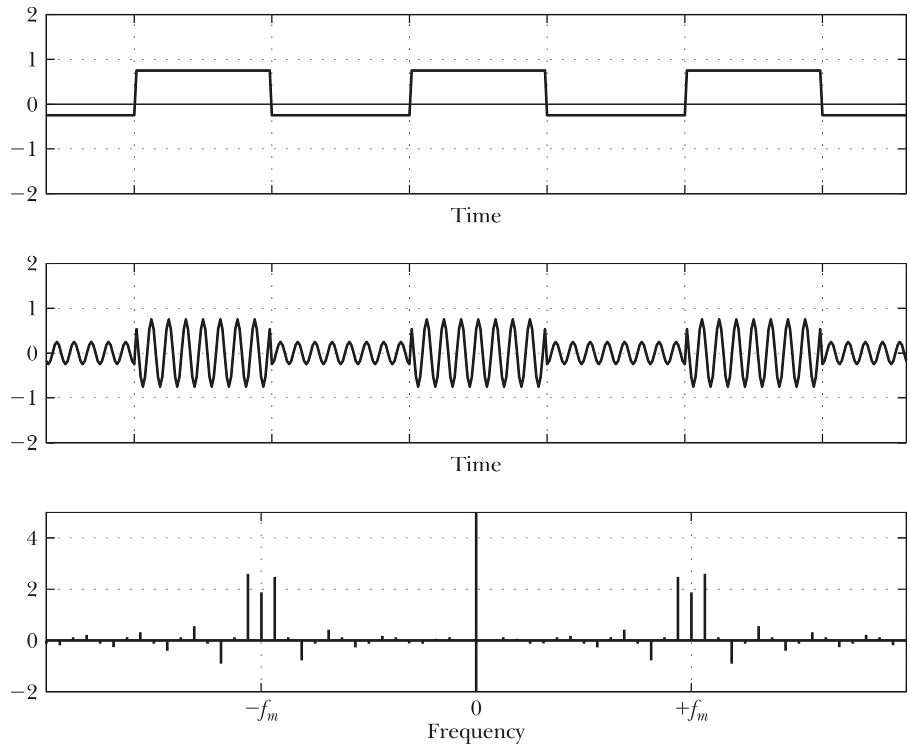



**FIGURE 10.29 Bandwidth of Periodic Digital Pulses—QAM** Both phase and amplitude variations can be used together without increasing the bandwidth of a modulated signal. Compare to Figure 10.27 and Figure 10.28.

# 10.7 Conclusions

An overview of communications signals illustrates the key role that signals theory plays in the transmission and reception of message signals, both digital and analog. Beginning with a *carrier signal* that is at a frequency able to travel through the channel of interest, various modulation techniques have been explored that impart a *message signal* onto the carrier such that the message also travels through the channel where it can be recovered in a distant receiver.

# End-of-Chapter Exercises

**10.1** Identify the frequency band (e.g., HF, VHF, UHF) used by each of the following radio signals.

**(a)** An aircraft navigation beacon on 326 kHz
**(b)** An AM broadcast station on 1260 kHz
**(c)** An audio baby monitor on 49.845 MHz
**(d)** An FM broadcast station on 99.5 MHz
**(e)** An aircraft control tower on 119.0 MHz
**(f)** A digital television station on 680 MHz
**(g)** A GPS navigation satellite on 1.57542 GHz
**(h)** A cellular telephone at 1.8502 GHz
**(i)** A microwave oven at 2.45 GHz
**(j)** A *Bluetooth* wireless link on 2.480 GHz
**(k)** A police speed radar gun at 24.15 GHz

**10.2** What is the wavelength of the carrier frequency of an FM radio broadcast signal at 102.3 MHz?

**10.3** What is the wavelength of the carrier frequency of an AM radio broadcast signal at 1010 kHz?

**10.4** A message signal $m(t)$ with 5 kHz bandwidth is modulated using double sideband AM on a carrier of 1 MHz. What is the bandwidth of the transmitted signal?

**10.5** A message signal $m(t)$ with 5 kHz bandwidth is modulated using single sideband AM on a carrier of 10 MHz. What is the bandwidth of the transmitted signal?

**10.6** Repeat the exercise of Figure 10.14 using a lower sideband (LSB) signal. Redraw the figure and state clearly how the apparent frequency of the demodulated signal varies if the mixing frequency is too high, or too low.

**10.7** An FM pocket radio has an intermediate frequency of 10.7 MHz. What mixing frequency is used if a station to be received is broadcasting on 101.3 MHz?

**10.8** A radio beacon transmits a 1 MHz carrier signal that is turned on for 1 second out of every 10 seconds. Accurately sketch the appearance of the transmitted signal in the frequency domain for frequencies close to 1 MHz.

**10.9** Consider a message signal $d(t) = \cos(8000\pi t)$. This message is to be transmitted on an AM broadcast radio station with a carrier frequency of $f_c = 970$ MHz using AM-TCDSB modulation.

**(a)** Accurately sketch the transmitted signal in the time domain.
**(b)** Accurately sketch the transmitted signal in the frequency domain.
**(c)** What is the total width of the modulated signal (Hz)?

**10.10** Consider a message signal $d(t) = \cos(8000\pi t)$. This message is to be transmitted on an AM broadcast radio station with a carrier frequency of $f_c = 970$ MHz using AM-TCDSB modulation.

**(a)** Accurately sketch the product signal in the time domain.
**(b)** Accurately sketch the product signal in the frequency domain.
**(c)** How can the original message be recovered from the product signal?

An AMDSB-TC signal may be modelled in MATLAB, beginning as:

```
t = 1 : 512;               % time axis
c = cos(2 * pi * t/8)      % carrier
m = cos(2 * pi * t/128)    % message
tx = (1 + m). * c; % modulated TC-DSB
```

**10.11** Use MATLAB to plot the above signal in the time domain and in the frequency domain near the carrier frequency. Identify the carrier and sidebands.

**10.12** Use MATLAB to apply a diode (half-wave rectifier) to the above signal and plot the resulting signal in the time domain and in the frequency domain.

**10.13** Identify the original message in the above demodulated signal and describe the type of filter that would best isolate the message.

**10.14** Review the example in Section 10.4.1 and sketch the signals in the frequency domain. With the aid of your sketch, identify the source of the squeal.

**10.15** On-off keying (OOK) may be modelled by multiplying a cosine carrier by a squarewave (0 to 1 V) representing some binary data. Binary phase shift keying (BPSK) may be modelled by a squarewave ($-1$ V to 1 V). Describe the difference between these two modulated signals in the frequency domain.

APPENDIX A

# The Illustrated Fourier Transform

In the pages that follow, a series of *Fourier transform pairs* are presented, each showing a function of time $s(t)$ on the left and the corresponding Fourier transform function of frequency $S(f)$ on the right, where:

$$s(t) \xleftrightarrow{\mathcal{F}} S(f) = \int_{-\infty}^{+\infty} s(t)\; e^{-j2\pi ft}\; dt$$

In each graph, real components are shown with a solid line, while imaginary components are plotted with a dashed line. The time- and frequency-domain sketches may be switched directly with at most a change of sign in the imaginary component, since:

$$S(f) \xleftrightarrow{\mathcal{F}} s(t) = \int_{-\infty}^{+\infty} S(f)\; e^{+j2\pi ft}\; df$$

This illustrated list provides a ready reference for common Fourier transform pairs from which more complicated Fourier transforms may be derived by applying the appropriate rules, often by inspection and without the need for further computations.

As a study guide, careful examination of each transform pair confirms and reveals a wealth of information about the general properties characterizing the Fourier transform, particularly those relating to odd and even functions, real and imaginary components, and both periodic and nonperiodic signals. For example, in each case, expect that the value at origin in one domain equals the area in the other.

