



**Modern Digital  
and Analog  
Communication  
Systems** THIRD EDITION

**B. P. Lathi**

THE OXFORD SERIES IN ELECTRICAL AND COMPUTER ENGINEERING

SERIES EDITORS

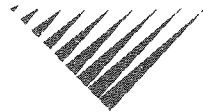
Adel S. Sedra, *Electrical Engineering*  
Michael R. Lightner, *Computer Engineering*

SERIES TITLES

Allen and Holberg, *CMOS Analog Circuit Design*  
Bobrow, *Elementary Linear Circuit Analysis, 2nd Ed.*  
Bobrow, *Fundamentals of Electrical Engineering, 2nd Ed.*  
Campbell, *The Science and Engineering of Microelectronic Fabrication*  
Chen, *Linear System Theory and Design, 3rd Ed.*  
Chen, *System and Signal Analysis, 2nd Ed.*  
Comer, *Digital Logic and State Machine Design, 3rd Ed.*  
Cooper and McGillem, *Probabilistic Methods of Signal and System Analysis, 3rd Ed.*  
Franco, *Electric Circuits Fundamentals*  
Jones, *Introduction to Optical Fiber Communication Systems*  
Krein, *Elements of Power Electronics*  
Kuo, *Digital Control Systems, 3rd Ed.*  
Lathi, *Modern Digital and Analog Communications Systems, 3rd Ed.*  
McGillem and Cooper, *Continuous and Discrete Signal and System Analysis, 3rd Ed.*  
Miner, *Lines and Electromagnetic Fields for Engineers*  
Roberts, *SPICE, 2nd Ed.*  
Santina, Stubberud and Hostetter, *Digital Control System Design, 2nd Ed.*  
Schwarz, *Electromagnetics for Engineers*  
Schwarz and Oldham, *Electrical Engineering: An Introduction, 2nd Ed.*  
Sedra and Smith, *Microelectronic Circuits, 4th Ed.*  
Stefani, Savant, and Hostetter, *Design of Feedback Control Systems, 3rd Ed.*  
Van Valkenburg, *Analog Filter Design*  
Warner and Grung, *Semiconductor Device Electronics*  
Wolovich, *Automatic Control Systems*  
Yariv, *Optical Electronics in Modern Communications, 5th Ed.*

# MODERN DIGITAL AND ANALOG COMMUNICATION SYSTEMS

Third Edition



B. P. LATHI

New York Oxford  
OXFORD UNIVERSITY PRESS  
1998

Oxford University Press

Oxford New York

Athens Auckland Bangkok Bogota Bombay Buenos Aires  
Calcutta Cape Town Dar es Salaam Delhi Florence Hong Kong  
Istanbul Karachi Kuala Lumpur Madras Madrid Melbourne  
Mexico City Nairobi Paris Singapore Taipei Tokyo Toronto Warsaw  
*and associated companies in*  
Berlin Ibadan

Copyright © 1998 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.,  
198 Madison Avenue, New York, New York 10016  
<http://www.oup-usa.org>  
1-800-334-4249

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Oxford University Press.

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

Lathi, B. P. (Bhagwandas Pannalal)

Modern digital and analog communication systems / B.P. Lathi.—  
3rd ed.

p. cm.—(The Oxford series in electrical and computer  
engineering)

Includes bibliographical references (p. )

ISBN 0-19-511009-9 (cloth)

1. Telecommunication systems. 2. Digital communications.  
3. Statistical communication theory. I. Title II. Series.

TK5101.L333 1998

621.382—dc21

97-16040

CIP

Printing (last digit): 9 8 7 6 5 4 3 2 1

Printed in the United States of America  
on acid-free paper

In Memory of  
M.E. Van Valkenburg  
1921–1997



# CONTENTS

PREFACE *xi*

x1 INTRODUCTION 1

Communication System 1

Analog and Digital Messages 3

Signal-to-Noise Ratio, Channel Bandwidth, and the Rate of Communication 8

Modulation 10

Randomness, Redundancy, and Coding 12

2 INTRODUCTION TO SIGNALS 14

2.1 Size of a Signal 14

2.2 Classification of Signals 20

2.3 Some Useful Signal Operations 24

2.4 Unit Impulse Function 28

2.5 Signals and Vectors 30

2.6 Signal Comparison: Correlation 35

2.7 Signal Representation by Orthogonal Signal Set 40

2.8 Trigonometric Fourier Series 44

2.9 Exponential Fourier Series 53

2.10 Numerical Computation of  $D_n$  60

3 ANALYSIS AND TRANSMISSION OF SIGNALS 71

3.1 Aperiodic Signal Representation by Fourier Integral 71

3.2 Transforms of Some Useful Functions 78

3.3 Some Properties of the Fourier Transform 84

3.4 Signal Transmission through a Linear System 101

3.5 Ideal and Practical Filters 106

3.6 Signal Distortion over a Communication Channel 110

	3.7 Signal Energy and Energy Spectral Density	115
	3.8 Signal Power and Power Spectral Density	123
	3.9 Numerical Computation of Fourier Transform: The DFT	130
<del>X</del> 4	AMPLITUDE (LINEAR) MODULATION	151
	4.1 Baseband and Carrier Communication	151
	4.2 Amplitude Modulation: Double Sideband (DSB)	152
	4.3 Amplitude Modulation (AM)	162
	4.4 Quadrature Amplitude Modulation (QAM)	170
	4.5 Amplitude Modulation: Single Sideband (SSB)	171
	4.6 Amplitude Modulation: Vestigial Sideband (VSB)	179
	4.7 Carrier Acquisition	183
	4.8 Superheterodyne AM Receiver	189
	4.9 Television	191
5	ANGLE (EXPONENTIAL) MODULATION	208
	5.1 Concept of Instantaneous Frequency	208
	5.2 Bandwidth of Angle-Modulated Waves	215
	5.3 Generation of FM Waves	229
	5.4 Demodulation of FM	233
	5.5 Interference in Angle-Modulated Systems	241
	5.6 FM Receiver	245
<del>X</del> 6	SAMPLING AND PULSE CODE MODULATION	251
	6.1 Sampling Theorem	251
	6.2 Pulse-Code Modulation (PCM)	262
	6.3 Differential Pulse Code Modulation (DPCM)	278
	6.4 Delta Modulation	281
<del>X</del> 7	PRINCIPLES OF DIGITAL DATA TRANSMISSION	294
	7.1 A Digital Communication System	294
	7.2 Line Coding	297
	7.3 Pulse Shaping	310
	7.4 Scrambling	319
	7.5 Regenerative Repeater	322
	7.6 Detection-Error Probability	329
	7.7 <i>M</i> -ary Communication	334
	7.8 Digital Carrier Systems	337
	7.9 Digital Multiplexing	342
8	EMERGING DIGITAL COMMUNICATIONS TECHNOLOGIES	354
	8.1 The North American Hierarchy	354
	8.2 Digital Services	368
	8.3 Broadband Digital Communication: SONET	377



8.4	Digital Switching Technologies	383
8.5	Broadband Services for Entertainment and Home Office Applications	392
8.6	Video Compression	395
8.7	High-Definition Television (HDTV)	400
<b>9</b>	<b>SOME RECENT DEVELOPMENTS AND MISCELLANEOUS TOPICS</b>	<b>404</b>
9.1	Cellular Telephone (Mobile Radio) System	404
9.2	Spread Spectrum Systems	406
9.3	Transmission Media	416
9.4	Hybrid Circuit: 2-Wire to 4-Wire Conversions	427
9.5	Public Switched Telephone Network	430
<b>10</b>	<b>INTRODUCTION TO THEORY OF PROBABILITY</b>	<b>434</b>
10.1	Concept of Probability	434
10.2	Random Variables	445
10.3	Statistical Averages (Means)	463
10.4	Central-Limit Theorem	472
10.5	Correlation	473
10.6	Linear Mean Square Estimation	476
<b>11</b>	<b>RANDOM PROCESSES</b>	<b>487</b>
11.1	From Random Variable to Random Process	487
11.2	Power Spectral Density of a Random Process	496
11.3	Multiple Random Processes	509
11.4	Transmission of Random Processes through Linear Systems	510
11.5	Bandpass Random Processes	514
11.6	Optimum Filtering: Wiener-Hopf Filter	522
<b>12</b>	<b>BEHAVIOR OF ANALOG SYSTEMS IN THE PRESENCE OF NOISE</b>	<b>532</b>
12.1	Baseband Systems	532
12.2	Amplitude-Modulated Systems	534
12.3	Angle-Modulated Systems	541
12.4	Pulse-Modulated Systems	557
12.5	Optimum Preemphasis-Deemphasis Systems	567
<b>13</b>	<b>BEHAVIOR OF DIGITAL COMMUNICATION SYSTEMS IN THE PRESENCE OF NOISE</b>	<b>577</b>
13.1	Optimum Threshold Detection	577
13.2	General Analysis: Optimum Binary Receiver	582
13.3	Carrier Systems: ASK, FSK, PSK, and DPSK	590
13.4	Performance of Spread Spectrum Systems	601

13.5 <i>M</i> -ary Communication	608
13.6 Synchronization	622

14 OPTIMUM SIGNAL DETECTION 626

14.1 Geometrical Representation of Signals: Signal Space	626
14.2 Gaussian Random Process	632
14.3 Optimum Receiver	637
14.4 Equivalent Signal Sets	662
14.5 Nonwhite (Colored) Channel Noise	669
14.6 Other Useful Performance Criteria	670

15 INTRODUCTION TO INFORMATION THEORY 679

15.1 Measure of Information	679
15.2 Source Encoding	684
15.3 Error-Free Communication over a Noisy Channel	690
15.4 Channel Capacity of a Discrete Memoryless Channel	693
15.5 Channel Capacity of a Continuous Channel	701
15.6 Practical Communication Systems in Light of Shannon's Equation	717

16 ERROR CORRECTING CODES 728

16.1 Introduction	728
16.2 Linear Block Codes	731
16.3 Cyclic Codes	737
16.4 Burst-Error Detecting and Correcting Codes	745
16.5 Interlaced Codes for Burst- and Random-Error Correction	746
16.6 Convolutional Codes	747
16.7 Comparison of Coded and Uncoded Systems	755

APPENDIXES 764

A. Orthogonality of Some Signal Sets	764
B. Schwarz Inequality	766
C. Gram-Schmidt Orthogonalization of a Vector Set	768
D. Miscellaneous	771

INDEX 775

# PREFACE

The study of communication systems can be divided into two distinct areas:

1. How communication systems work.
2. How they perform in the presence of noise.

The study of each of these two areas, in turn, requires specific tools. To study the first area, the students must be familiar with signal analysis (Fourier techniques), and to study the second area, a basic understanding of probability theory and random processes is essential. For a meaningful comparison of various communication systems, it is necessary to have some understanding of the second area. For this reason many instructors feel that the study of communication systems is not complete unless both of the areas are covered reasonably well. However, it poses one serious problem: the material to be covered is enormous. The two areas along with their tools are overwhelming; it is difficult to cover this material in depth in one course.

The current trend in teaching communication systems is to study the tools in early chapters and then proceed with the study of the two areas of communication. Because too much time is spent in the beginning in studying the tools (without much motivation), there is little time left to study the two proper areas of communication. Consequently, teaching a course in communication systems poses a real dilemma. The second area (statistical aspects) of communication theory is a degree harder than the first area, and it can be properly understood only if the first area is well assimilated. One of the reasons for the dilemma mentioned earlier is our attempt to cover both areas at the same time. The students are forced to grapple with the statistical aspects while also trying to become familiar with how communication systems work. This practice is most unsound pedagogically because it violates the basic fact that one must learn to walk before one can run. The ideal solution would be to offer two courses in sequence, the first course dealing with how communication systems function and the second course dealing with statistical aspects and noise. But in the present curriculum, with so many competing courses, it is difficult to squeeze in two basic courses in the communications area. Some schools do require a course in probability and random processes as a prerequisite. In this case, it is possible to cover both areas reasonably well in a one-semester course. This book,

I hope, can be adopted to either case. It can be used as a one-semester survey course in which the deterministic aspects of communication systems are emphasized. It can also be used for a course that deals with deterministic and probabilistic aspects of communication systems. The book provides all the necessary background in probabilities and random processes. However, as stated earlier, it is highly desirable for students to have a good background in probabilities if the course is to be covered in one semester.

The first nine chapters discuss in depth how digital and analog communication systems work, and thus, form a sound, well-rounded, comprehensive survey course in communication systems that is within the reach of an average undergraduate and that can be taught in a one-semester course (about 40 to 45 hours). However, if the students have an adequate background in Fourier analysis and probabilities, it should be possible to cover the first 13 chapters.

Chapter 1 introduces the students to a panoramic view of communication systems. All the important concepts of communication theory are explained qualitatively in a heuristic way. This gets the students deeply interested so that they are encouraged to study the subject. Because of this momentum, they are motivated to study the tool of signal analysis in Chapters 2 and 3, where a student is encouraged to see a signal as a vector, and to think of the Fourier spectrum as a way of representing a signal in terms of its vector components. Chapters 4 and 5 discuss amplitude (linear) and angle (nonlinear) modulation, respectively. Many instructors feel that in this digital age, modulation should be deemphasized with a minimal presence. I feel that modulation is not so much a method of communication as a basic tool of signal processing; it will always be needed not only in the area of communication (digital or analog), but also in many other areas of electrical engineering. Hence, neglecting modulation may prove to be rather shortsighted. Chapter 6 deals with sampling, pulse code modulation (including DPCM), and delta modulation. Chapter 7 discusses transmission of digital data. Some emerging digital technologies in digital data transmission are the subject of Chapter 8. Chapter 9 discusses some recent developments (such as cellular telephone, spread spectrum, global positioning systems), along with miscellaneous topics such as communication media, optical communication, satellite communication, and hybrid circuits. Chapters 10 and 11 provide a reasonably thorough treatment of the theory of probability and random processes. This is the second tool required for the study of communication systems. Every attempt is made to motivate students and sustain their interest through these chapters by providing applications to communications problems wherever possible. Chapters 12 and 13 discuss the behavior of communication systems in the presence of noise. Optimum signal detection is presented in Chapter 14, and information theory is the subject of Chapter 15. Finally, error-control coding is introduced in Chapter 16.

Analog pulse modulation systems such as PAM, PPM, and PWM are deemphasized in comparison to digital schemes (PCM, DPCM, and DM) because the applications of the former in communications are hard to find. In the treatment of angle modulation, rather than compartmentalizing FM and PM, we have provided a generalized treatment of angle modulation, where FM and PM are merely two (of the infinite) special cases. Tone-modulated FM is deemphasized for a sound reason. Since angle modulation is nonlinear, the conclusions derived from tone modulation cannot be blindly applied to modulation by other baseband signals. In fact, these conclusions are misleading in many instances. For example, in the literature PM gets short shrift as being inferior to FM, a conclusion based on tone-modulation

analysis.\* It is shown in Chapter 12 that PM is, in fact, superior to FM for all practical cases (including audio).

One of the aims in writing this book has been to make learning a pleasant or at least a less intimidating experience for the student by presenting the subject in a clear, understandable, and logically organized manner. Every effort has been made to give an insight—rather than just an understanding—as well as heuristic explanations of theoretical results wherever possible. Many examples are provided for further clarification of abstract results. Even a partial success in achieving my stated goal would make all my toils worthwhile.

## ACKNOWLEDGMENTS

It is a pleasure to acknowledge the assistance received from several individuals during the preparation of this book. I am greatly indebted to Mr. Maynard Wright, who is a member of several standards committees, for his valuable help in several areas of data transmission. He also contributed Secs. 9.4, 9.5, and part of Sec. 9.3. I greatly appreciate the help of Professor William Jameson from Montana State University, who contributed Chapter 8 (Emerging Digital Communication Technologies). I am much obliged to Prof. Brian Woerner and R.M. Buehrer from Virginia Polytechnic Institute for their contribution. The analysis of spread spectrum systems in Section 13.4 and some parts of Sec. 9.2 are based solely on their contribution. I appreciate the enthusiastic help of Jerry Olup in preparation of the solutions manual. Thanks are also due to several reviewers, especially Profs. W. Green, James Kang, Dan Murphy, W. Jameson, Jeff Reed, R. Vaz, S. Bibyk, C. Alexander and S. Mousavinezhad. I am obliged to Berkeley-Cambridge Press for their permission to use the material (Chapters 2 and 3) from their forthcoming publication *Signal Processing and Linear Systems* by B. P. Lathi. Finally, I owe a debt of gratitude to my wife Rajani for her patience and understanding.

B. P. LATHI

---

\* Another reason given for the alleged inferiority of PM is that the phase deviation has to be restricted to a value less than  $\pi$ . It has been shown in Chapter 5 that this is simply not true of band-limited analog signals.

## IEEE CODE OF ETHICS

We, the members of the IEEE, in recognition of the importance of our technologies in affecting the quality of life throughout the world, and in accepting a personal obligation to our profession, its members and the communities we serve, do hereby commit ourselves to conduct of the highest ethical and professional manner and agree:

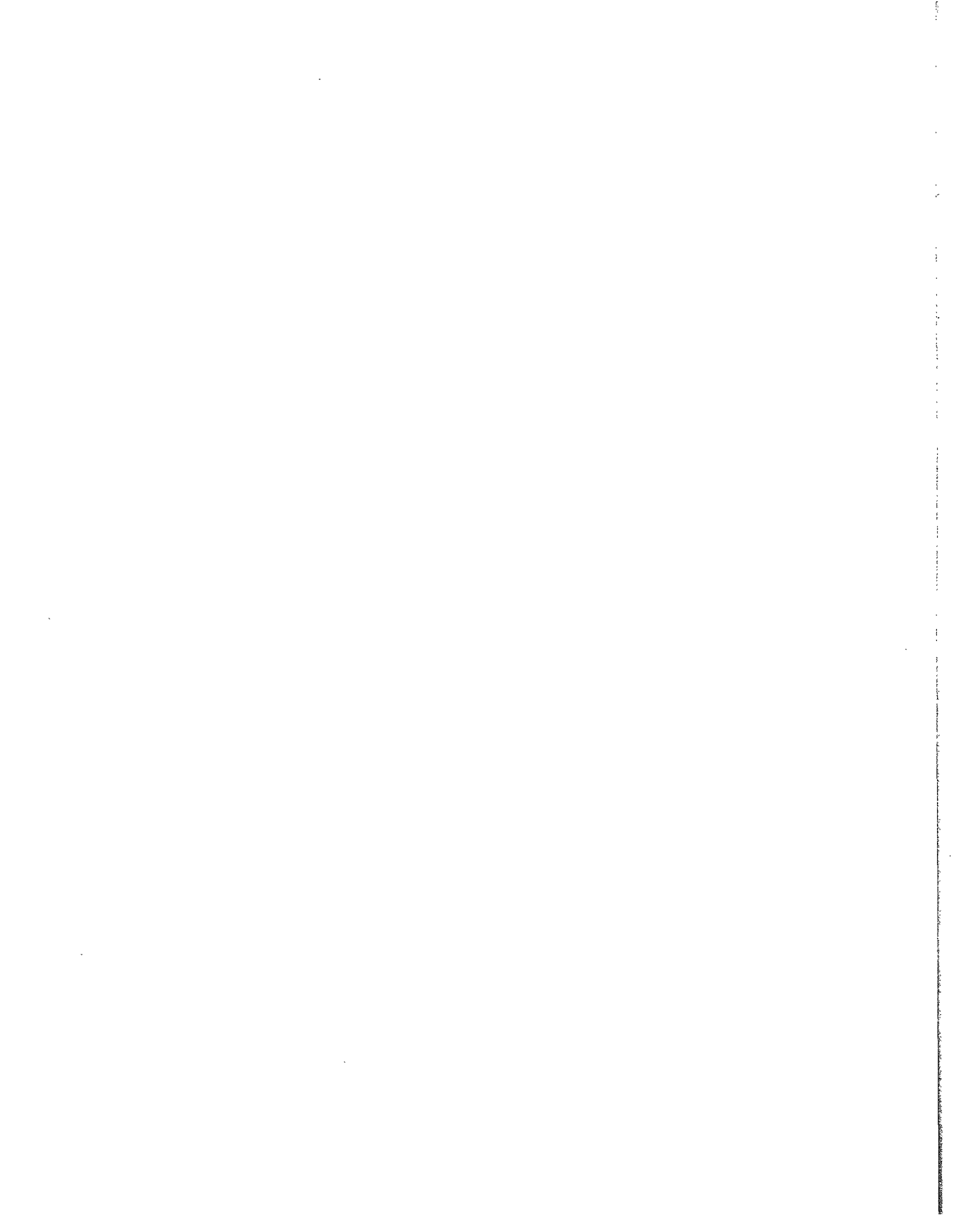
1. to accept responsibility in making engineering decisions consistent with the safety, health, and welfare of the public, and to disclose promptly factors that might endanger the public or the environment;
2. to avoid real or perceived conflicts of interest whenever possible, and to disclose them to affected parties when they do exist;
3. to be honest and realistic in stating claims or estimates based on available data;
4. to reject bribery in all of its forms;
5. to improve understanding of technology; its appropriate application, and potential consequences;
6. to maintain and improve our technical competence and to undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations;
7. to seek, accept, and offer honest criticism of technical work, to acknowledge and correct errors, and to credit properly the contributions of others;
8. to treat fairly all persons regardless of such factors as race, religion, gender, disability, age, or national origin;
9. to avoid injuring others, their property, reputation, or employment by false or malicious action;
10. to assist colleagues and co-workers in their professional development and to support them in following this code of ethics.