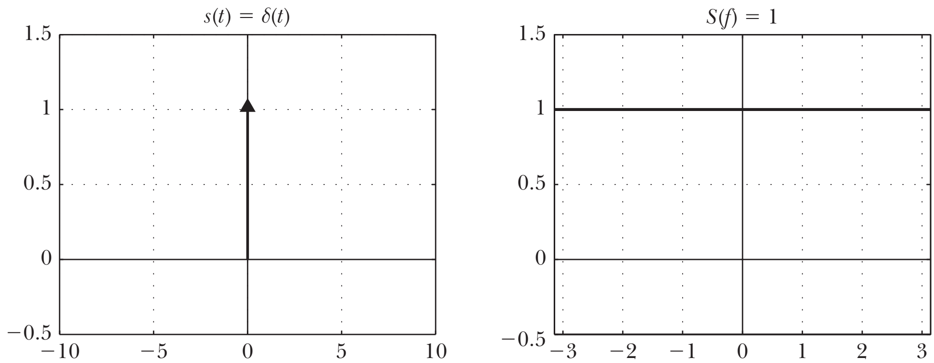



**FIGURE A.1** $\delta(t) \Leftrightarrow 1$.

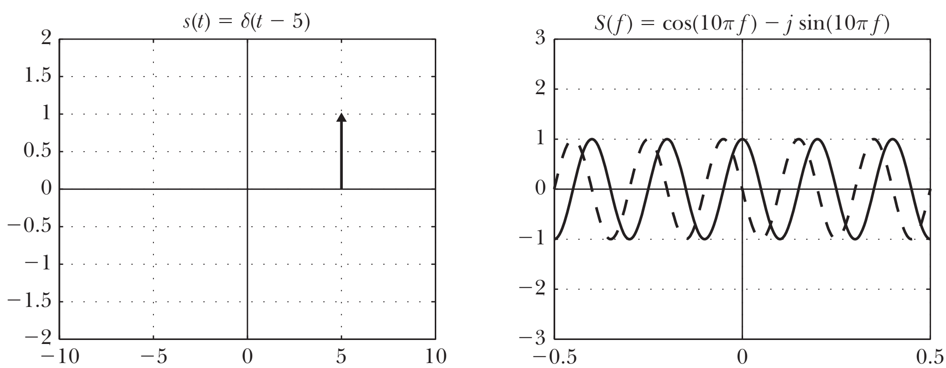



**FIGURE A.2** $\delta(t-5) \Leftrightarrow \cos(10\pi f) - j\sin(10\pi f)$.

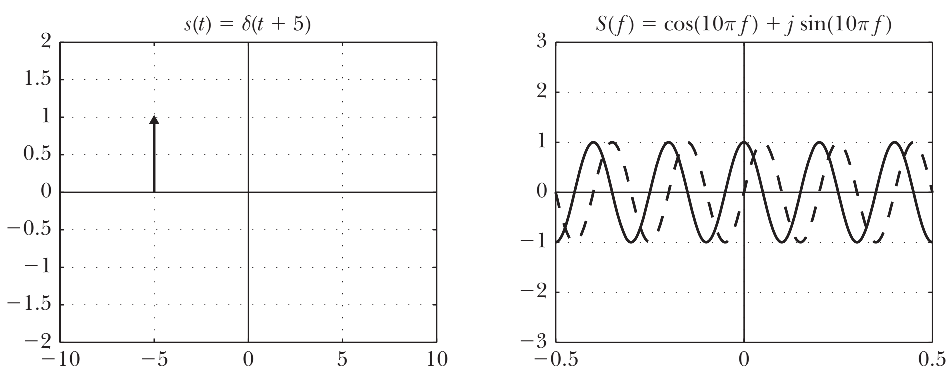



**FIGURE A.3** $\delta(t+5) \Leftrightarrow \cos(10\pi f) + j\sin(10\pi f)$.

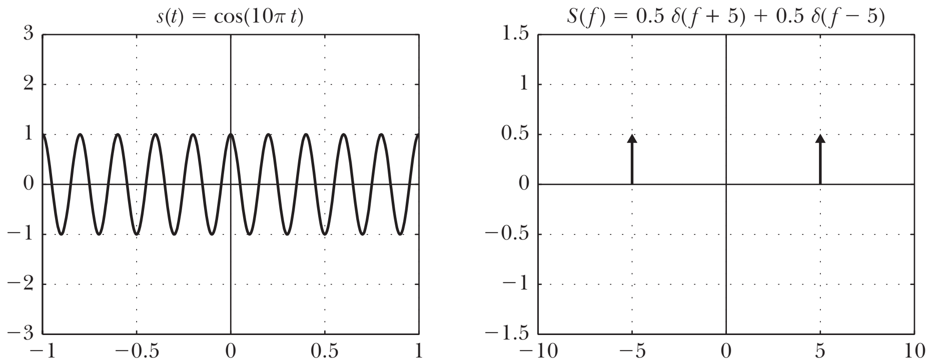



**FIGURE A.4** $\cos(10\pi\, t) \Leftrightarrow \frac{1}{2}\delta(f+5) + \frac{1}{2}\delta(f-5)$.

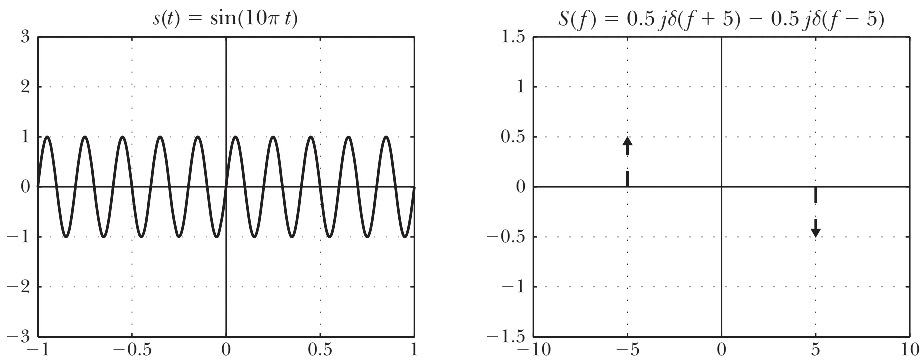



**FIGURE A.5** $\sin(10\pi\, t) \Leftrightarrow j\frac{1}{2}\delta(f+5) - j\frac{1}{2}\delta(f-5)$.

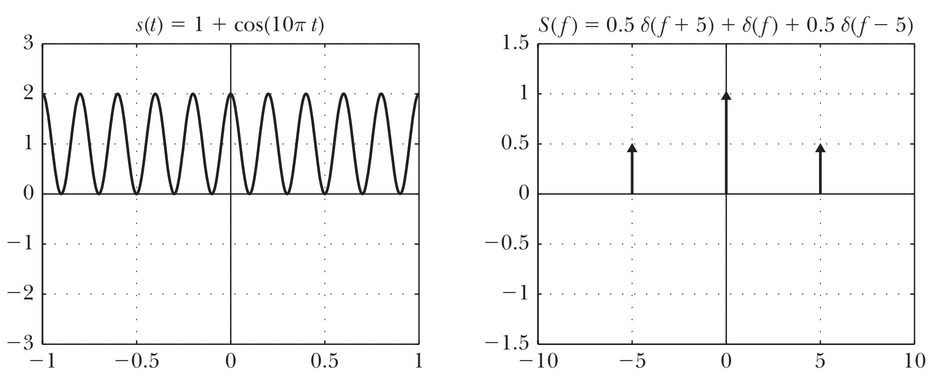



**FIGURE A.6** $1 + \cos(10\pi\, t) \Leftrightarrow \frac{1}{2}\delta(f+5) + \delta(f) + \frac{1}{2}\delta(f-5)$.

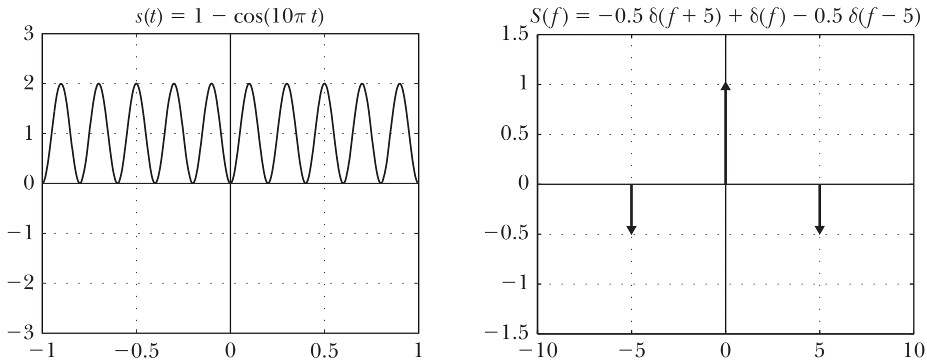



**FIGURE A.7** $1 - \cos(10\pi\, t) \Leftrightarrow -\frac{1}{2}\delta(f+5) + \delta(f) - \frac{1}{2}\delta(f-5)$.

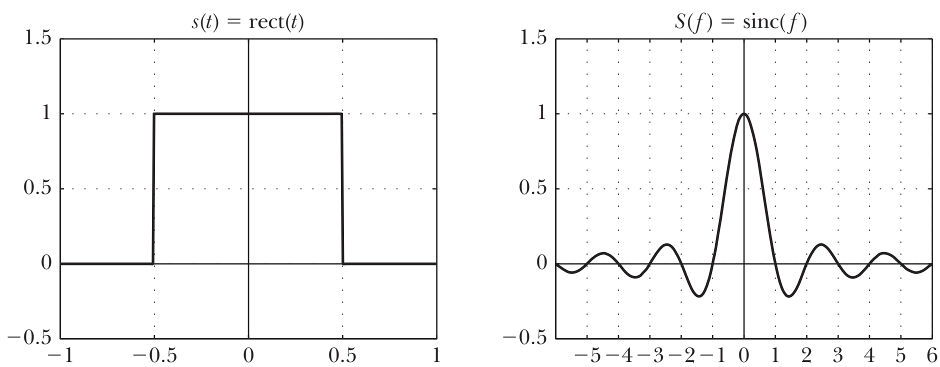



**FIGURE A.8** $\text{rect}(t) \Leftrightarrow \text{sinc}(f)$: Compare to Fig. A.13.

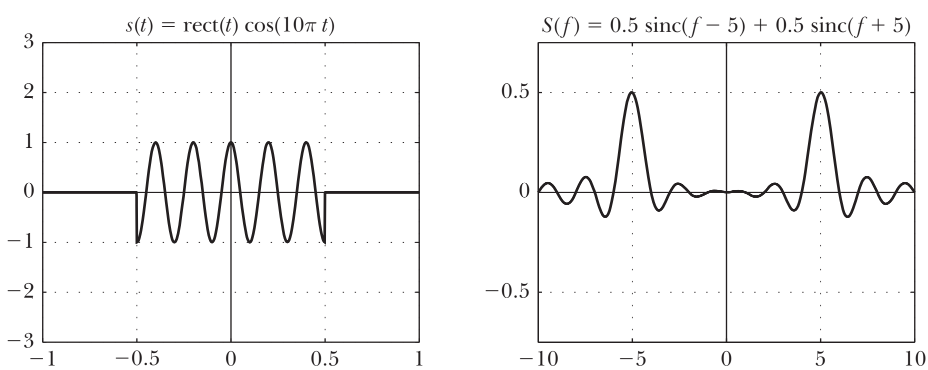



**FIGURE A.9** $\text{rect}(t)\cos(10\pi\, t) \Leftrightarrow \frac{1}{2}\text{sinc}(f+5) + \frac{1}{2}\text{sinc}(f-5)$.

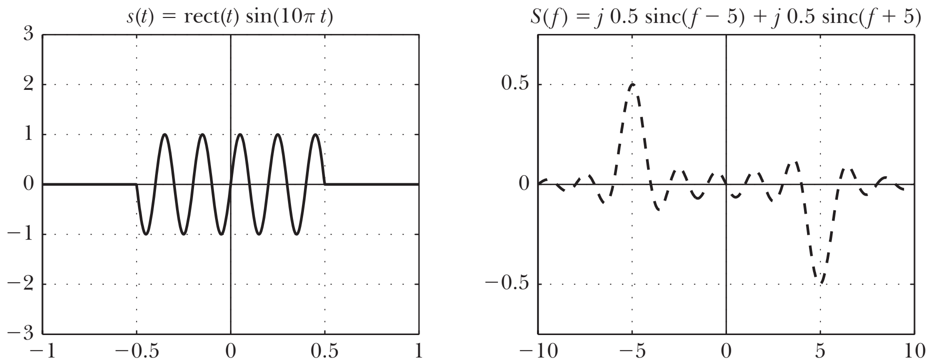



**FIGURE A.10** $\mathrm{rect}(t)\sin(10\pi\, t) \Leftrightarrow j\frac{1}{2}\mathrm{sinc}(f+5) - j\frac{1}{2}\mathrm{sinc}(f-5)$.

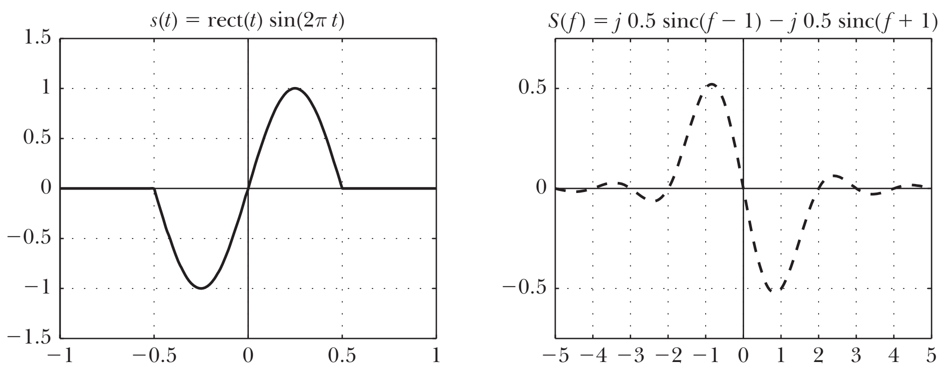



**FIGURE A.11** $\mathrm{rect}(t)\sin(2\pi\, t) \Leftrightarrow j\frac{1}{2}\mathrm{sinc}(f+1) - j\frac{1}{2}\mathrm{sinc}(f-1)$.

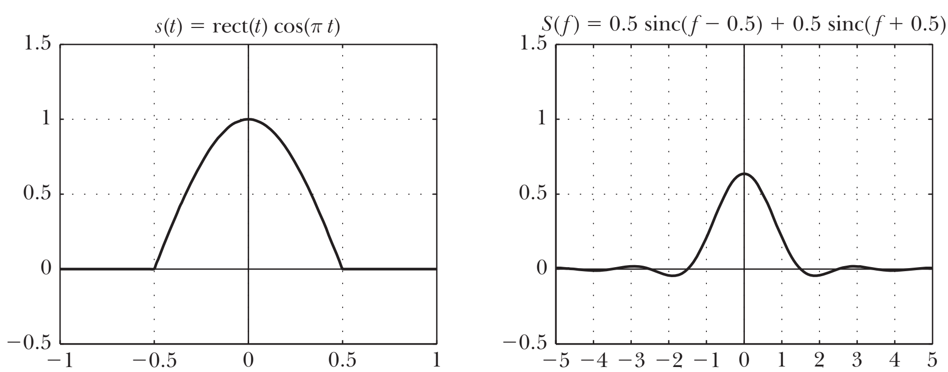



**FIGURE A.12** $\mathrm{rect}(t)\cos(\pi t) \Leftrightarrow \frac{1}{2}\mathrm{sinc}(f+\frac{1}{2}) + \frac{1}{2}\mathrm{sinc}(f-\frac{1}{2})$.

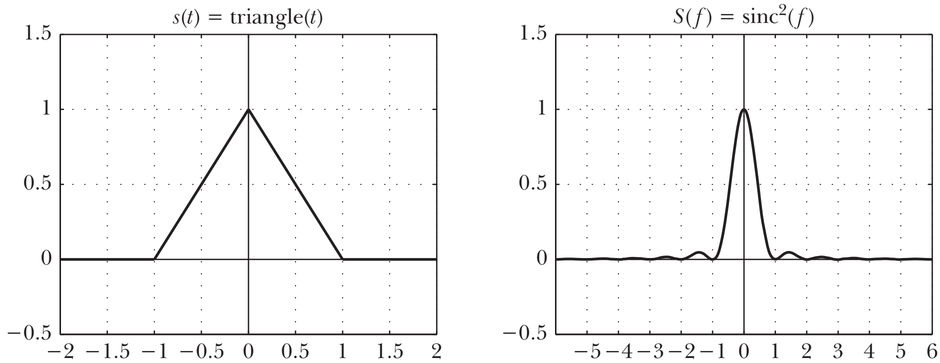



**FIGURE A.13** $\mathrm{rect}(t) * \mathrm{rect}(t) \Leftrightarrow \mathrm{sinc}(f) \times \mathrm{sinc}(f)$: Compare to Figure A.8.