Approved by IEEE Board of Directors, August 1990

For further information please consult the IEEE Ethics Committee WWW page:  
<http://www.ieee.org/committee.ethics>

Copyright © 1997 by IEEE

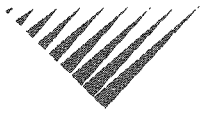
MODERN DIGITAL AND  
ANALOG  
COMMUNICATION  
SYSTEMS





# 1

# INTRODUCTION



**T**his book examines communication by electrical signals. In the past, messages have been carried by runners, carrier pigeons, drum beats, and torches. These schemes were adequate for the distances and “data rates” of the age. In most parts of the world, these modes of communication have been superseded by electrical communication systems,\* which can transmit signals over much longer distances (even to distant planets and galaxies) and at the speed of light.

Electrical communication is reliable and economical; communication technology is alleviating the energy crisis by trading information processing for a more rational use of energy resources. Some examples: Important discussions now mostly communicated face to face in meetings or conferences, often requiring travel, are increasingly using “teleconferencing.” Similarly, teleshopping and telebanking will provide services by electronic communication, and newspapers may be replaced by electronic news services.

## COMMUNICATION SYSTEM

Figure 1.1 shows three examples of communication systems. A typical communication system can be modeled as shown in Fig. 1.2. The components of a communication system are as follows:

The **source** originates a message, such as a human voice, a television picture, a teletype message, or data. If the data is nonelectrical (human voice, teletype message, television picture), it must be converted by an **input transducer** into an electrical waveform referred to as the **baseband signal** or **message signal**.

The **transmitter** modifies the baseband signal for efficient transmission.†

The **channel** is a medium—such as wire, coaxial cable, a waveguide, an optical fiber, or a radio link—through which the transmitter output is sent.

---

\* With the exception of the postal service.

† The transmitter consists of one or more of the following subsystems: a preemphasizer, a sampler, a quantizer, a coder, and a modulator. Similarly, the receiver may consist of a demodulator, a decoder, a filter, and a deemphasizer.



Figure 1.1 Some examples of communications systems.

The **receiver** reprocesses the signal received from the channel by undoing the signal modifications made at the transmitter and the channel. The receiver output is fed to the **output transducer**, which converts the electrical signal to its original form—the message.

The **destination** is the unit to which the message is communicated.

A channel acts partly as a filter to attenuate the signal and distort its waveform. The signal attenuation increases with the length of the channel, varying from a few percent for short distances to orders of magnitude for interplanetary communication. The waveform is distorted because of different amounts of attenuation and phase shift suffered by different frequency components of the signal. For example, a square pulse is rounded or “spread out” during the transmission. This type of distortion, called **linear distortion**, can be partly corrected at the receiver by an equalizer with gain and phase characteristics complementary to those of the channel. The channel may also cause **nonlinear distortion** through attenuation that varies with the signal amplitude. Such distortion can also be partly corrected by a complementary equalizer at the receiver.

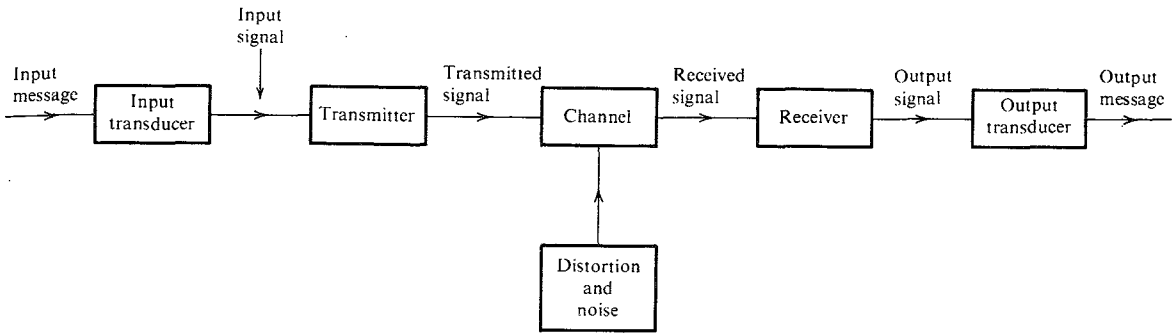


Figure 1.2 Communication system.

The signal is not only distorted by the channel, but it is also contaminated along the path by undesirable signals lumped under the broad term **noise**, which are random and unpredictable signals from causes external and internal. External noise includes interference from signals transmitted on nearby channels, human-made noise generated by faulty contact switches for electrical equipment, automobile ignition radiation, fluorescent lights or natural noise from lightning, as well as electrical storms and solar and intergalactic radiation. With proper care, external noise can be minimized or even eliminated. Internal noise results from thermal motion of electrons in conductors, random emission, and diffusion or recombination of charged carriers in electronic devices. Proper care can reduce the effect of internal noise but can never eliminate it. Noise is one of the basic factors that set limits on the rate of communication.

The **signal-to-noise ratio (SNR)** is defined as the ratio of signal power to noise power. The channel distorts the signal, and noise accumulates along the path. Worse yet, the signal strength decreases while the noise level increases with distance from the transmitter. Thus, the SNR is continuously decreasing along the length of the channel. Amplification of the received signal to make up for the attenuation is of no avail because the noise will be amplified in the same proportion, and the SNR remains, at best, unchanged.\*

## ANALOG AND DIGITAL MESSAGES

Messages are digital or analog. Digital messages are constructed with a finite number of symbols. For example, printed language consists of 26 letters, 10 numbers, a space, and several punctuation marks. Thus, a text is a digital message constructed from about 50 symbols. Human speech is also a digital message, because it is made up from a finite vocabulary in a language.† Similarly, a Morse-coded telegraph message is a digital message constructed from a set of only **two** symbols—mark and space. It is therefore a **binary** message, implying only two symbols. A digital message constructed with  $M$  symbols is called an **M-ary** message.

Analog messages, on the other hand, are characterized by data whose values vary over a continuous range. For example, the temperature or the atmospheric pressure of a certain

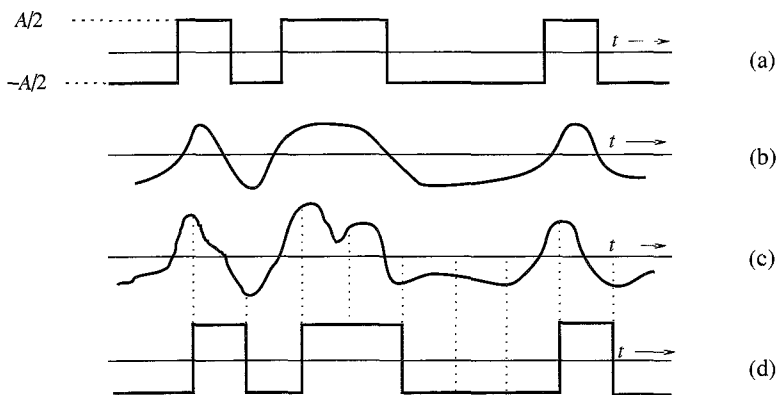
\* Actually, amplification further deteriorates the SNR because of the amplifier noise.

† Here we imply the printed text of the speech rather than its details such as the pronunciation of words and varying inflections, pitch, emphasis, and so on. The speech signal from a microphone contains all these details. This signal is an analog signal, and its information content is more than a thousand times the information in the written text of the same speech.

location can vary over a continuous range and can assume an infinite number of possible values. Similarly, a speech waveform has amplitudes that vary over a continuous range. Over a given time interval, an infinite number of possible different speech waveforms exist, in contrast to only a finite number of possible digital messages.