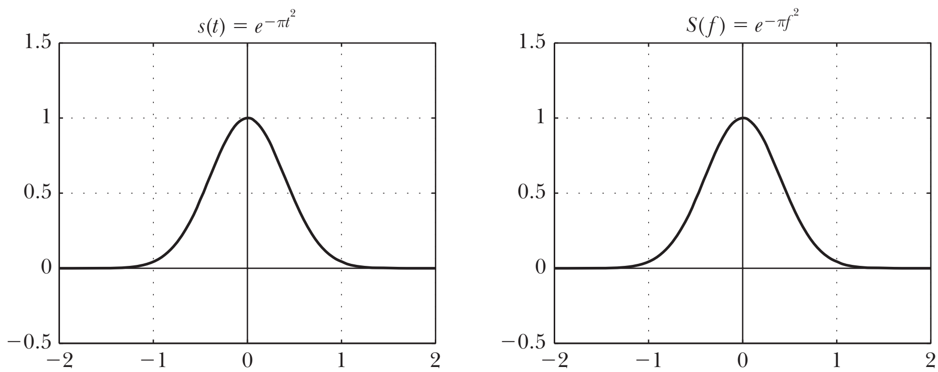



**FIGURE A.14** Gaussian: $e^{-\pi t^2} \Leftrightarrow e^{-\pi f^2}$: The two curves are identical.

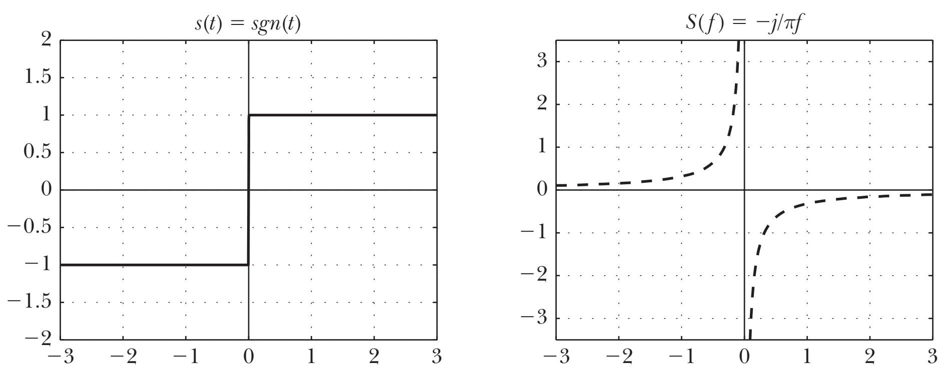



**FIGURE A.15** $sgn(t) \Leftrightarrow j\frac{1}{\pi f}$: Compare to Figure A.16.

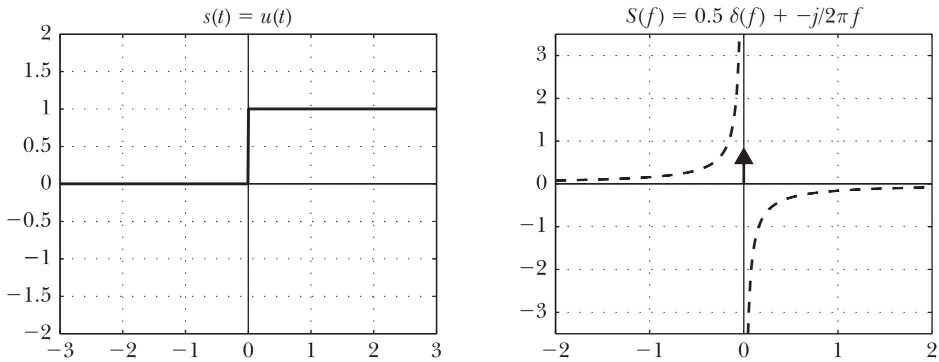



**FIGURE A.16** $u(t) \Leftrightarrow \frac{1}{2}\delta(f) + j\frac{1}{2\pi f}$: Compare to Figure A.15.

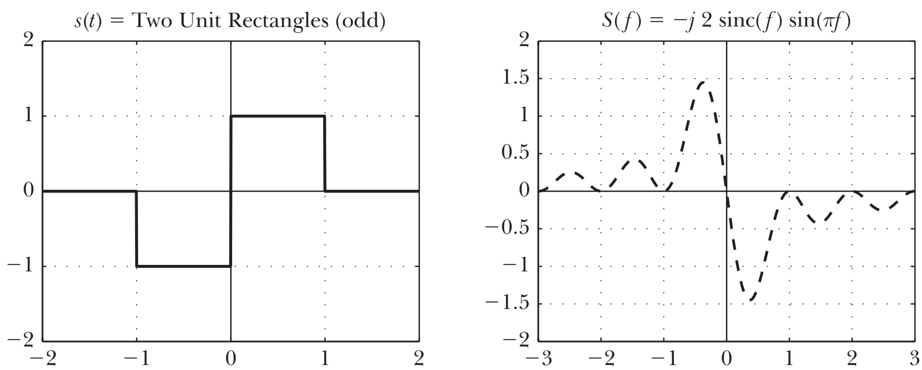



**FIGURE A.17** $\mathrm{rect}(t - \frac{1}{2}) - \mathrm{rect}(t + \frac{1}{2}) \Leftrightarrow -j\,2\,\mathrm{sinc}(f)\sin(\pi f)$.

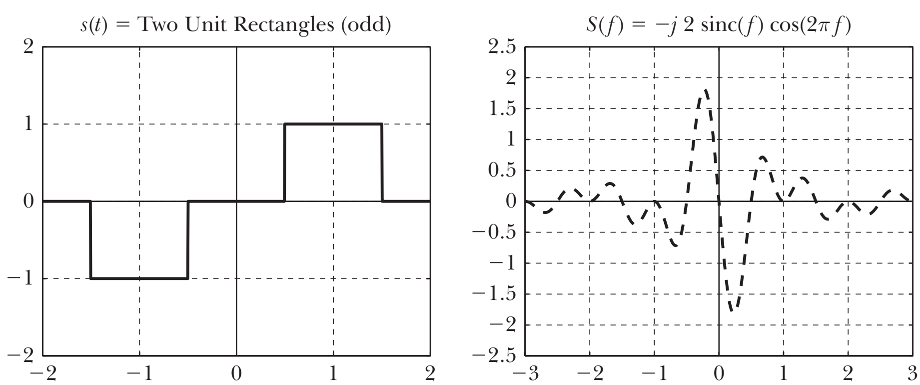



**FIGURE A.18** $\mathrm{rect}(t - 1) - \mathrm{rect}(t + 1) \Leftrightarrow -j\,2\,\mathrm{sinc}(f)\,\sin(2\pi f)$.

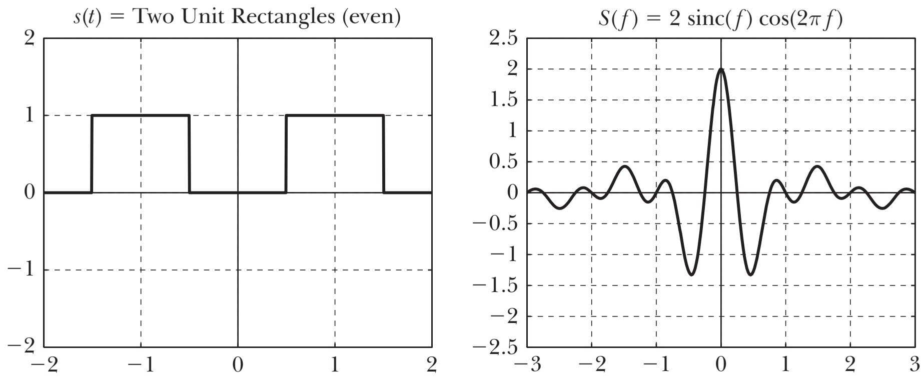



**FIGURE A.19** $\text{rect}(t-1)+\text{rect}(t+1) \Leftrightarrow 2\,\text{sinc}(f)\cos(2\pi f)$.

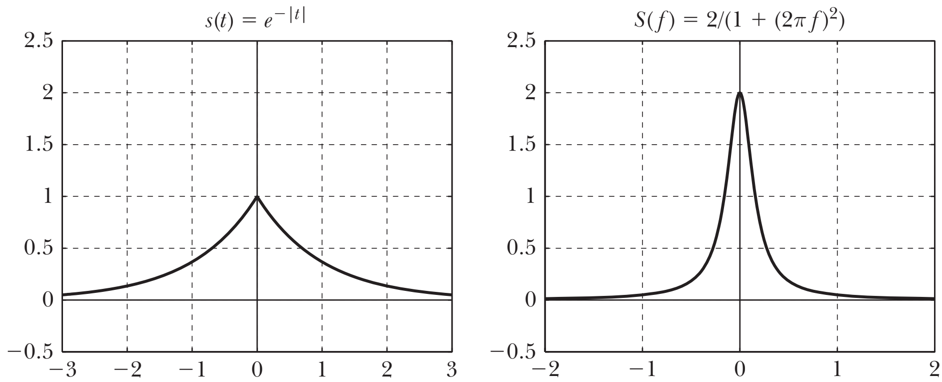



**FIGURE A.20** $e^{-|t|} \Leftrightarrow \frac{2}{1+(2\pi f)^2}$.

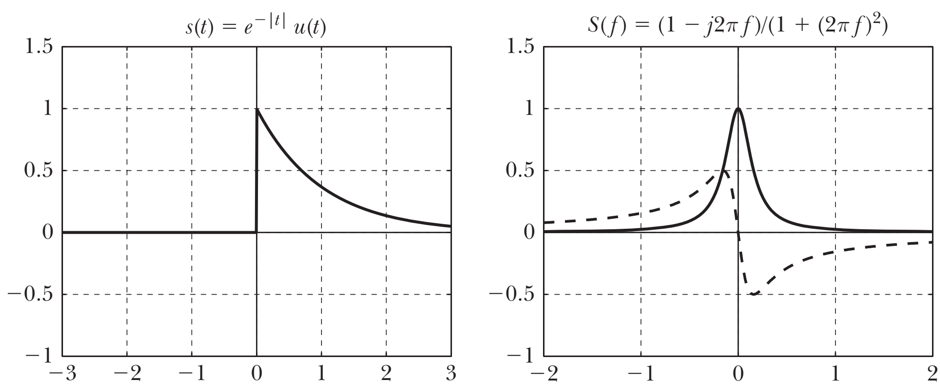



**FIGURE A.21** $e^{-|t|}u(t) \Leftrightarrow \frac{1-j2\pi f}{1+(2\pi f)^2}$.

# APPENDIX B

# The Illustrated Laplace Transform

In the pages that follow, a series of *Laplace transform pairs* are presented, each showing a function of time $h(t)$ on the left and the corresponding pole-zero diagram for $H(s)$ on the right, where:

$$h(t) \xleftrightarrow{\mathcal{L}} H(s) = \int_{-\infty}^{+\infty} h(t)e^{-st}\, dt$$

Each of the $h(t)$ may be considered as the impulse response of an LTI system. In each pole-zero diagram, poles are shown with an **X** while zeros are plotted with an **O**. The region of convergence shown shaded in each pole-zero diagram is generally the entire *z*-plane to the right of the rightmost pole.

The inverse Laplace transform $h(t)$ is only generally found from the pole-zero diagram of $H(s)$ as constant values or gain factors are not visible when poles and zeros are plotted. Similarly, the sum of two pole-zero diagrams is *not* generally the sum of the corresponding Laplace transforms. In particular, the presence of zeros will be affected, as seen when sine and cosine are added together in Figure B.8 and in the summation of Figure B.12.

The importance of the pole-zero diagram to system stability can also be seen, where poles on the right-hand plane are associated with divergent impulse response functions. If the region of convergence includes the line $\sigma = 0$, then the related Fourier transform can be found along this line. Otherwise, the signal $h(t)$ has a Laplace transform but does not have a Fourier transform. The line $\sigma = 0$ corresponds to the unit circle in the *z*-transform.

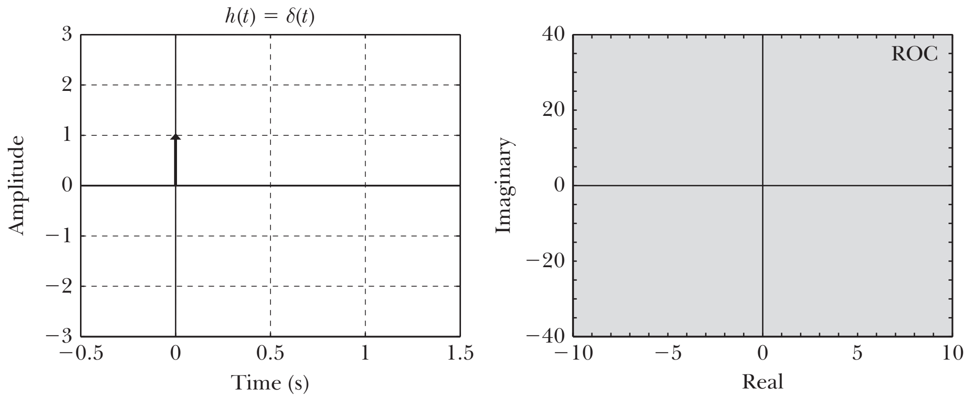



**FIGURE B.1** $\delta(t) \Leftrightarrow 1$.

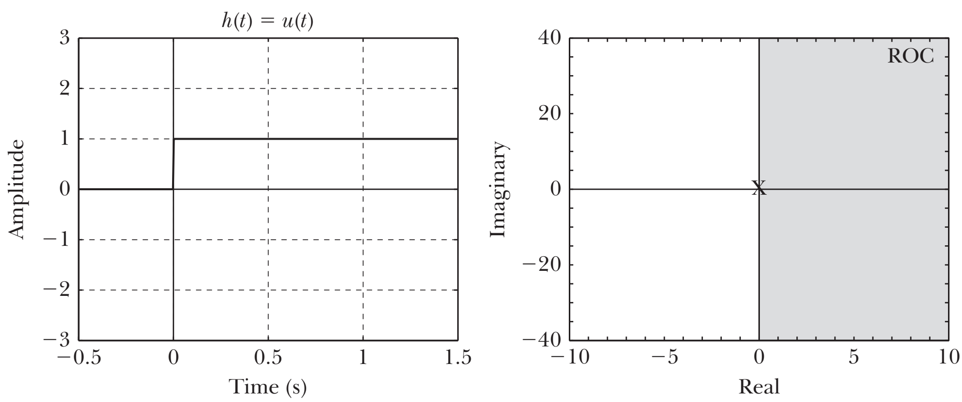



**FIGURE B.2** $u(t) \Leftrightarrow \frac{1}{s}$.

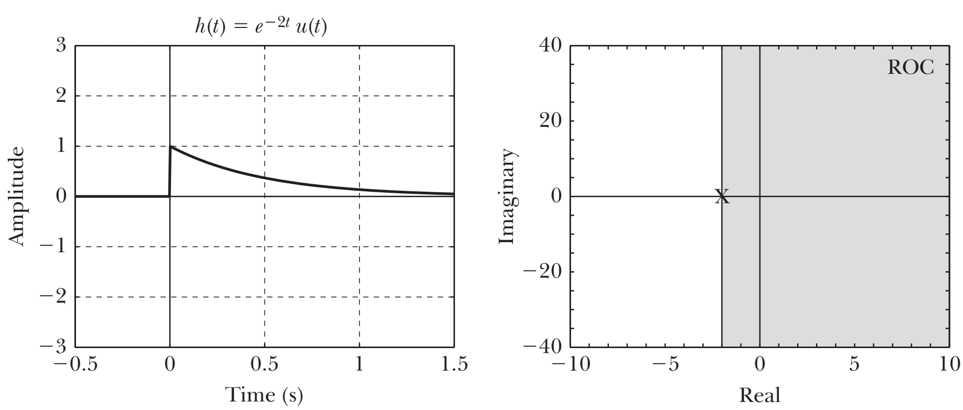



**FIGURE B.3** $e^{-2t}\,u(t) \Leftrightarrow \frac{1}{s+2}$.

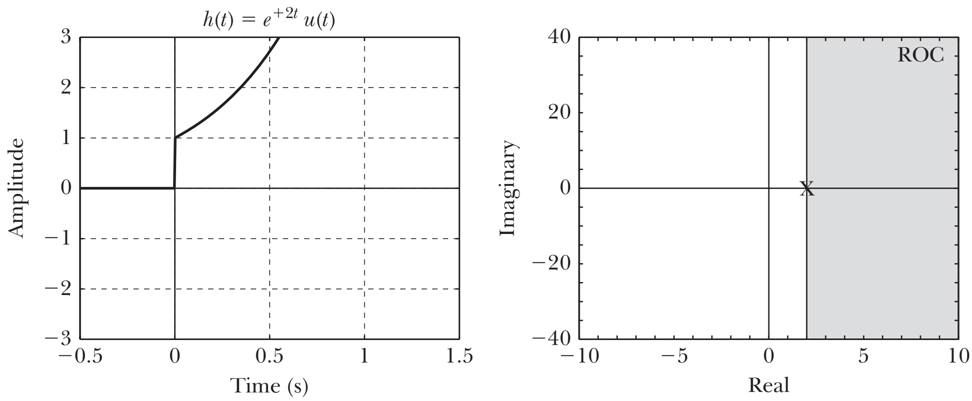



**FIGURE B.4** $e^{+2t}\,u(t) \Leftrightarrow \frac{1}{s-2}$.

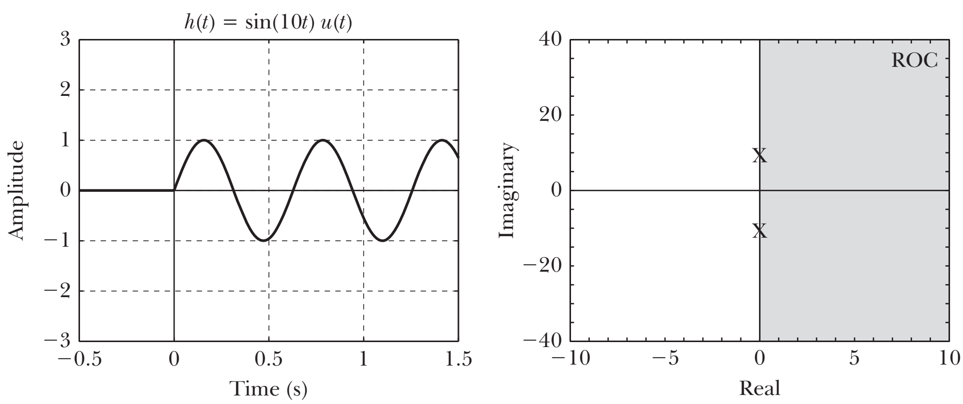



**FIGURE B.5** $\sin(10t)\,u(t) \Leftrightarrow \frac{10}{s^2+10^2}$.

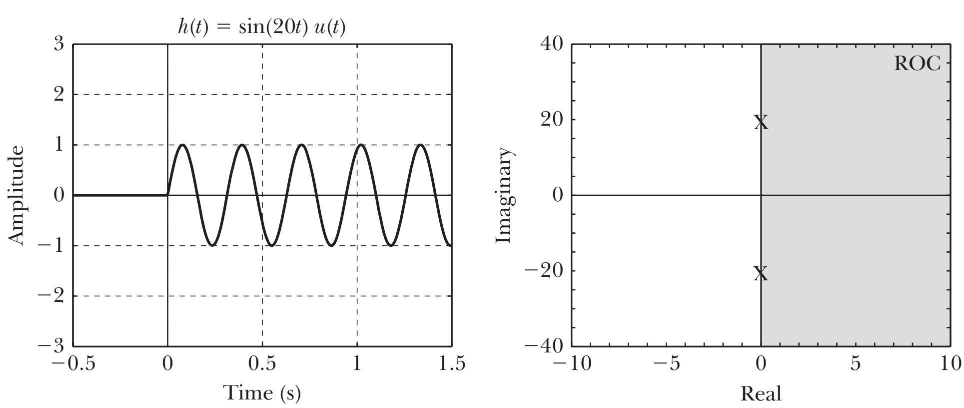



**FIGURE B.6** $\sin(20t)\,u(t) \Leftrightarrow \frac{20}{s^2+20^2}$.

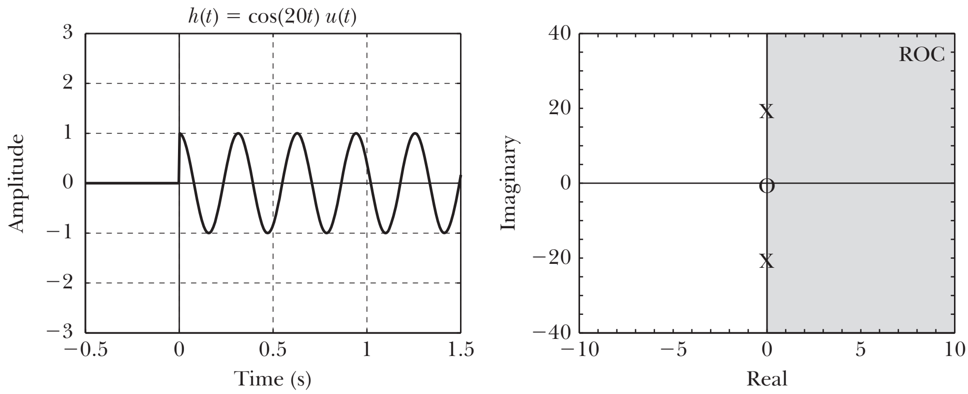



**FIGURE B.7** $\cos(20t)\; u(t) \Leftrightarrow \frac{s}{s^2+20^2}$.

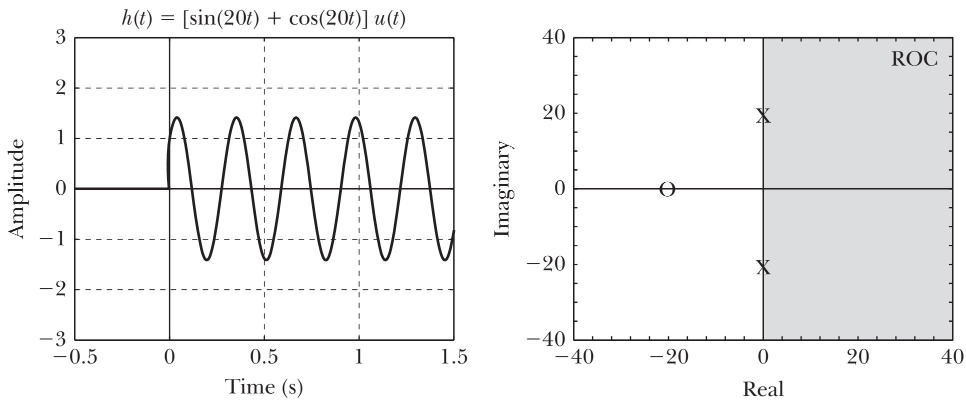



**FIGURE B.8** $[\sin(20t) + \cos(20t)]\; u(t) \Leftrightarrow \frac{s+20}{s^2+20^2}$.

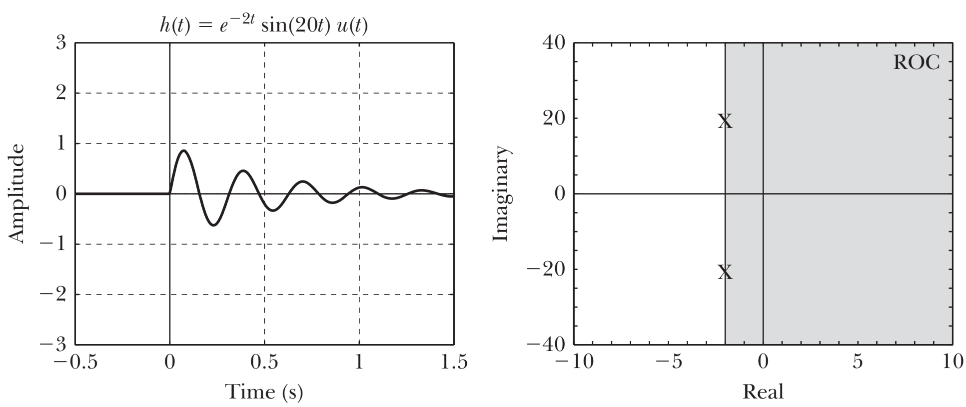



**FIGURE B.9** $e^{-2t} \sin(20t)\; u(t) \Leftrightarrow \frac{20}{(s+2)^2+20^2}$.

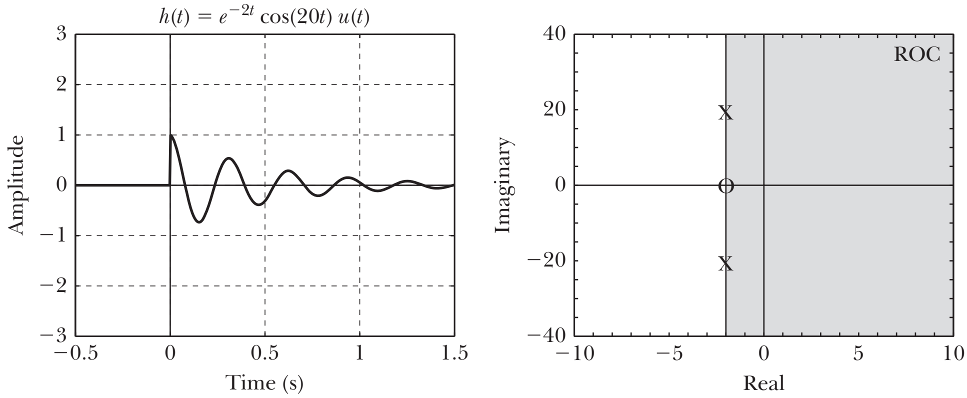



**FIGURE B.10** $e^{-2t} \cos(20t)\, u(t) \Leftrightarrow \frac{s+2}{(s+2)^2+20^2}$.

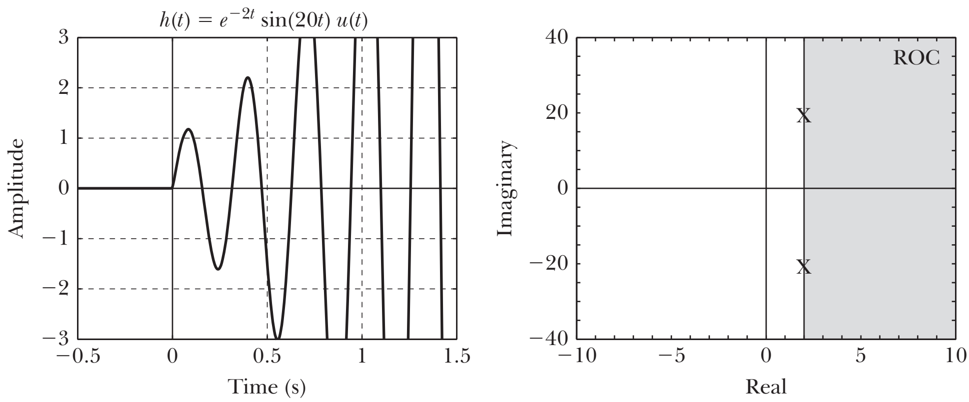



**FIGURE B.11** $e^{+2t} \sin(20t)\, u(t) \Leftrightarrow \frac{20}{(s-2)^2+20^2}$.

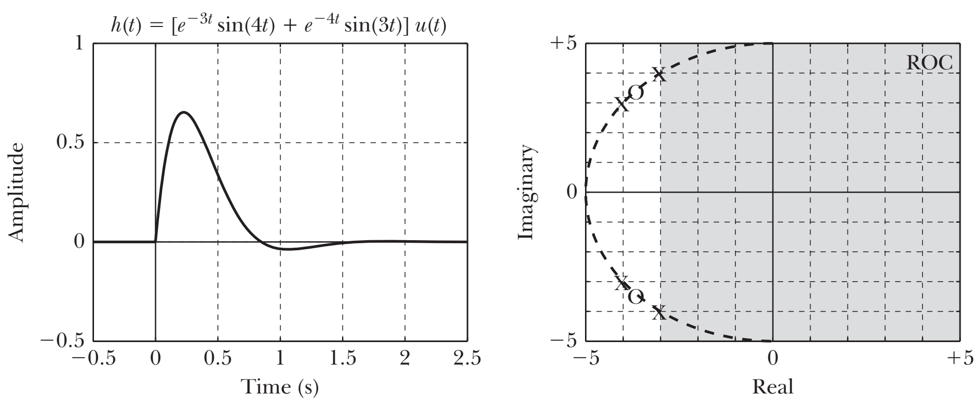



**FIGURE B.12** $[e^{-3t} \sin(4t) + e^{-4t} \sin(3t)]\, u(t) \Leftrightarrow \frac{4}{(s-3)^2+4^2} + \frac{3}{(s-4)^2+3^2}$.