### Noise Immunity of Digital Signals

Digital messages are transmitted by using a finite set of electrical waveforms. For example, in the Morse code, a mark can be transmitted by an electrical pulse of amplitude  $A/2$ , and a space can be transmitted by a pulse of amplitude  $-A/2$ . In an  $M$ -ary case,  $M$  distinct electrical pulses (or waveforms) are used; each of the  $M$  pulses represents one of the  $M$  possible symbols. The task of the receiver is to extract a message from a distorted and noisy signal at the channel output. Message extraction is often easier from digital signals than from analog signals. Consider a binary case: Two symbols are encoded as rectangular pulses of amplitudes  $A/2$  and  $-A/2$ . The only decision at the receiver is the selection between two possible pulses received, not the details of the pulse shape. The decision is readily made with reasonable certainty even if the pulses are distorted and noisy (Fig. 1.3). The digital message in Fig. 1.3a is distorted by the channel, as shown in Fig. 1.3b. Yet, if the distortion is within limits, we can recover the data without error because we need only to make a simple binary decision as to whether the received pulse is positive or negative. Figure 1.3c shows the same data with channel distortion and noise. Here again, the data can be recovered correctly as long as the distortion and the noise are within limits. In contrast, the waveform in an analog message is important, and even a slight distortion or interference in the waveform will cause an error in the received signal. Clearly, a digital communication system is more rugged than an analog communication system in the sense that it can better withstand noise and distortion (as long as they are within a limit).



**Figure 1.3** (a) Transmitted signal. (b) Received distorted signal (without noise). (c) Received distorted signal (with noise). (d) Regenerated signal (delayed).

### Viability of Regenerative Repeaters in Digital Communication

The main reason for the superiority of digital systems over analog ones is the viability of **regenerative repeaters** in the former. Repeater stations are placed along the communication path of a digital system at distances short enough to ensure that noise and distortion remain within a limit. This allows pulse detection with high accuracy. At each repeater station the incoming pulses are detected and new clean pulses are transmitted to the next repeater station. This process prevents the accumulation of noise and distortion along the path by cleaning the pulses periodically at the repeater stations. We can thus transmit messages over longer

distances with greater accuracy. For analog systems, there is no way to avoid accumulation of noise and distortion along the path. As a result, the distortion and the noise interference are cumulative over the entire transmission path. To compound the difficulty, the signal is attenuated continuously over the transmission path. Thus, with increasing distance the signal becomes weaker, whereas the distortion and the noise become stronger. Ultimately, the signal, overwhelmed by the distortion and noise, is mutilated. Amplification is of little help, because it enhances the signal and the noise in the same proportion. Consequently, the distance over which an analog message can be transmitted is limited by the transmitter power. Despite these problems, analog communication was used widely and successfully in the past. Because of the advent of optical fiber and the dramatic cost reduction achieved in the fabrication of digital circuitry, almost all new communication systems being installed are digital. But the old analog communication facilities are also in use.

### Analog-to-Digital (A/D) Conversion

A meeting ground exists for analog and digital signals: conversion of analog signals to digital signals (A/D conversion). The frequency spectrum of a signal indicates relative magnitudes of various frequency components. The **sampling theorem** (to be proved in Chapter 6) states that if the highest frequency in the signal spectrum is  $B$  (in hertz), the signal can be reconstructed from its samples, taken at a rate not less than  $2B$  samples per second. This means that in order to transmit the information in a continuous-time signal, we need only transmit its samples (Fig. 1.4). Unfortunately, the sample values are still not digital because they lie in a continuous range and can take on any one of the infinite values in the range. We are back where we started! This difficulty is neatly resolved by what is known as **quantization**, where each sample is approximated, or “rounded off,” to the nearest quantized level, as shown in Fig. 1.4. Amplitudes of the signal  $m(t)$  lie in the range  $(-m_p, m_p)$ , which is partitioned into  $L$  intervals, each of magnitude  $\Delta v = 2m_p/L$ . Each sample amplitude is approximated to the midpoint of the interval in which the sample value falls. Each sample is now approximated to one of the  $L$  numbers. The information is thus digitized.

The quantized signal is an approximation of the original one. We can improve the accuracy of the quantized signal to any desired degree by increasing the number of levels  $L$ .

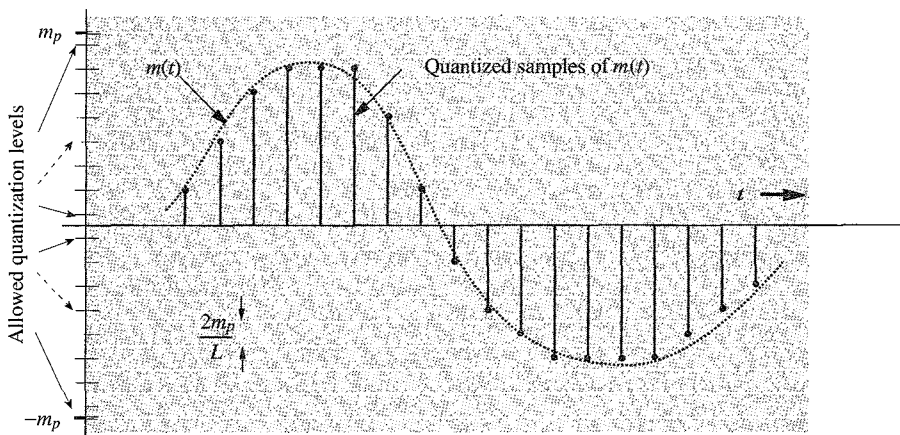


Figure 1.4 Analog-to-digital conversion of a signal.

For intelligibility of voice signals, for example,  $L = 8$  or  $16$  is sufficient. For commercial use,  $L = 32$  is a minimum, and for telephone communication,  $L = 128$  or  $256$  is commonly used.

During each sampling interval, we transmit one quantized sample, which takes on one of the  $L$  values. This requires  $L$  distinct waveforms, which may be constructed, for example, by using a basic rectangular pulse of amplitude  $A/2$  and its multiples (for instance,  $\pm A/2, \pm 3A/2, \pm 5A/2, \dots, \pm[(L-1)A/2]$ , as shown in Fig. 1.5) to form  $L$  distinct waveforms to be assigned to the  $L$  values to be transmitted. Amplitudes of any two of these waveforms are separated by at least  $A$  to guard against noise interference and channel distortion. Another possibility is to use fewer than  $L$  waveforms and form their combinations (codes) to yield  $L$  distinct patterns. As an example, for the case  $L = 16$  we may use 16 pulses ( $\pm A/2, \pm 3A/2, \dots, \pm 15A/2$ , as shown in Fig. 1.5). The second alternative is to use combinations of only two basic pulses,  $A/2$  and  $-A/2$ . A sequence of four such pulses gives  $2 \times 2 \times 2 \times 2 = 16$  distinct patterns, as shown in Fig. 1.6. We can assign one pattern to each of the 16 quantized values to be transmitted. Each quantized sample is now coded into a sequence of four binary pulses. This is the so-called binary case, where signaling is carried out by means of only two basic pulses (or symbols).\*

The binary case is of great practical importance because of its simplicity and ease of detection. Virtually all digital communication today is binary. This scheme of transmitting data by digitizing and then using pulse codes to transmit the digitized data is known as **pulse-code modulation (PCM)**.

A typical distorted binary signal with noise acquired over the channel is shown in Fig. 1.3. If  $A$  is sufficiently large compared to typical noise amplitudes, the receiver can still correctly distinguish between the two pulses. The pulse amplitude is typically 5 to 10 times the rms noise amplitude. For such a high SNR, the probability of error at the receiver is less than  $10^{-6}$ ; that is, on the average, the receiver will make less than one error per million pulses. The effect

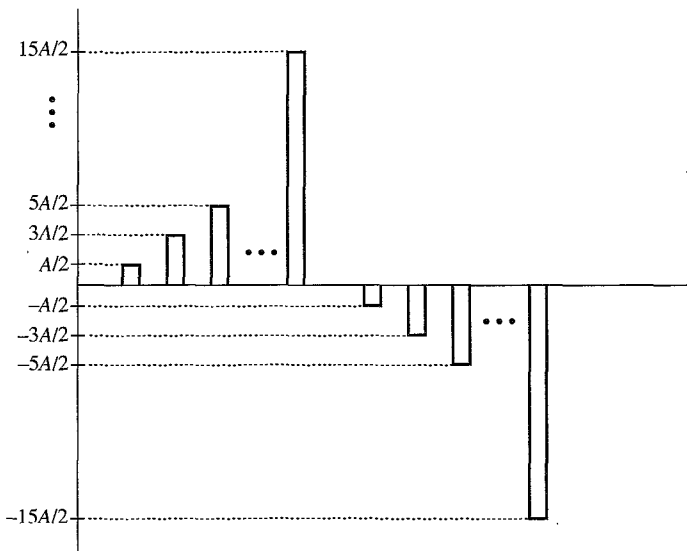


Figure 1.5 Multi-amplitude pulse code that uses  $L$  amplitude levels.

\* An intermediate case exists where we use four basic pulses (quaternary pulses) of amplitudes  $\pm A/2$  and  $\pm 3A/2$ . A sequence of two quaternary pulses can form  $4 \times 4 = 16$  distinct levels of values.

Digit	Binary equivalent	Pulse code waveform
0	0000	
1	0001	
2	0010	
3	0011	
4	0100	
5	0101	
6	0110	
7	0111	
8	1000	
9	1001	
10	1010	
11	1011	
12	1100	
13	1101	
14	1110	
15	1111	

Figure 1.6 Example of a binary pulse code.

of random channel noise and distortion is thus practically eliminated. Hence, when analog signals are transmitted by digital means, the only error, or uncertainty, in the received signal is that caused by quantization. By increasing  $L$ , we can reduce the uncertainty, or error, caused by quantization to any desired amount. At the same time, because of the use of regenerative repeaters, we can transmit signals over a much longer distance than would have been possible for the analog signal. As will be seen later in this chapter, the price for all these benefits of digital communication is paid in terms of increased bandwidth of transmission.