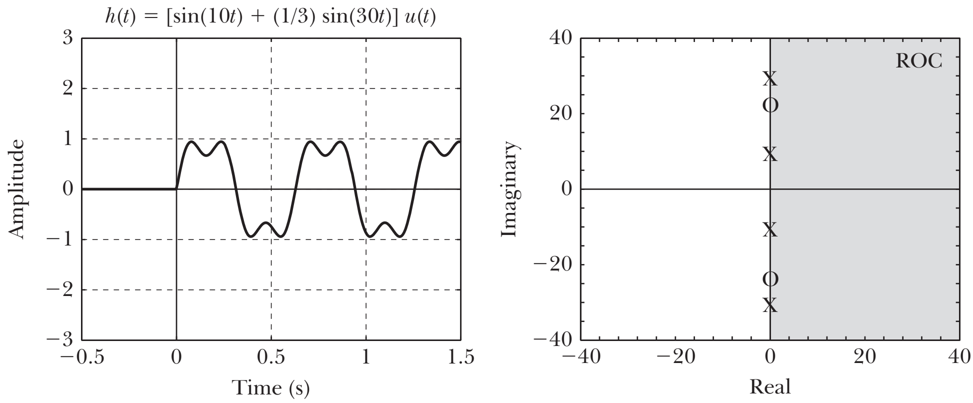



**FIGURE B.13** $[\sin(10t) + \frac{1}{3}\sin(30t)]\ u(t) \Leftrightarrow \frac{10}{[s^2+10^2]} + \frac{1}{3}\,\frac{30}{[s^2+30^2]}$.

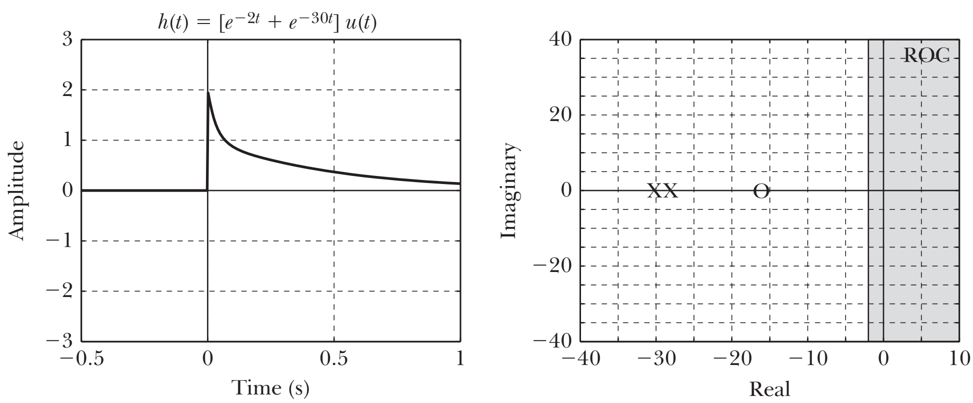



**FIGURE B.14** $[e^{-2t} + e^{-30t}]\ u(t) \Leftrightarrow \frac{1}{[s+2]} + \frac{1}{[s+30]}$.

# APPENDIX C

# The Illustrated *z*-Transform

In the pages that follow, a series of *z-transform pairs* are presented, each showing a discrete function $h[n]$ on the left and the corresponding pole-zero diagram for $H(z)$ on the right, where:

$$h[n] \overset{\mathcal{Z}}{\longleftrightarrow} H(z) = \sum_{n=0}^{+\infty} z^n \, h[n]$$

Each of the $h[n]$ may be considered as the impulse response of a discrete LTI system. In each pole-zero diagram, poles are shown with an **X** while zeros are plotted with an **O**. The region of convergence shown shaded in each pole-zero diagram is generally the entire *z*-plane outside of a circle (centered on the origin) that encompasses all the poles.

The inverse *z*-transform $h[n]$ is only generally found from the pole-zero diagram of $H(z)$ as constant values or gain factors are not visible when poles and zeros are plotted. Similarly, the sum of two pole-zero diagrams is *not* generally the sum of the corresponding *z*-transforms. Because the *z*-transform takes place on discrete (sampled) signals, the angular position of poles and zeros around the origin varies with sampling rate for a given frequency component.

The importance of the pole-zero diagram to system stability can also be seen, where poles outside the unit circle are associated with divergent impulse response functions. If the region of convergence includes the circle $r = 1$, then the related discrete time Fourier transform can be found around its circumference. Otherwise the discrete signal $h[n]$ has a *z*-transform but does not have a (discrete time) Fourier transform. The unit circle corresponds to the vertical line $\sigma = 0$ in the Laplace transform.

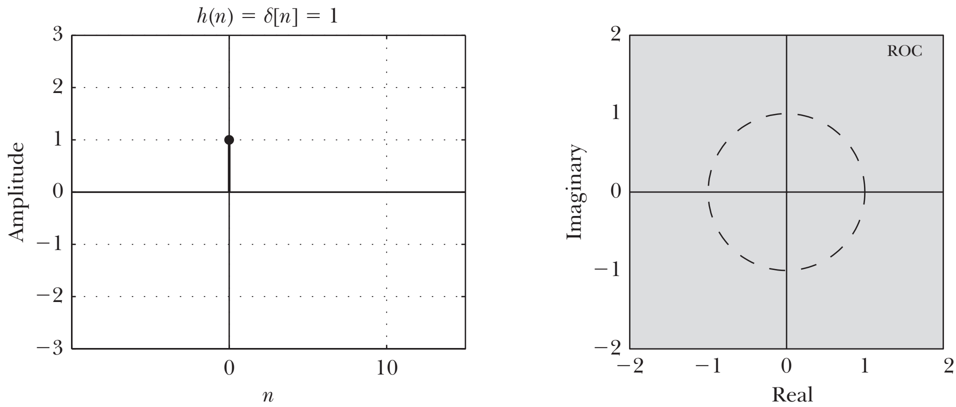



**FIGURE C.1** $\delta[n] \Leftrightarrow 1$.

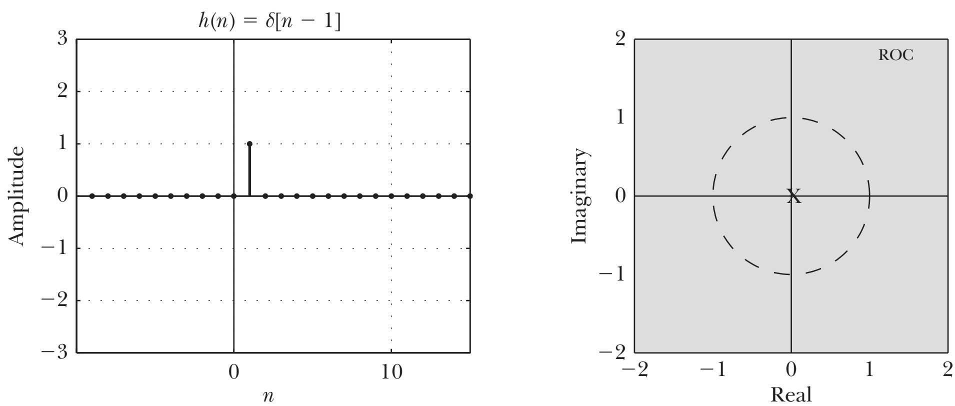



**FIGURE C.2** $\delta[n-1] \Leftrightarrow z^{-1}$.

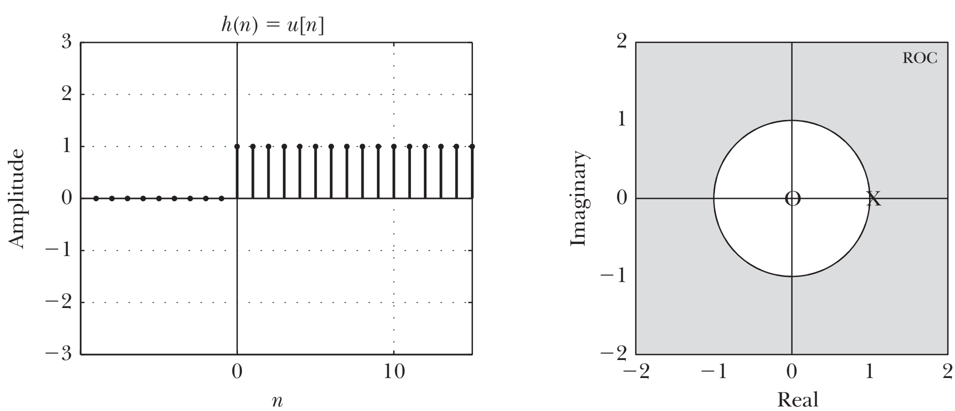
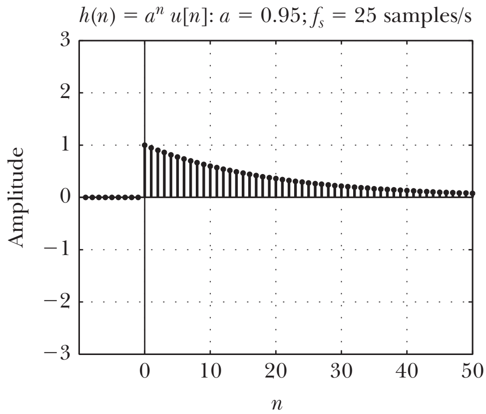



**FIGURE C.3** $u[n] \Leftrightarrow \frac{z}{z-1}$.



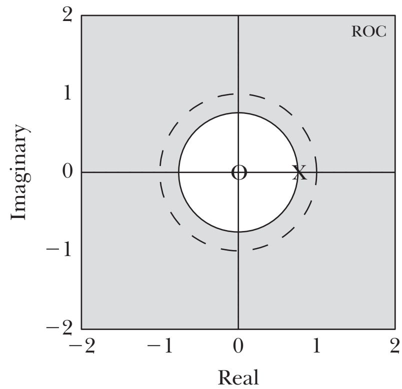
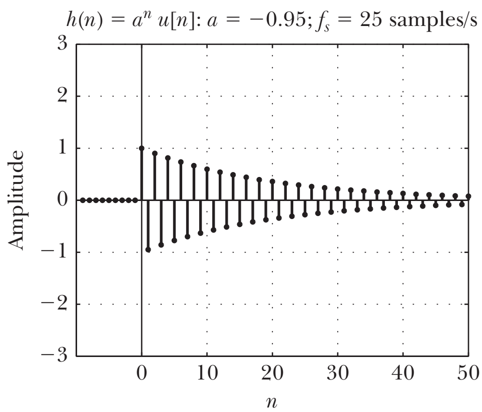




**FIGURE C.4** $a^n\, u[n] \Leftrightarrow \frac{z}{z-a}$; $a = 0.95$.



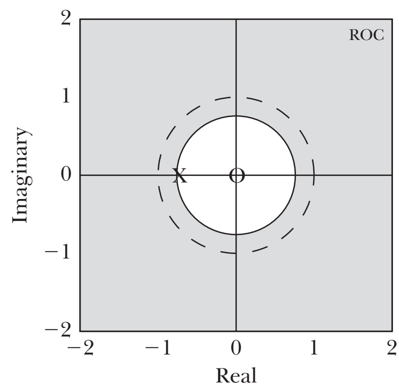
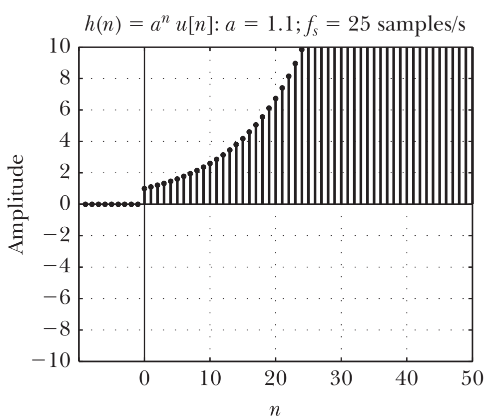




**FIGURE C.5** $a^n\, u[n] \Leftrightarrow \frac{z}{z-a}$; $a = -0.95$.



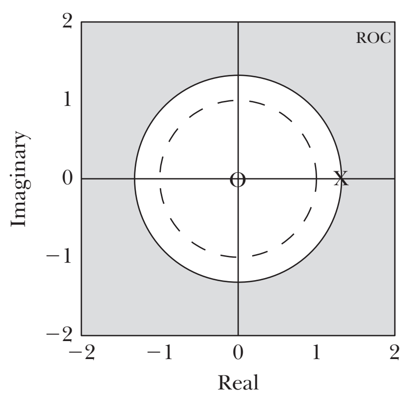




**FIGURE C.6** $a^n\, u[n] \Leftrightarrow \frac{z}{z-a}$; $a = 1.1$.

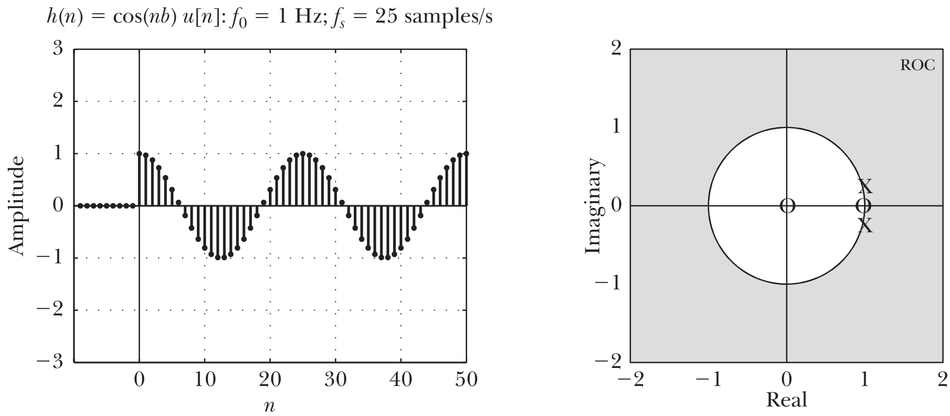



**FIGURE C.7** $\cos(nb)\,u[n] \Leftrightarrow \frac{z(z-\cos(b))}{z^2-2z\cos(b)+1}$; $b = 1$; $f_s = 25$ samples/s.

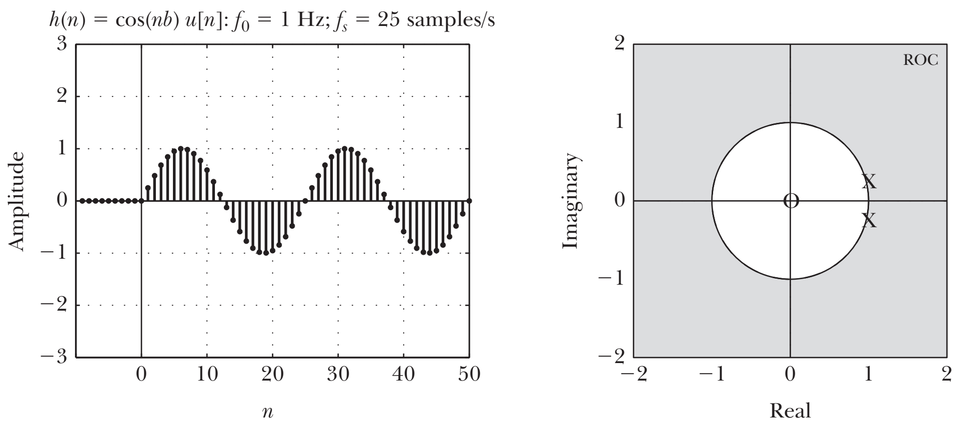



**FIGURE C.8** $\sin(nb)\,u[n] \Leftrightarrow \frac{z\sin(b)}{z^2-2z\cos(b)+1}$; $b = 1$; $f_s = 25$ samples/s.

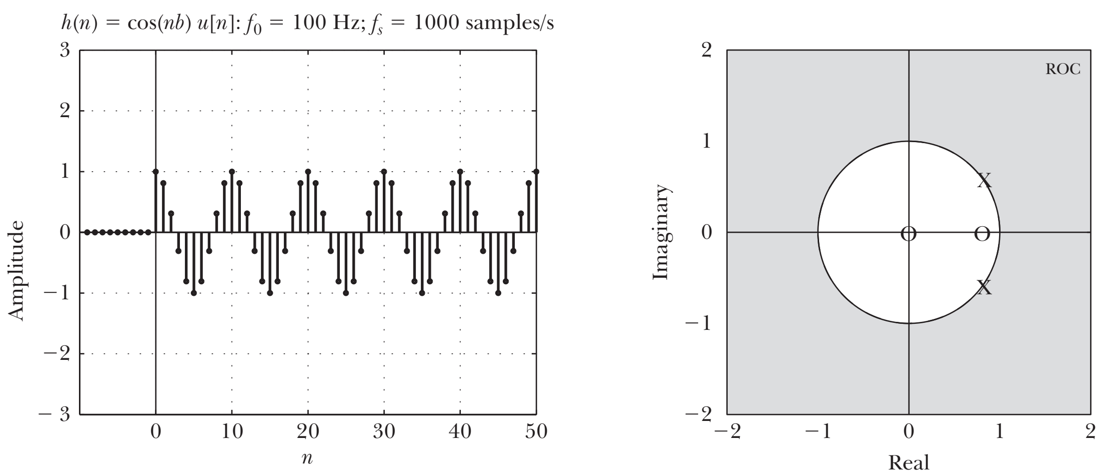



**FIGURE C.9** $\cos(nb)\,u[n] \Leftrightarrow \frac{z(z-\cos(b))}{z^2-2z\cos(b)+1}$; $b = 1/100$; $f_s = 1000$ samples/s.

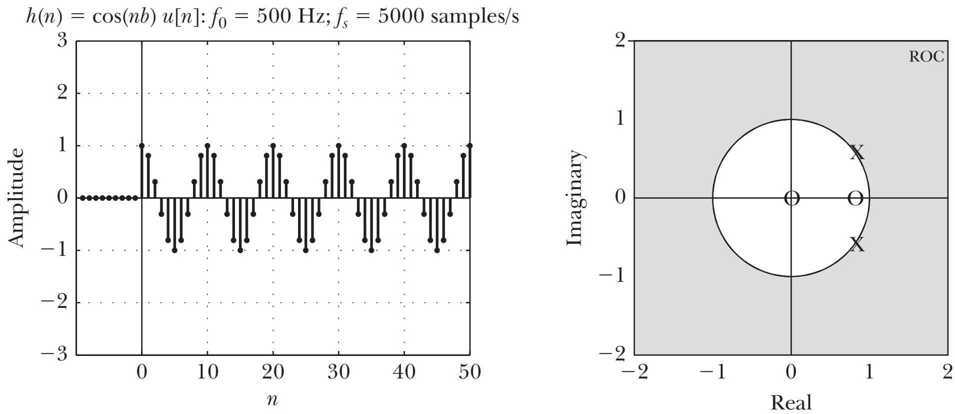



**FIGURE C.10** $\cos(nb)\,u[n] \Leftrightarrow \frac{z(z-\cos(b))}{z^2-2z\cos(b)+1}$; $b = 1/500$; $f_s = 5000$ samples/s.

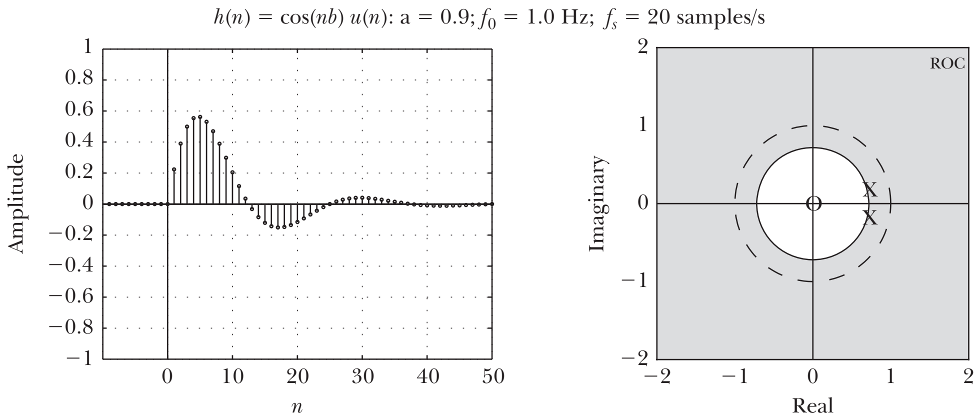



**FIGURE C.11** $a^n \sin(bn)\,u[n] \Leftrightarrow \frac{az\sin(b)}{s^2+2az\cos(b)+a^2}$; $a = 0.9$.

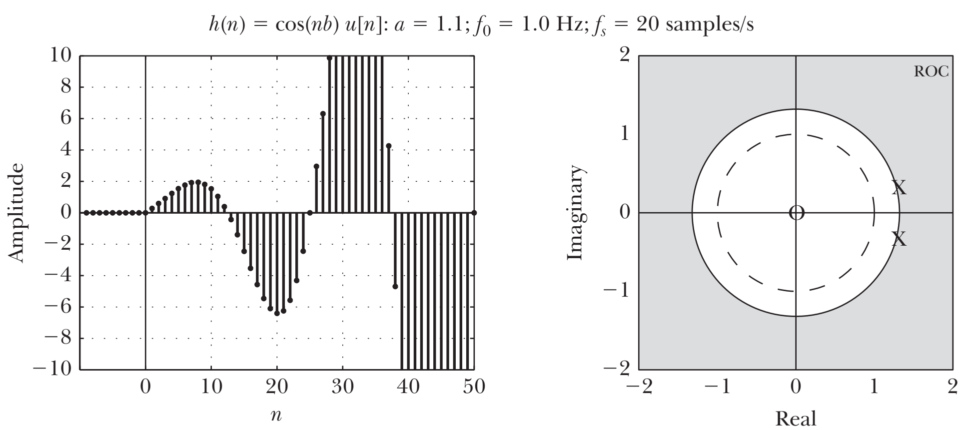



**FIGURE C.12** $a^n \sin(bn)\,u[n] \Leftrightarrow \frac{az\sin(b)}{s^2+2az\cos(b)+a^2}$; $a = 1.1$.

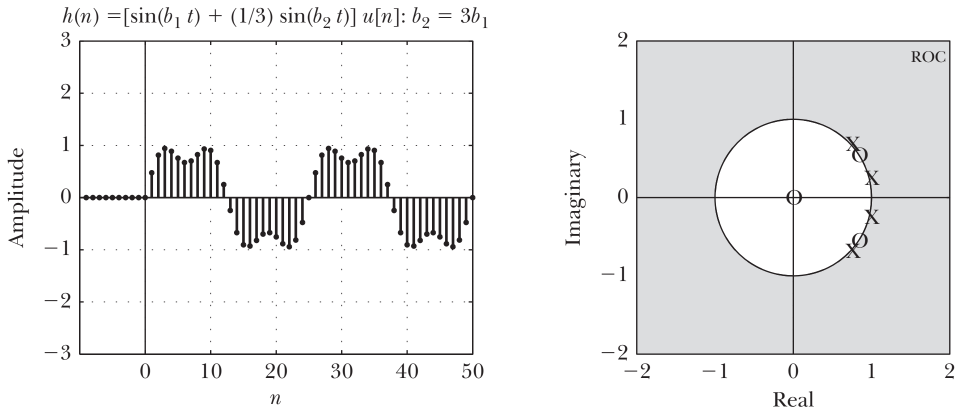



**FIGURE C.13** $[\sin(b_1 n) + \frac{1}{3}\sin(b_2 n)]\ u[n] \Leftrightarrow \frac{z\sin(b_1)}{z^2-2z\cos(b_1)+1} + \frac{1}{3}\frac{z\sin(b_2)}{z^2-2z\cos(b_2)+1};\ b_2 = 3b_1.$

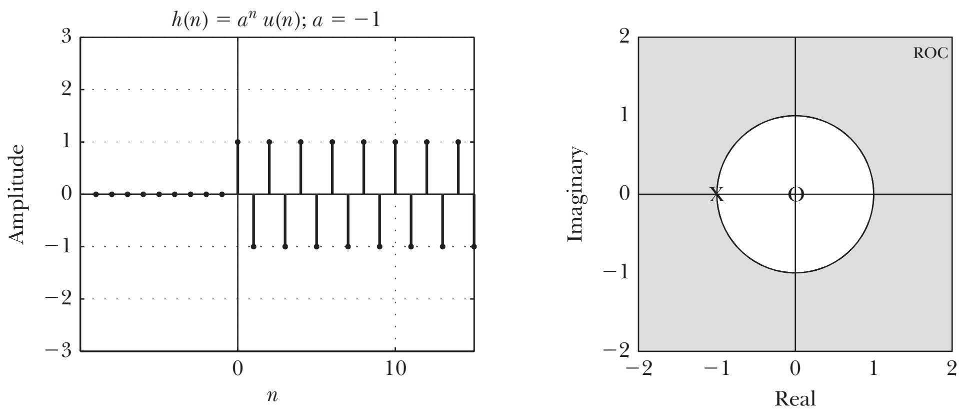



**FIGURE C.14** $a^n u[n] \Leftrightarrow \frac{z}{z-a};\ a = -1.$

# APPENDIX D

# MATLAB Reference Guide

This appendix summarizes MATLAB commands and functions that are especially useful to the study of signals and systems and that have been used throughout the text. A number of these functions are specific to some optional MATLAB toolbox;[1] all the commands used here may be found in the MATLAB *Student Version*. MATLAB is a rich and powerful tool that boasts many more commands and options than can be listed here. For more advanced use of system tools, the MATLAB documentation or help screens should be referenced.

Using only the commands found here, a variety of signals can be defined and plotted, a system can be defined in one of several ways, the system behavior can be confirmed, and a test signal can be input to the system to observe the output. Each of these operations can be readily accomplished in a few lines.

```
% MATLAB comments are text like this after a percent sign.
```

## D.1 Defining Signals

### D.1.1 MATLAB Variables

Variable names in MATLAB are *case sensitive*, such that `Power` is different from `power`.

A few mathematical constants are predefined in MATLAB, including:

```
i    % = square root of -1
j    % = square root of -1
pi   % = 3.14159 ···
```

These predefined values may be overwritten or redefined, and care should taken to not use these variable names in other contexts. The default value of any of the above constants can be restored using the command `clear`. For example:

```
pi = 3;      % sets the variable pi equal to 3
clear pi;    % restores the default value of pi
```

The command `clear all` deletes all user-defined variable names and restores all the default values.

[1] Primarily the *Signal Processing Toolbox* and *Control System Toolbox.*

## D.1.2 The Time Axis

The time axis must be defined as a time interval and then by either by specifying a spacing that determines the total number of points or by specifying a total number of points that determines the spacing. Both of the commands below define the same 1001 points on a time interval $[-5, 5]$ seconds with 0.01 second spacing between points.

```
t = -5:0.01:5;                    % t = [-5,5] spacing = 0.01
t = linspace(-5,5,1001)           % t = [-5,5] points = 1001

size(t)   % return the array dimensions

ans =
     1      1001

length(t) % return the largest dimension

ans =
      1001
```

## D.1.3 Common Signals

Once the time axis is defined, certain commonplace signals are found as built-in commands or as simple constructs. Note that `pi` is predefined in MATLAB.