From all this discussion, we arrive at a rather interesting (and by no means obvious) conclusion—that every possible communication can be carried on with a minimum of two symbols. Thus, merely by using a proper sequence of a wink of the eye, one can convey any message, be it a conversation, a book, a movie, or an opera star’s singing. Every possible detail (such as various shades of colors of the objects and tones of the voice, etc.) that is reproducible on a movie screen or on the best quality color television can be conveyed with no less accuracy, merely by a wink of an eye.\*

Although PCM was invented by P. M. Rainey in 1926 and rediscovered by A. H. Reeves in 1939, it was not until the early sixties that Bell Laboratories installed the first communication

\* Of course, to convey the information in a movie or a television program in real time, the winking would have to be at an inhumanly high speed.

link using PCM. The cost and size of vacuum tube circuits were the chief impediments to the use of PCM in the early days. It was the transistor that made PCM practicable.

## SIGNAL-TO-NOISE RATIO, CHANNEL BANDWIDTH, AND THE RATE OF COMMUNICATION

The fundamental parameters that control the rate and quality of information transmission are the channel bandwidth  $B$  and the signal power  $S$ . The appropriate quantitative relationships will be derived later. Here we shall demonstrate these relationships qualitatively.

The **bandwidth** of a channel is the range of frequencies that it can transmit with reasonable fidelity. For example, if a channel can transmit with reasonable fidelity a signal whose frequency components occupy a range from 0 (dc) up to a maximum of 5000 Hz (5 kHz), the channel bandwidth  $B$  is 5 kHz.

To understand the role of  $B$ , consider the possibility of increasing the speed of information transmission by time compression of the signal. If a signal is compressed in time by a factor of 2, it can be transmitted in half the time, and the speed of transmission is doubled. Compression by a factor of 2, however, causes the signal to “wobble” twice as fast, implying that the frequencies of its components are doubled. To transmit this compressed signal without distortion, the channel bandwidth must also be doubled. Thus, the rate of information transmission is directly proportional to  $B$ . More generally, if a channel of bandwidth  $B$  can transmit  $N$  pulses per second, then to transmit  $KN$  pulses per second we need a channel of bandwidth  $KB$ . To reiterate, the number of pulses per second that can be transmitted over a channel is directly proportional to its bandwidth  $B$ .

The **signal power**  $S$  plays a dual role in information transmission. First,  $S$  is related to the quality of transmission. Increasing  $S$ , the signal power, reduces the effect of channel noise, and the information is received more accurately, or with less uncertainty. A larger signal-to-noise ratio (SNR) also allows transmission over a longer distance. In any event, a certain minimum SNR is necessary for communication.

The second role of the signal power is not as obvious, although it is very important. We shall demonstrate that the channel bandwidth  $B$  and the signal power  $S$  are exchangeable; that is, to maintain a given rate and accuracy of information transmission, we can trade  $S$  for  $B$ , and vice versa. Thus, one may reduce  $B$  if one is willing to increase  $S$ , or one may reduce  $S$  if one is willing to increase  $B$ . The rigorous proof of this will be provided in Chapter 15. Here we shall give only a “plausibility argument.”

Consider the PCM scheme discussed earlier, with 16 quantization levels ( $L = 16$ ). Here we may use 16 distinct pulses of amplitudes  $\pm A/2, \pm 3A/2, \dots, \pm 15A/2$  to represent the 16 levels (a 16-ary case). Each sample is transmitted by one of the 16 pulses during the sampling interval  $T_s$ . The amplitudes of these pulses range from  $-15A/2$  to  $15A/2$ . Alternately, we may use the binary scheme, where a group of four binary pulses is used to transmit each sample during the sampling interval  $T_s$ . In the latter case, the transmitted power is reduced considerably because the peak amplitude of transmitted pulses is only  $A/2$ , as compared to the peak amplitude  $15A/2$  in the 16-ary case. In the binary case, however, we need to transmit four pulses in each interval  $T_s$  instead of just one pulse required in the 16-ary case. Thus, the required channel bandwidth in the binary case is 4 times as great as that for the 16-ary case. Despite the fact that the binary case requires 4 times as many pulses, its power is less



than the power required for the 16-ary case by a factor of  $255/12 = 21.25$ , as shown later in Eq. 13.51a.\* In both cases, the minimum amplitude separation between transmitted pulses is  $A$ , and we therefore have about the same error probability at the receiver.† This means the quality of the received signal is about the same in both cases. In the binary case, the transmitted signal power is reduced at the cost of increased bandwidth. We have demonstrated here the exchangeability of  $S$  with  $B$ . Later we shall see that relatively little increase in  $B$  enables a significant reduction in  $S$ .

In conclusion, the two primary communication resources are the bandwidth and the transmitted power. In a given communication channel, one resource may be more valuable than the other, and the communication scheme should be designed accordingly. A typical telephone channel, for example, has a limited bandwidth (3 kHz), but a lot of power is available. On the other hand, in space vehicles, infinite bandwidth is available but the power is limited. Hence, the communication schemes required in the two cases are radically different.

Since the SNR is proportional to the power  $S$ , we can say that SNR and bandwidth are exchangeable. It will be shown in Chapter 15 that the relationship between the bandwidth expansion factor and the SNR is exponential. Thus, if a given rate of information transmission requires a channel bandwidth  $B_1$  and a signal-to-noise ratio  $\text{SNR}_1$ , then it is possible to transmit the same information over a channel bandwidth  $B_2$  and a signal-to-noise ratio  $\text{SNR}_2$ , where

$$\text{SNR}_2 \simeq \text{SNR}_1^{B_1/B_2} \quad (1.1)$$

Thus, if we double the channel bandwidth, the required SNR is just a square root of the former SNR, and tripling the channel bandwidth reduces the corresponding SNR to just a cube root of the former SNR. Therefore, a relatively small increase in channel bandwidth buys a large advantage in terms of reduced transmission power. But a large increase in transmitted power buys a meager advantage in bandwidth reduction. Hence, in practice, the exchange between  $B$  and SNR is usually in the sense of increasing  $B$  to reduce transmitted power, and rarely the other way around.

Equation (1.1) gives the upper bound on the exchange between SNR and  $B$ . Not all systems are capable of achieving this bound. For example, frequency modulation (FM) is one scheme that is commonly used in radio broadcasting for improving the signal quality at the receiver by increasing the transmission bandwidth. We shall see that an FM system does not make efficient use of bandwidth in reducing the required SNR, and its performance falls far short of that in Eq. (1.1). PCM, on the other hand, comes close (within 10 dB) to realizing the performance in Eq. (1.1). Generally speaking, the transmission of signals in digital form comes much closer to the realization of the limit in Eq. (1.1) than does the transmission of signals in analog form.

The limitation imposed on communication by the channel bandwidth and the SNR is dramatically highlighted by Shannon's equation,‡

$$C = B \log_2(1 + \text{SNR}) \quad \text{bit/s} \quad (1.2)$$

---

\* To explain this behavior qualitatively, let the number of symbols used be  $M$  ( $M = 16$  in the present case) instead of 2 (binary case). We shall see later that the power of a pulse is proportional to its amplitude. Hence, the signal power increases as  $(M - 1)^2$ , but  $n$ , the number of binary pulses per sample, increases only as the logarithm of  $M$ .

† Not quite true! We use this approximation to keep our argument simple and nonquantitative at this point.

‡ This is true for a certain kind of noise—white gaussian noise.

Here  $C$  is the rate of information transmission per second. This rate  $C$  (known as the channel capacity) is the maximum number of binary symbols (bits) that can be transmitted per second with a probability of error arbitrarily close to zero. In other words, a channel can transmit  $B \log_2 (1 + \text{SNR})$  binary digits, or symbols, per second as accurately as one desires. Moreover, it is impossible to transmit at a rate higher than this without incurring errors. Shannon's equation clearly brings out the limitation on the rate of communication imposed by  $B$  and SNR. If there were no noise on the channel ( $N = 0$ ),  $C = \infty$ , and communication would cease to be a problem. We could then transmit any amount of information in the world over a channel. This can be readily verified. If noise were zero, there would be no uncertainty in the received pulse amplitude, and the receiver would be able to detect any pulse amplitude without ambiguity. The minimum pulse-amplitude separation  $A$  can be arbitrarily small, and for any given pulse, we have an infinite number of levels available. We can assign one level to every possible message. For example, the contents of this book will be assigned one level; if it is desired to transmit this book, all that is needed is to transmit one pulse of that level. Because an infinite number of levels are available, it is possible to assign one level to any conceivable message. Cataloging of such a code may not be practical, but that is beside the point. The point is that if the noise is zero, communication ceases to be a problem, at least theoretically. Implementation of such a scheme would be difficult because of the requirement of generation and detection of pulses of precise amplitudes. Such practical difficulties would then set a limit on the rate of communication.

In conclusion, we have demonstrated qualitatively the basic role played by  $B$  and SNR in limiting the performance of a communication system. These two parameters then represent the ultimate limitation on a rate of communication. We have also demonstrated the possibility of trade or exchange between these two basic parameters.

Equation (1.1) can be derived from Eq. (1.2). It should be remembered that Shannon's result represents the upper limit on the rate of communication over a channel and can be achieved only with a system of monstrous and impractical complexity, and with a time delay in reception approaching infinity. Practical systems operate at rates below the Shannon rate. In Chapter 15, we shall derive Shannon's result and compare the efficiencies of various communication systems.

## MODULATION

Baseband signals produced by various information sources are not always suitable for direct transmission over a given channel. These signals are usually further modified to facilitate transmission. This conversion process is known as **modulation**. In this process, the baseband signal is used to modify some parameter of a high-frequency carrier signal.

A **carrier** is a sinusoid of high frequency, and one of its parameters—such as amplitude, frequency, or phase—is varied in proportion to the baseband signal  $m(t)$ . Accordingly, we have amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM). Figure 1.7 shows a baseband signal  $m(t)$  and the corresponding AM and FM waveforms. In AM, the carrier amplitude varies in proportion to  $m(t)$ , and in FM, the carrier frequency varies in proportion to  $m(t)$ .

At the receiver, the modulated signal must pass through a reverse process called **demodulation** in order to reconstruct the baseband signal.

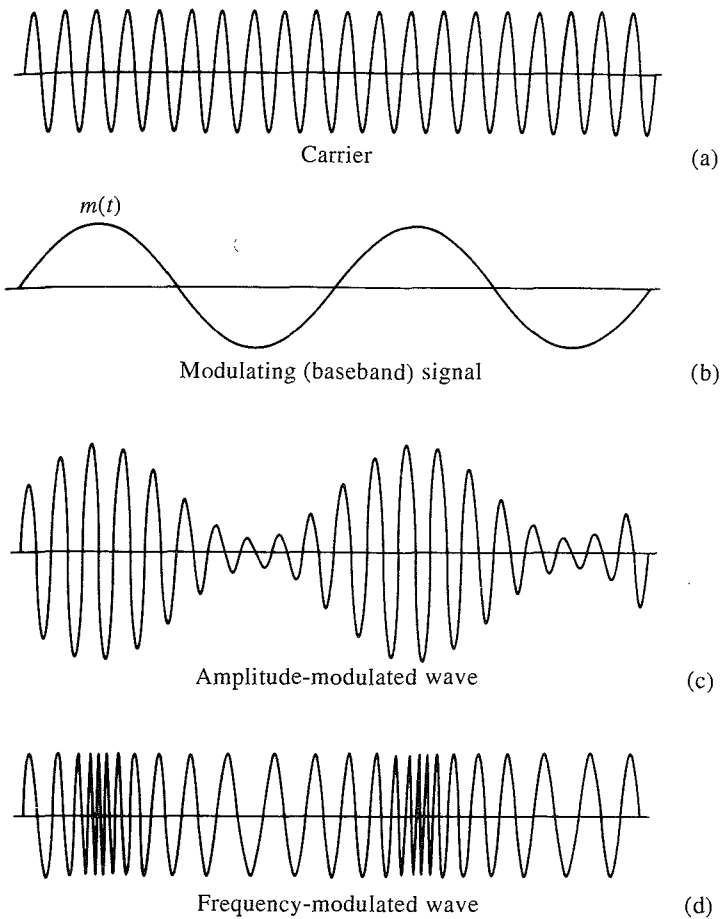


Figure 1.7 Modulation.

As mentioned earlier, modulation is used to facilitate transmission. Some of the important reasons for modulation are given next.

### Ease of Radiation

For efficient radiation of electromagnetic energy, the radiating antenna should be on the order of one-tenth or more of the wavelength of the signal radiated. For many baseband signals, the wavelengths are too large for reasonable antenna dimensions. For example, the power in a speech signal is concentrated at frequencies in the range of 100 to 3000 Hz. The corresponding wavelength is 100 to 3000 km. This long wavelength would necessitate an impracticably large antenna. Instead, we modulate a high-frequency carrier, thus translating the signal spectrum to the region of carrier frequencies that corresponds to a much smaller wavelength. For example, a 1-MHz carrier has a wavelength of only 300 m and requires an antenna whose size is on the order of 30 m. In this aspect, modulation is like letting the baseband signal hitchhike on a high-frequency sinusoid (carrier). The carrier and the baseband signal may be compared to a stone and a piece of paper. If we wish to throw a piece of paper, it cannot go too far by itself. But by wrapping it around a stone (a carrier), it can be thrown over a longer distance.

### Simultaneous Transmission of Several Signals

Consider the case of several radio stations broadcasting audio baseband signals directly, without any modification. They would interfere with each other because the spectra of all the signals occupy more or less the same bandwidth. Thus, it would be possible to broadcast from only one radio or television station at a time. This is wasteful because the channel bandwidth may be much larger than that of the signal. One way to solve this problem is to use modulation. We can use various audio signals to modulate different carrier frequencies, thus translating each signal to a different frequency range. If the various carriers are chosen sufficiently far apart in frequency, the spectra of the modulated signals will not overlap and thus will not interfere with each other. At the receiver, one can use a tunable bandpass filter to select the desired station or signal. This method of transmitting several signals simultaneously is known as **frequency-division multiplexing (FDM)**. Here the bandwidth of the channel is shared by various signals without any overlapping.

Another method of multiplexing several signals is known as **time-division multiplexing (TDM)**. This method is suitable when a signal is in the form of a pulse train (as in PCM). The pulses are made narrower, and the spaces that are left between pulses are used for pulses from other signals. Thus, in effect, the transmission time is shared by a number of signals by interleaving the pulse trains of various signals in a specified order. At the receiver, the pulse trains corresponding to various signals are separated.

### Effecting the Exchange of SNR with $B$

We have shown earlier that it is possible to exchange SNR with the bandwidth of transmission. FM or PM can effect such an exchange. The amount of modulation (to be defined later) used controls the exchange of SNR and the transmission bandwidth.

## RANDOMNESS, REDUNDANCY, AND CODING

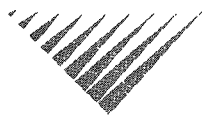
**Randomness** plays an important role in communication. As noted earlier, one of the limiting factors in the rate of communication is noise, which is a random signal. Randomness is also closely associated with information. Indeed, randomness is the essence of communication. Randomness means unpredictability, or uncertainty, of the outcome. If a source had no unpredictability, or uncertainty, it would be known beforehand and would convey no information. Probability is the measure of certainty, and information is associated with probability. If a person winks, it conveys some information in a given context. But if a person were to wink continuously with the regularity of a clock, it would convey no meaning. The unpredictability of the winking is what gives the information to the signal. What is more interesting, however, is that from the engineering point of view, also, information is associated with uncertainty. The information of a message, from the engineering point of view, is defined as a quantity proportional to the minimum time needed to transmit it. Consider the Morse code, for example. In this code, various combinations of marks and spaces (code words) are assigned to each letter. In order to minimize the transmission time, shorter code words are assigned to more frequently occurring (more probable) letters (such as  $e$ ,  $t$ , and  $a$ ) and longer code words are assigned to rarely occurring (less probable) letters (such as  $x$ ,  $q$ , and  $z$ ). Thus, the time required to transmit a message is closely related to the probability of its occurrence. It will be shown in Chapter 15 that for digital signals, the overall transmission time is minimized if a message (or symbol)

of probability  $P$  is assigned a code word with a length proportional to  $\log(1/P)$ . Hence, from an engineering point of view, the information of a message with probability  $P$  is proportional to  $\log(1/P)$ .

**Redundancy** also plays an important role in communication. It is essential for reliable communication. Because of redundancy, we are able to decode a message accurately despite errors in the received message. Redundancy thus helps combat noise. All languages are redundant. For example, English is about 50 percent redundant; that is, on the average, we may throw out half of the letters or words without destroying the message. This also means that in any English message, the speaker or the writer has free choice over half the letters or words, on the average. The remaining half is determined by the statistical structure of the language. If all the redundancy of English were removed, it would take about half the time to transmit a telegram or telephone conversation. If an error occurs at the receiver, however, it would be rather difficult to make sense out of the received message. The redundancy in a message, therefore, plays a useful role in combating the noise in the channel. This same principle of redundancy applies in coding messages. A deliberate redundancy is used to combat the noise. For example, in order to transmit samples with  $L = 16$  quantizing levels, we may use a group of four binary pulses, as shown in Fig. 1.6. In this coding scheme, no redundancy exists. If an error occurs in the reception of even one of the pulses, the receiver will produce a wrong value. Here we may use redundancy to eliminate the effect of possible errors caused by channel noise or imperfections. Thus, if we add to each code word one more pulse of such polarity as to make the number of positive pulses even, we have a code that can detect a single error in any place. Thus, to the code words **0001** we add a fifth pulse, of positive polarity, to make a new code word, **00011**. Now the number of positive pulses is 2 (even). If a single error occurs in any position, this parity will be violated. The receiver knows that an error has been made and can request retransmission of the message. This is a very simple coding scheme. It can only detect an error, but cannot locate it. Moreover, it cannot detect an even number of errors. By introducing more redundancy, it is possible not only to detect but also to correct errors. For example, for  $L = 16$ , it can be shown that properly adding three pulses will not only detect but also correct a single error occurring at any location. This subject of error-correcting codes will be discussed in Chapter 16.

# 2

# INTRODUCTION TO SIGNALS



**I**n this chapter we discuss certain basic signal concepts. Signals are processed by systems. We shall start with explaining the terms *signals* and *systems*.

## Signals

A **signal**, as the term implies, is a set of information or data. Examples include a telephone or a television signal, monthly sales of a corporation, or the daily closing prices of a stock market (e.g., the Dow Jones averages). In all these examples, the signals are functions of the independent variable *time*. This is not always the case, however. When an electrical charge is distributed over a surface, for instance, the signal is the charge density, a function of *space* rather than time. In this book we deal almost exclusively with signals that are functions of time. The discussion, however, applies equally well to other independent variables.

## Systems

Signals may be processed further by **systems**, which may modify them or extract additional information from them. For example, an antiaircraft gun operator may want to know the future location of a hostile moving target, which is being tracked by a radar. Knowing the radar signal, the antiaircraft gun operator knows the past location and velocity of the target. By properly processing the radar signal (the input), we can approximately estimate the future location of the target. Thus, a system is an entity that *processes* a set of signals (**inputs**) to yield another set of signals (**outputs**). A system may be made up of physical components, as in electrical, mechanical, or hydraulic systems (hardware realization), or it may be an algorithm that computes an output from an input signal (software realization).