**Nonperiodic**

```
y = rectpuls(t);       % unit rectangle
y = double(t>=0);      % unit step
y = heaviside(t);      % unit step
y = exp(t);            % exponential
y = sinc(t);           % sinc
```

**Periodic**

```
y = sin(t);          % sine
y = cos(t);          % cosine
y = square(t);       % square wave
y = sawtooth(t);     % sawtooth wave
y = sawtooth(t, 0.5);                 % triangle wave
y = square(t,duty);                   % pulse wave
y = pulstran(t,d,'rectpuls',w);       % pulse wave
y = pulstran(t,d,'tripuls', w);       % triangle wave
```

Many built-in signal definitions are part of the *Signal Processing Toolbox* that is included with the Student Version of MATLAB. In MATLAB versions prior to 2011a, the command `rectpuls()` returns a vector of logical type that may not be compatible with other operations. The command `double()` converts an input variable to the double format.

## D.2 Complex Numbers

Complex variables in MATLAB are expressed as real and imaginary parts in the form $a + bi$ where $i$ represents the square root of $-1$. In electrical engineering, $j$ is used to represent the square root of $-1$, and both $i$ and $j$ are predefined as such in MATLAB. While MATLAB allows a user to set both $i$ and $j$ to other values, this would be undesirable when complex manipulations are anticipated. In any case, MATLAB always returns results using $i$. For example, the same complex result is returned using $i$ in each of the following three expressions.

```
x = i    % default i equals square root of -1
x =
   0 + 1.0000i

x = j    % default j equals square root of -1
x =
   0 + 1.0000i

x = sqrt(-1)   % using the square root of -1
x =
   0 + 1.0000i
```

Complex values may be defined directly so that the value $x = 3 + 4j$ may be represented as:

```
y = complex(3,4)    % define y as real and imaginary parts
y =
          3.0000 + 4.0000i

y = 3 + 4j     % define y as a complex expression
y =
          3.0000 + 4.0000i

real(y)    % the real part of y
ans=
           3

imag(y)    % the imaginary part of y
ans=
           4

abs(y)     % the magnitude of y
ans=
           5

angle(y)   % the phase angle of y (radians)
ans=
           0.9273
```

The phase and magnitude of the complex value $y$ are given by:

$$|y| = \sqrt{\mathrm{Im}(y)^2 + \mathrm{Re}(y)^2} \qquad \Phi(y) = \tan^{-1}\left[\frac{\mathrm{Im}(y)}{\mathrm{Re}(y)}\right]$$

so the same results can be obtained from first principles as:

```
sqrt(imag(y)^2 + real(y)^2)    % magnitude of y, same as abs(y)
ans=
          5

atan2(imag(y), real(y))        % phase angle of y (radians)
ans=
          0.9273
```

# D.3 Plot Commands

The simplest graph of a MATLAB array $s$ is obtained using `plot(s)`, which creates a graph automatically scaled in the $y$-axis to fully fit the values in $s$, and the horizontal axis will be labelled with the full set of indices spanned by $s$. The option `plot(t,s)` uses values from $t$ to fill the $x$-axis; both $s$ and $t$ must have the same number of elements. Some of the most used functions to enhance a plot are listed below.

While `plot(s)` connects points with a line, another useful graph type with similar options is `stem(s)`, which connects a vertical line to each point from the $x$-axis. By default, a small circle is also drawn on each point. The option `stem(s,'marker','none')` omits the circles.

1. **Plot Window** By default, the first plot window will be named *Figure 1* and further calls to the `plot` function will replace the current graph with a new plot.
   - **a.** `figure(N);` Use figure $N$ for future calls to plot.
   - **b.** `clf;` Clear the current figure.
   - **c.** `subplot(A,B,C);` Use plot $C$ from among $A \times B$ subplots.
   - **d.** `hold on;` Further calls to `plot` will *not* delete current contents.
2. **Axis Labels**
   - **a.** `title('text');` Adds a text title to a plot.
   - **b.** `xlabel('text');` Adds a label to the $x$-axis.
   - **c.** `ylabel('text');` Adds a label to the $y$-axis.
3. **Plot Markup**
   - **a.** `text(x,y,'text');` Inserts text within a plot at $(x, y)$.
   - **b.** `line([x1,y1],[x2,y2]);` Draws a line within a plot.
4. **Plot Adjustment**
   - **a.** `axis([x1, x2, y1, y2]);` Set the horizontal and vertical limits.
   - **b.** `set(gca,'XTick',0:15);` Set $x$-axis tick marks 0:15.
   - **c.** `set(gca,'XTickLabel',0:15);` Label tick marks as above.
   - **d.** `grid on;` Turn on plot grid lines.

# D.4 Signal Operations

1. **Numerical integration** `trapz()` (trapezoid approximation) essentially determines "the area under the curve" for a signal over a given interval. To find the

integral of a signal $s(t)$, define an interval for the integration and a step size that accurately reflects the variations in the signal. For example, find the definite integral:

$$y = \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \cos(t)\,dt$$

```
t = -pi/2:0.01:pi/2      % define the time interval
y = trapz(t,cos(t))      % compute the integration
y =
      2.0000
```

Alternatively, `trapz()` may be used directly and the result scaled by the step size:

```
t = -pi/2:0.01:pi/2      % define the time interval
y = trapz(cos(t))*0.01   % compute the integration
y =
        2.0000
```

2. **Convolution** `conv(a,b)` finds the convolution of two signals.

# D.5 Defining Systems

An LTI system can be described in MATLAB in several ways and, once defined, the system behavior can be examined using a number of different MATLAB functions. Figure D.1 gives an overview of this approach and the associated commands. These same functions are used to define and to examine both continuous time and discrete time systems.

## D.5.1 System Definition

Define a system named `sys` in several ways:

1. `sys = tf(···)`; Define from transfer function.
2. `sys = zpk(···)`; Define from zeros and poles and gain.
3. `sys = ss(···)`; Define from state-space description.
4. Discrete time forms of the above.

**1. Transfer Function** A continuous time system defined by its transfer function $H(s)$ can use the `tf` function in one of two ways. For example, consider the Laplace transfer function:

$$H(s) = 5\left[\frac{2s+10}{(s+5)^2+10^2}\right]$$

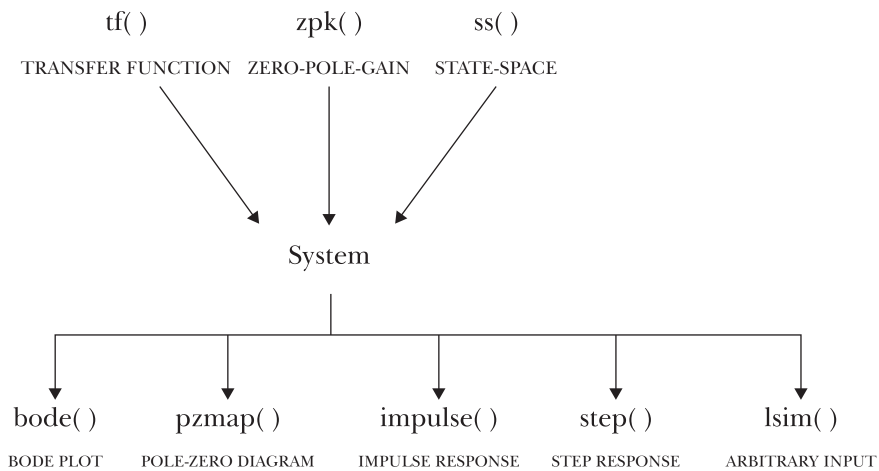



**FIGURE D.1** A *system* defined in MATLAB in one of several different ways can then be studied using a variety of common tools. These same functions are used to define and to examine both continuous time and discrete time systems.

To define this system, the variable $s$ can be assigned as a transfer function variable, after which $H(s)$ can be defined exactly as the equation is written.

```
s = tf('s');
sys = 5 * (2*s+10) / ((s+5)^2 + 10^2)
Transfer function:
   10s + 50
-----------------
s^2 + 10 s + 125
```

It can be seen that MATLAB has identified `sys` as a transfer function and expanded the numerator and denominator into polynomials in terms of $s$. Alternatively, the transfer function can be specified directly in the form [*numerator*, *denominator*] where the three denominator terms in $s^N$ are in the form `[1 10 125]` while the numerator terms are `[10 50]`. In this case, $s$ need not be defined; instead, using `tf(num, den)` gives:

```
sys= tf([10 50], [1 10 125])
Transfer function:
   10s + 50
-----------------
s^2 + 10 s + 125
```

Using either method, the system object named `sys` has been established. The `tf(sys)` function can also be used to convert an existing system `sys` to a transfer function model.

**2. Zeros and Poles and Gain** The same system as above can be defined directly from a knowledge of the poles and zeros. An additional gain term is required to establish the proper scaling for the system, so this is called the *zero-pole-gain* approach. In this case, the gain is 10 so that the zero at $s = -5$ defines the numerator $(10s + 50)$. There is also a pair of complex poles, and the system is written as:

```
zpk([-5],[(-5.0000 +10.0000j), (-5.0000 -10.0000j)], 10)
Zero/pole/gain:
   10 (s+5)
----------------
(s^2 + 10s + 125)
```

Defining a system model in this way allows a direct study of the effect as poles and zeros are moved about the *s*-plane. The `zpk(sys)` function can be used to convert an existing system `sys` to a zero-pole-gain model.

**3. State Space Model** A system can also be defined directly from a state space model using `ss(a,b,c,d)` and where $\{a, b, c, d\}$ are the system matrices. The function `ss(sys)` can also be used to convert a defined system `sys` to a state space model.

**4. Discrete Time Systems** All of the above commands can be used for a discrete time LTI system by including an additional parameter to specify the sample time. For example consider a discrete time system having a sample rate of 1000 samples per second, or a sample time of 1/1000 seconds. The above transfer function becomes:

```
sys= tf([10 50], [1 10 125], 1/1000)
Transfer function:
   10 z + 50
----------------
z^2 + 10 z + 125
```

where the transfer function is now automatically expressed as *z*-transform.

## D.5.2 System Analysis

In examining the system behavior, the appropriate forms of the Bode plots and pole-zero diagram are automatically chosen depending on whether `sys` was defined as a continuous time or as a discrete time system. Available functions include:

1. `bode(sys);` Plot the transfer function magnitude and phase.
2. `bodemag(sys);` Plot the transfer function magnitude only.
3. `pzmap(sys);` Plot the pole-zero diagram.
   a. `[p,z] = pzmap(sys);` Also return any poles ($p$) and zeros ($z$).
   b. `pole(sys);` Returns the value of any poles.
   c. `zero(sys);` Returns the value of any zeros.

4. `impulse(sys);` Plot the impulse response.
5. `step(sys);` Plot the step response.
6. `lsim(sys,input,t);` Plot the arbitrary input response.

## D.6 Example System Definition and Test

In this example, a simple system will be defined and examined, and then tested using a sinusoidal input signal. First, define a system using its transfer function description.

```
sys = tf([1],[1 1]);     % define a system (lowpass filter)
bode(sys);               % examine the system Bode plot
grid on;                 % include grid lines
```

The expected behavior of this system is seen in the Bode plot of Figure D.2. This Bode plot shows that the system is a first-order lowpass filter with cutoff frequency $f_c = 1$ rad/s (half-power point) where the phase shift is $-45$ degrees.