## 2.1 SIZE OF A SIGNAL

The size of any entity is a number that indicates the largeness or strength of that entity. Generally speaking, the signal amplitude varies with time. How can a signal that exists over a certain time interval with varying amplitude be measured by one number that will indicate the signal

size or signal strength? Such a measure must consider not only the signal amplitude, but also its duration. For instance, if we are to devise a single number  $V$  as a measure of the size of a human being, we must consider not only his or her width (girth), but also the height. The product of girth and height is a reasonable measure of the size of a person. If we wish to be a little more precise, we should average this product over the entire length of the person. If we make the simplifying assumption that the shape of a person is a cylinder of radius  $r$ , which varies with the height  $h$  of the person, then a reasonable measure of the size of a person of height  $H$  is the person's volume  $V$ , given by

$$V = \pi \int_0^H r^2(h) dh$$

### Signal Energy

Arguing in this manner, we may consider the area under a signal  $g(t)$  as a possible measure of its size, because it takes account of not only the amplitude, but also the duration. However, this will be a defective measure because  $g(t)$  could be a large signal, yet its positive and negative areas could cancel each other, indicating a signal of small size. This difficulty can be corrected by defining the signal size as the area under  $g^2(t)$ , which is always positive. We call this measure the **signal energy**  $E_g$ , defined (for a real signals) as

$$E_g = \int_{-\infty}^{\infty} g^2(t) dt \quad (2.1)$$

This definition can be generalized to a complex valued signal  $g(t)$  as

$$E_g = \int_{-\infty}^{\infty} |g(t)|^2 dt \quad (2.2)$$

There are also other possible measures of signal size, such as the area under  $|g(t)|$ . The above energy measure, however, is not only more tractable mathematically, but is also more meaningful (as shown later) in the sense that it is indicative of the energy that can be extracted from the signal.

### Signal Power

The signal energy must be finite for it to be a meaningful measure of the signal size. A necessary condition for the energy to be finite is that the signal amplitude  $\rightarrow 0$  as  $|t| \rightarrow \infty$  (Fig. 2.1a). Otherwise the integral in Eq. (2.1) will not converge.

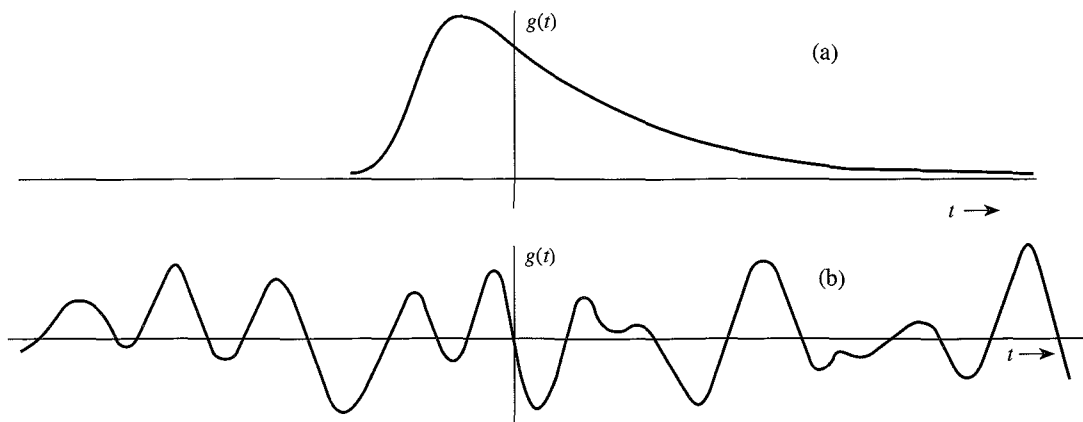
If the amplitude of  $g(t)$  does not  $\rightarrow 0$  as  $|t| \rightarrow \infty$  (Fig. 2.1b), the signal energy is infinite. A more meaningful measure of the signal size in such a case would be the time average of the energy (if it exists), which is the average power  $P_g$  defined (for a real signal) by

$$P_g = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t) dt \quad (2.3)$$

We can generalize this definition for a complex signal  $g(t)$  as

$$P_g = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} |g(t)|^2 dt \quad (2.4)$$

Observe that the signal power  $P_g$  is the time average (mean) of the signal amplitude squared, that is the **mean-squared** value of  $g(t)$ . Indeed, the square root of  $P_g$  is the familiar **root mean square (rms)** value of  $g(t)$ .



**Figure 2.1** Examples of signals. (a) Signal with finite energy. (b) Signal with finite power.

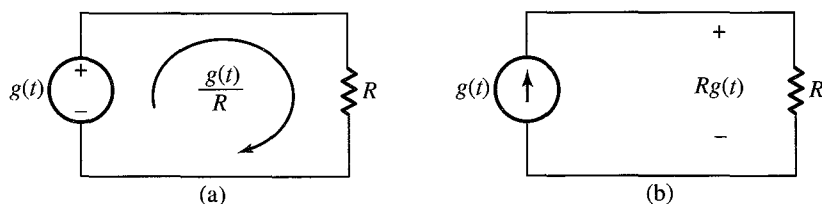
The mean of an entity averaged over a large time interval approaching infinity exists if the entity is either periodic or has a statistical regularity. If such a condition is not satisfied, the average may not exist. For instance, a ramp signal  $g(t) = t$  increases indefinitely as  $|t| \rightarrow \infty$ , and neither the energy nor the power exists for this signal.

### Comments

The signal energy as defined in Eq. (2.1) or Eq. (2.2) does not indicate the actual energy of the signal because the signal energy depends not only on the signal, but also on the load. It can, however, be interpreted as the energy dissipated in a normalized load of a 1-ohm resistor. If a voltage  $g(t)$  is applied across an  $R$ -ohm resistor, the current through the resistor is  $g(t)/R$  (Fig. 2.2a), and the instantaneous power dissipated would be  $v(t)i(t) = g^2(t)/R$ . The energy dissipated, being the integral of the instantaneous power, is

$$\text{Energy dissipated} = \int_{-\infty}^{\infty} \frac{g^2(t)}{R} dt = \frac{E_g}{R} \quad (2.5)$$

If  $R = 1$ , the energy dissipated in the resistor is  $E_g$ . Thus, the signal energy  $E_g$  could be interpreted as the energy dissipated in a unit resistor if a voltage  $g(t)$  were applied across this unit resistor. From Fig. 2.2b, it follows that  $E_g$  may also be interpreted as the energy dissipated in a unit resistor if a current  $g(t)$  were passed through this unit resistor. Parallel observation applies to signal power as defined in Eq. (2.3) or Eq. (2.4).



**Figure 2.2** Computation of the actual energy dissipated across a load.



The measure of “energy” (or “power”) is therefore indicative of the energy (or power) capability of the signal, and not the actual energy. For this reason the concepts of conservation of energy should not be applied to the measure of signal energy. These measures are but convenient indicators of the signal size. For instance, if we approximate a signal  $g(t)$  by another signal  $z(t)$ , the error in the approximation is  $e(t) = g(t) - z(t)$ . The energy (or power) of  $e(t)$  is a convenient indicator of the goodness of the approximation. It provides us with a quantitative measure of determining the closeness of the approximation. It also allows us to determine if one approximation is better than the other. In communication systems, during transmission over a channel, message signals are corrupted by unwanted signals (noise). The quality of the received signal is judged by the relative sizes of the desired signal and the unwanted signal (noise). In this case the ratio of the message signal and the noise signal powers (SNR) is a good indication of the received signal quality.

**Units of Energy and Power:** Equations (2.1) and (2.2) are not correct dimensionally. This is because here we are using the term *energy* not in its conventional sense, but to indicate the signal size. The same observations apply to Eqs. (2.3) and (2.4) for power. In the present context the units of energy and power depend on the nature of the signal  $g(t)$ . If  $g(t)$  is a voltage signal, its energy  $E_g$  has units of volts squared seconds, and its power  $P_g$  has units of volts squared. If  $g(t)$  is a current signal, these units will be amperes squared seconds, and amperes squared, respectively.

**EXAMPLE 2.1** Determine the suitable measures of the signals in Fig. 2.3.

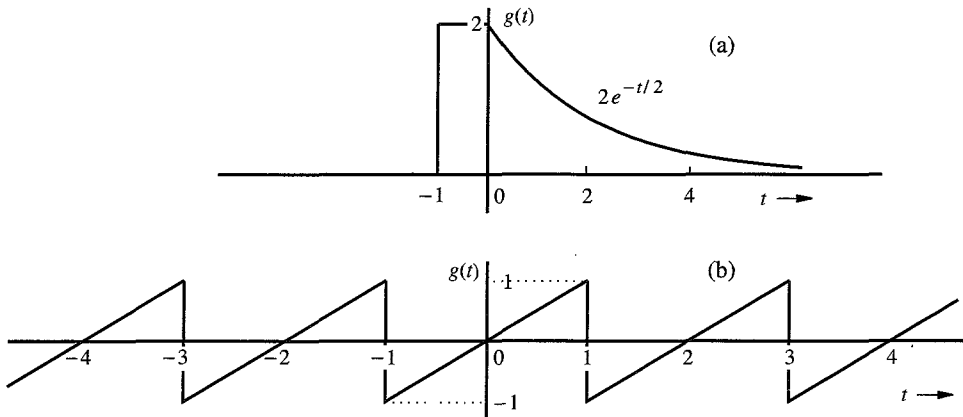


Figure 2.3 Signal for Example 2.1.

The signal in Fig. 2.3a  $\rightarrow 0$  as  $|t| \rightarrow \infty$ . Therefore, the suitable measure for this signal is its energy  $E_g$  given by

$$E_g = \int_{-\infty}^{\infty} g^2(t) dt = \int_{-1}^0 (2)^2 dt + \int_0^{\infty} 4e^{-t} dt = 4 + 4 = 8$$

The signal in Fig. 2.3b does not  $\rightarrow 0$  as  $|t| \rightarrow \infty$ . However, it is periodic, and therefore its power exists. We can use Eq. (2.3) to determine its power. We can simplify the procedure

for periodic signals by observing that a periodic signal repeats regularly each period (2 seconds in this case). Therefore, averaging  $g^2(t)$  over an infinitely large interval is identical to averaging it over one period (2 seconds in this case). Thus,

$$P_g = \frac{1}{2} \int_{-1}^1 g^2(t) dt = \frac{1}{2} \int_{-1}^1 t^2 dt = \frac{1}{3}$$

Recall that the signal power is the square of its rms value. Therefore, the rms value of this signal is  $1/\sqrt{3}$ .