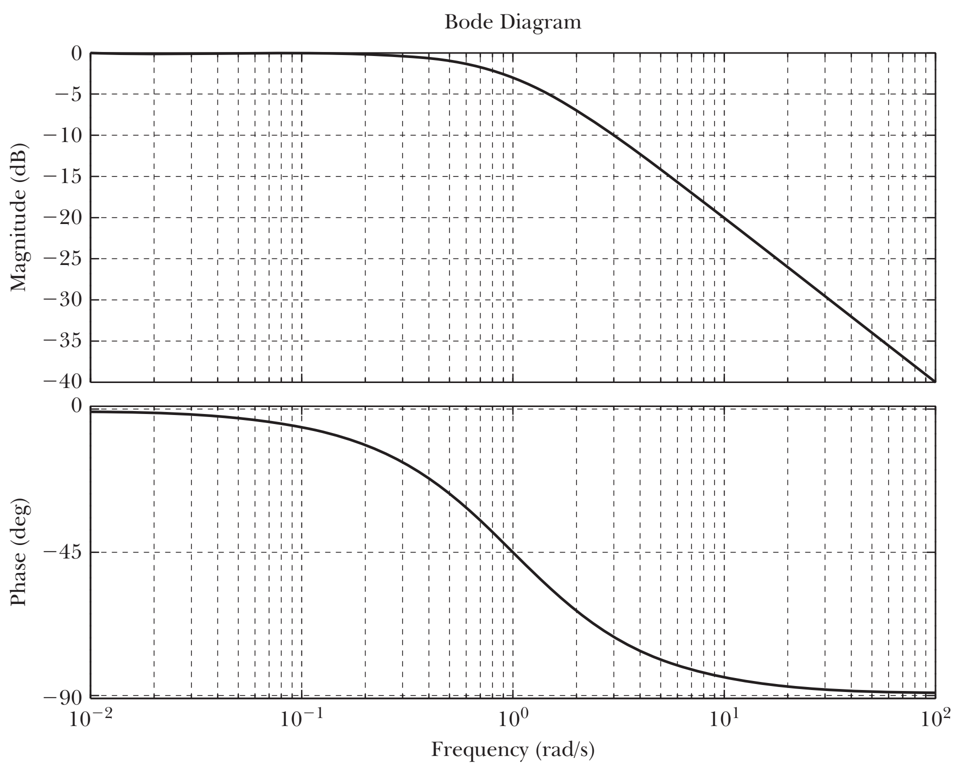



**FIGURE D.2** A *system* is defined in MATLAB by a transfer function and then the system is examined using a Bode plot.

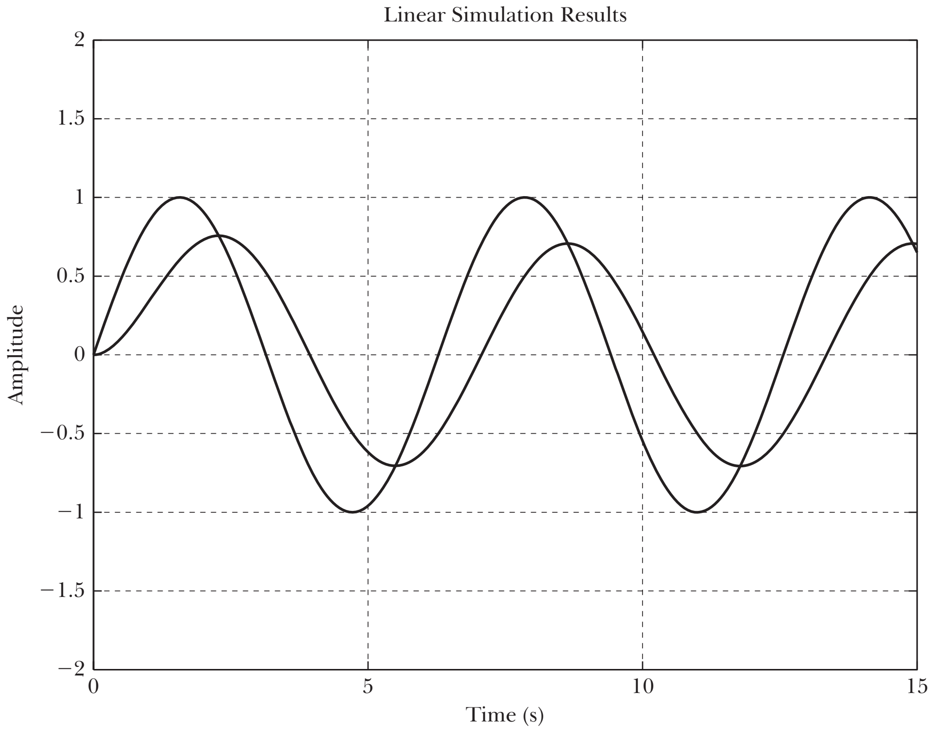



**FIGURE D.3** A lowpass filter *system* is defined in MATLAB by a transfer function, then a sinusoidal test signal is applied at the cutoff frequency. This graph shows the input and the attenuated, phase shifted, output signal.

To observe the response of this system `sys` to some input signal, first define a time axis, then a test signal, then use `lsim()` to view the resulting system output.

```
time = 0:0.001:15;          % define a time axis
input = sin(time);          % define a test input
lsim(sys,input,time);       % plot the test input response
axis([0 15 -2 2]);          % adjust the graph axes
grid on;                    % include grid lines
```

The response of this system to a sinusoidal test input at the cutoff frequency is seen in Figure D.3. The graph shows the sinusoidal input and simulated output signals in which the output is found to be attenuated to about 0.707 (half power) and shifted in phase by 45 degrees. This is the same response that was predicted in the Bode plot.

Alternatively, when using `lsim()` with a sinusoidal test signal, the function `[input,time]=gensig('sin',period);` with a constant `period` returns both the test input signal and a suitable time variable to span about five periods of the waveform. This is a very convenient way to rapidly create different test input signals.

APPENDIX E

# Reference Tables

## E.1 Fourier Transform

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft}\, dt$$

$$s(t) = \int_{-\infty}^{+\infty} S(f) e^{+j2\pi ft}\, df$$

### E.1.1 Fourier Transform Theorems

linearity:

$$k_1 s_1(t) + k_2 s_2(t) \xleftrightarrow{\mathcal{F}} k_1 S_1(f) + k_2 S_2(f)$$

amplifying:

$$ks(t) \xleftrightarrow{\mathcal{F}} kS(f)$$

scaling:

$$s(kt) \xleftrightarrow{\mathcal{F}} \frac{1}{|k|} S\left(\frac{f}{k}\right)$$

time shift:

$$s(t-k) \xleftrightarrow{\mathcal{F}} e^{-j2\pi fk} S(f)$$

convolution:

$$a(t) * b(t) \xleftrightarrow{\mathcal{F}} A(f) \times B(f)$$

multiplication:

$$a(t) \times b(t) \xleftrightarrow{\mathcal{F}} A(f) * B(f)$$

modulation:

$$s(t) \cos(2\pi f_0 t) \xleftrightarrow{\mathcal{F}} \frac{1}{2} S(f+f_0) + \frac{1}{2} S(f-f_0)$$

derivative:

$$\frac{d}{dt}s(t) = s'(t) - \overset{\mathcal{F}}{\longleftrightarrow} j2\pi f\ S(f)$$

integral:

$$\int_{-\infty}^{t} s(x)dx \overset{\mathcal{F}}{\longleftrightarrow} \frac{1}{j2\pi}f\ S(f) + \frac{1}{2}S(0)\delta(f)$$

sampling:

$$\sum_{n=-\infty}^{+\infty} \delta(t - nT) \overset{\mathcal{F}}{\longleftrightarrow} \frac{1}{T}\sum_{n=-\infty}^{+\infty} \delta(f - \frac{n}{T})$$

Parseval's:

$$\int_{-\infty}^{+\infty} |s(t)|^2\ dt = \int_{-\infty}^{+\infty} |S(f)|^2 df$$

# E.2 Laplace Transform

two-sided:

$$A(s) = \int_{-\infty}^{+\infty} a(t)e^{-st}dt$$

one-sided:

$$A(s) = \int_{0}^{+\infty} a(t)e^{-st}dt$$

## E.2.1 Laplace Transform Theorems

linearity:

$$k_1 a_1(t) + k_2 a_2(t) \overset{\mathcal{L}}{\longleftrightarrow} k_1\ A_1(s) + k_2\ A_2(s)$$

amplifying:

$$ka(t) \overset{\mathcal{L}}{\longleftrightarrow} kA(s)$$

scaling:

$$a(kt) \overset{\mathcal{L}}{\longleftrightarrow} \frac{1}{|k|}A(\frac{s}{k})$$

shifting:

$$a(t - k) \overset{\mathcal{L}}{\longleftrightarrow} e^{-sk}A(s)$$

convolution:

$$a(t) * b(t) \xleftrightarrow{\mathcal{L}} A(s) \times B(s)$$

derivative:

$$\frac{d}{dt}a(t) = a'(t) \xleftrightarrow{\mathcal{L}} sA(s) - a(0)$$

second derivative:

$$\frac{d^2}{dt}a(t) = a''(t) \xleftrightarrow{\mathcal{L}} s^2A(s) - s\ a(0) - a'(0)$$

integration:

$$\int_0^t a(\tau)\,d\tau \xleftrightarrow{\mathcal{L}} \frac{1}{s}A(s)$$

## E.3 *z*-Transform

two-sided:

$$X(z) = \sum_{n=-\infty}^{+\infty} x[n]z^{-n}$$

one-sided:

$$X(z) = \sum_{n=0}^{+\infty} x[n]z^{-n}$$

### E.3.1 *z*-Transform Theorems

linearity:

$$k_1 s_1[n] + k_2\ s_2[n] \xleftrightarrow{\mathcal{Z}} k_1 S_1(z) + k_2 S_2(z)$$

amplifying:

$$ks[n] \xleftrightarrow{\mathcal{Z}} kS(z)$$

scaling:

$$k^n s[n] \xleftrightarrow{\mathcal{Z}} S\left(\frac{z}{k}\right)$$

shifting:

$$s[n-k] \xleftrightarrow{\mathcal{Z}} z^{-k}S(z)$$

convolution:

$$a[n] * b[n] \xleftrightarrow{\mathcal{Z}} A(z) \times B(z)$$

derivative:

$$-z\frac{d}{dz}S(z) \longleftrightarrow n\ s[n]$$

# BIBLIOGRAPHY


[1] Kitcher, P. "Mathematical Rigor—Who Needs It?" *Noûs* 15, no. 4 (1981): 469–493.

[2] Institute of Electrical and Electronics Engineers and IEEE Standards Coordinating Committee on Definitions. *IEEE Standard Dictionary of Electrical and Electronics Terms.* 3rd ed. Edited by F. Jay and J. A. Goetz. New York, NY: Institute of Electrical and Electronics Engineers, 1984.

[3] Davey, K. "Is Mathematical Rigor Necessary in Physics?" *British Journal for the Philosophy of Science* 54, no. 3 (2003): 439–463.

[4] *Merriam-Webster's Collegiate Dictionary.* Springfield, MA: Merriam-Webster, Inc., 2003.

[5] James, R. C., and James, G. *Mathematics Dictionary.* 3rd ed. Princeton, NJ: Van Nostrand, 1968.

[6] Bracewell, R. *The Fourier Transform and Its Applications.* 3rd ed. McGraw-Hill Electrical and Electronic Engineering Series. New York, NY: McGraw-Hill, 1978.

[7] Abramowitz, M., and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* 4th ed. Applied Mathematics Series 55. New York, NY: National Bureau of Standards, 1964.

# INDEX

# 0.1 Useful Information

## 0.1.1 Identities

$$2\cos(x)\cos(y) = \cos(x-y) + \cos(x+y)$$

$$2\sin(x)\sin(y) = \cos(x-y) - \cos(x+y)$$

$$2\sin(x)\cos(y) = \sin(x-y) + \sin(x+y)$$

$$\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$$

$$\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$$

$$\sin(x) + \sin(y) = 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)$$

$$\cos(x) + \cos(y) = 2\cos\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right)$$

$$\operatorname{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

## 0.1.2 Definite Integrals

$$\int x\sin(ax)dx = \frac{\sin(ax)}{a^2} - \frac{x\cos(ax)}{a}$$

$$\int x\cos(ax)dx = \frac{\cos(ax)}{a^2} + \frac{x\sin(ax)}{a}$$

$$\int \frac{1}{a^2+x^2}\,dx = \frac{1}{a}\tan^{-1}\left(\frac{x}{a}\right)$$

### 0.1.3 Infinite Series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad \text{for } |x| < 1$$

### 0.1.4 Orthogonality

$$\int_{t_0}^{t_1} a(t)\ b^*(t)\ dt = 0$$

### 0.1.5 Signal Inner Product

$$a(t) \cdot b(t) = \int_{t_0}^{t_1} a(t)\ b^*(t)\ dt$$

### 0.1.6 Convolution

continuous:

$$a(t) * b(t) = \int_{-\infty}^{+\infty} a(t-x)\ b(x) dx$$

discrete:

$$a[n] * b[n] = \sum_{k=-\infty}^{+\infty} a[n-k]\ b[n]$$

### 0.1.7 Fourier Series

$$s(t) = A_0 + \sum_{n=1}^{+\infty} [A_n \cos(2\pi n f_0 t) + B_n \sin(2\pi n f_0 t)]$$

$$A_0 = \frac{1}{T} \int_{-T/2}^{+T/2} s(t) dt$$

$$A_n = \frac{2}{T} \int_{-T/2}^{+T/2} s(t) \cos(2\pi n f_0 t) dt$$

$$B_n = \frac{2}{T} \int_{-T/2}^{+T/2} s(t) \sin(2\pi n f_0 t) dt$$

### 0.1.8 Complex Fourier Series

$$s(t) = \sum_{n=-\infty}^{+\infty} C_n \; e^{+j2\pi n f_0 t}$$

$$C_n = \frac{1}{T} \int_{-T/2}^{+T/2} s(t) \; e^{-j2\pi n f_0 t} \, dt$$

### 0.1.9 Fourier Transform

$$s(t) \overset{\mathcal{F}}{\longleftrightarrow} S(f)$$

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi f t} \, dt$$

$$s(t) = \int_{-\infty}^{+\infty} S(f) e^{+j2\pi f t} \, df$$

### 0.1.10 Laplace Transform

$$x(t) \overset{\mathcal{L}}{\longleftrightarrow} X(s)$$

two-sided:

$$X(s) = \int_{-\infty}^{+\infty} x(t) \; e^{-st} \, dt$$

one-sided:

$$X(s) = \int_{0}^{+\infty} x(t) \; e^{-st} \, dt$$

### 0.1.11 $z$–Transform

$$x[n] \overset{\mathcal{Z}}{\longleftrightarrow} X(z)$$

two-sided:

$$X(z) = \sum_{n=-\infty}^{+\infty} x[n] z^{-n}$$

one-sided:

$$X(z) = \sum_{n=0}^{+\infty} x[n] z^{-n}$$

## 0.2 List of Acronyms

| | |
|---|---|
| BIBO | Bounded Input, Bounded Output |
| BPF | Band Pass Filter |
| DC | Direct Current |
| DFT | Discrete Fourier Transform |
| DSP | Digital Signal Processing |
| DTFT | Discrete Time Fourier Transform |
| FFT | Fast Fourier Transform |
| FIR | Finite Impulse Response |
| HPF | Highpass Filter |
| IIR | Infinite Impulse Response |
| LPF | Lowpass Filter |
| LTI | Linear Time Invariant |
| RC | Resistor, Capacitor (circuit) |
| RLC | Resistor, Inductor, Capacitor (circuit) |
| ROC | Region of Convergence |