**EXAMPLE 2.2** Determine the power and the rms value of:

- (a)  $g(t) = C \cos(\omega_0 t + \theta)$
- (b)  $g(t) = C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2) \quad \omega_1 \neq \omega_2$
- (c)  $g(t) = D e^{j\omega_0 t}$

(a) This is a periodic signal with period  $T_0 = 2\pi/\omega_0$ . The suitable measure of this signal is its power. Because it is a periodic signal, we may compute its power by averaging its energy over one period  $2\pi/\omega_0$ . However, for the sake of generality, we shall solve this problem by averaging over an infinitely large time interval using Eq (2.3),

$$\begin{aligned} P_g &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} C^2 \cos^2(\omega_0 t + \theta) dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \frac{C^2}{2} [1 + \cos(2\omega_0 t + 2\theta)] dt \\ &= \lim_{T \rightarrow \infty} \frac{C^2}{2T} \int_{-T/2}^{T/2} dt + \lim_{T \rightarrow \infty} \frac{C^2}{2T} \int_{-T/2}^{T/2} \cos(2\omega_0 t + 2\theta) dt \end{aligned}$$

The first term on the right-hand side is equal to  $C^2/2$ . Moreover, the second term is zero because the integral appearing in this term represents the area under a sinusoid over a very large time interval  $T$  with  $T \rightarrow \infty$ . This area is at most equal to the area of half the cycle because of cancellations of the positive and negative areas of a sinusoid. The second term is this area multiplied by  $C^2/2T$  with  $T \rightarrow \infty$ . Clearly this term is zero, and

$$P_g = \frac{C^2}{2} \tag{2.6a}$$

This shows that a sinusoid of amplitude  $C$  has a power  $C^2/2$  regardless of the value of its frequency  $\omega_0$  ( $\omega_0 \neq 0$ ) and phase  $\theta$ . The rms value is  $C/\sqrt{2}$ . If the signal frequency is zero (dc or a constant signal of amplitude  $C$ ), the reader can show that the power is  $C^2$ .

(b) In this case,

$$\begin{aligned} P_g &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} [C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2)]^2 dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} C_1^2 \cos^2(\omega_1 t + \theta_1) dt \end{aligned}$$

$$\begin{aligned}
& + \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} C_2^2 \cos^2(\omega_2 t + \theta_2) dt \\
& + \lim_{T \rightarrow \infty} \frac{2C_1 C_2}{T} \int_{-T/2}^{T/2} \cos(\omega_1 t + \theta_1) \cos(\omega_2 t + \theta_2) dt
\end{aligned}$$

Observe that the first and second integrals on the right-hand side are the powers of the two sinusoids, which are  $C_1^2/2$  and  $C_2^2/2$  as found in part (a). We now show that the third term on the right-hand side is zero. Using a trigonometric identity, this term, which is the product of the two sinusoids, is equal to the sum of two sinusoids or frequencies  $\omega_1 + \omega_2$  and  $\omega_1 - \omega_2$ . Thus, the third term is  $2C_1 C_2/T$  times the sum of the areas under two sinusoids. Now the area under any sinusoid over a large time interval is at most equal to the area under half the cycle because of cancellations of positive and negative areas as argued in part (a). So the third term vanishes because  $T \rightarrow \infty$ , and we have\*

$$P_g = \frac{C_1^2}{2} + \frac{C_2^2}{2} \quad (2.6b)$$

and the rms value is  $\sqrt{(C_1^2 + C_2^2)/2}$ .

We can readily extend this result to a sum of any number of sinusoids with distinct frequencies  $\omega_n$  ( $\omega_n \neq 0$ ). Thus, if

$$g(t) = \sum_{n=1}^{\infty} C_n \cos(\omega_n t + \theta_n)$$

where none of the two sinusoids have identical frequencies, then

$$P_g = \frac{1}{2} \sum_{n=1}^{\infty} C_n^2 \quad (2.6c)$$

(c) In this case the signal is complex, and we use Eq. (2.4) to compute the power. However, because this signal is periodic, we need average it only over a period  $T_0$ . Thus,

$$P_g = \frac{1}{T_0} \int_0^{T_0} |D e^{j\omega_0 t}|^2 dt$$

Recall that  $|e^{j\omega_0 t}| = 1$  so that  $|D e^{j\omega_0 t}|^2 = |D|^2$ , and

$$P_g = \frac{|D|^2}{T_0} \int_0^{T_0} dt = |D|^2 \quad (2.6d)$$

The rms value is  $|D|$ .

*Comments:* In part (b) we have shown that the power of the sum of two sinusoids is equal to the sum of the powers of the sinusoids. It appears that the power of  $g_1(t) + g_2(t)$  is  $P_{g_1} + P_{g_2}$ . Be cautioned against such a generalization. All we have proved here is that this is true if the two signals  $g_1(t)$  and  $g_2(t)$  happen to be sinusoids. It is not true in general. In fact, it is not true even for the sinusoids if the two sinusoids are of the same frequency.

\* This is true only if  $\omega_1 \neq \omega_2$ . If  $\omega_1 = \omega_2$ , the integrand of the third term is a nonnegative entity, and the integral in the third term  $\rightarrow \infty$  as  $T \rightarrow \infty$ .

We shall show in Sec. 2.5.3 that only under a certain condition (called orthogonality condition) the power (or energy) of  $g_1(t) + g_2(t)$  is equal to the sum of the powers (or energies) of  $g_1(t)$  and  $g_2(t)$ .

---

## 2.2 CLASSIFICATION OF SIGNALS

There are several classes of signals. Here we shall consider only the following classes, which are suitable for the scope of this book:

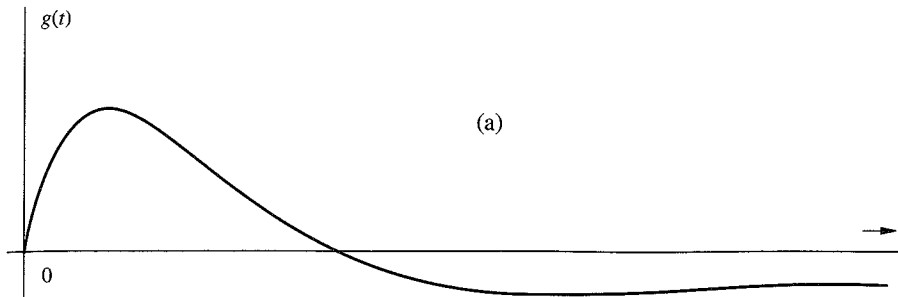
1. Continuous-time and discrete-time signals
2. Analog and digital signals
3. Periodic and aperiodic signals
4. Energy and power signals
5. Deterministic and probabilistic signals

### 2.2.1 Continuous-Time and Discrete-Time Signals

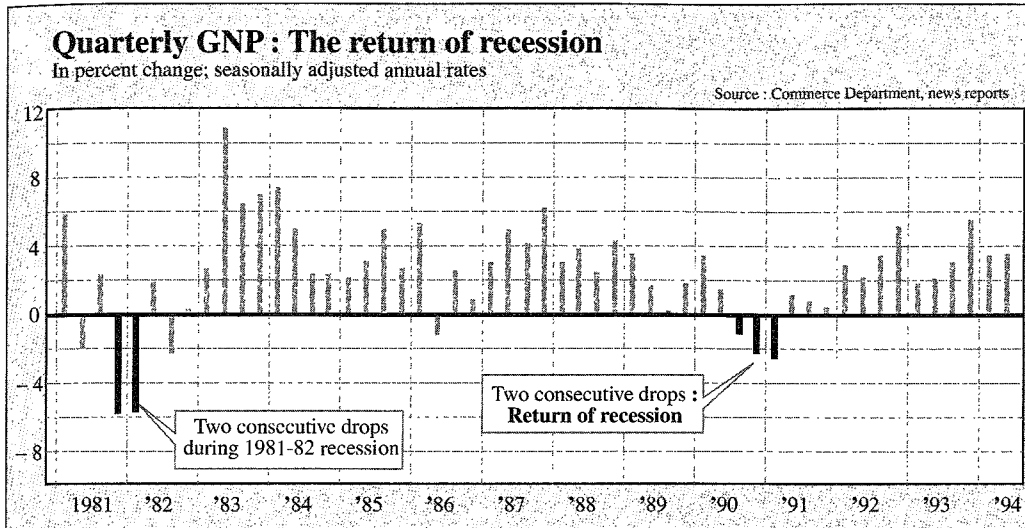
A signal that is specified for every value of time  $t$  (Fig. 2.4a) is a **continuous-time signal**, and a signal that is specified only at discrete values of  $t$  (Fig. 2.4b) is a **discrete-time signal**. Telephone and video camera outputs are continuous-time signals, whereas the quarterly gross national product (GNP), monthly sales of a corporation, and stock market daily averages are discrete-time signals.

### 2.2.2 Analog and Digital Signals

The concept of continuous time is often confused with that of analog. The two are not the same. The same is true of the concepts of discrete time and digital. A signal whose amplitude can take on any value in a continuous range is an **analog signal**. This means that an analog signal amplitude can take on an infinite number of values. A **digital signal**, on the other hand, is one whose amplitude can take on only a finite number of values. Signals associated with a digital computer are digital because they take on only two values (binary signals). For a signal to qualify as digital, the number of values need not be restricted to two. It can be any finite number. A digital signal whose amplitudes can take on  $M$  values is an  **$M$ -ary signal** of which binary ( $M = 2$ ) is a special case. The terms *continuous time* and *discrete time* qualify the nature of a signal along the time (horizontal) axis. The terms *analog* and *digital*, on the other hand, qualify the nature of the signal amplitude (vertical axis). Figure 2.5 shows examples of various types of signals. It is clear that analog is not necessarily continuous time and digital need not be discrete time. Figure 2.5c shows an example of an analog but discrete-time signal. An analog signal can be converted into a digital signal [analog-to-digital (A/D) conversion] through quantization (rounding off), as explained in Sec. 6.2.



(a)



(b)

Figure 2.4 Continuous-time and discrete-time signals.

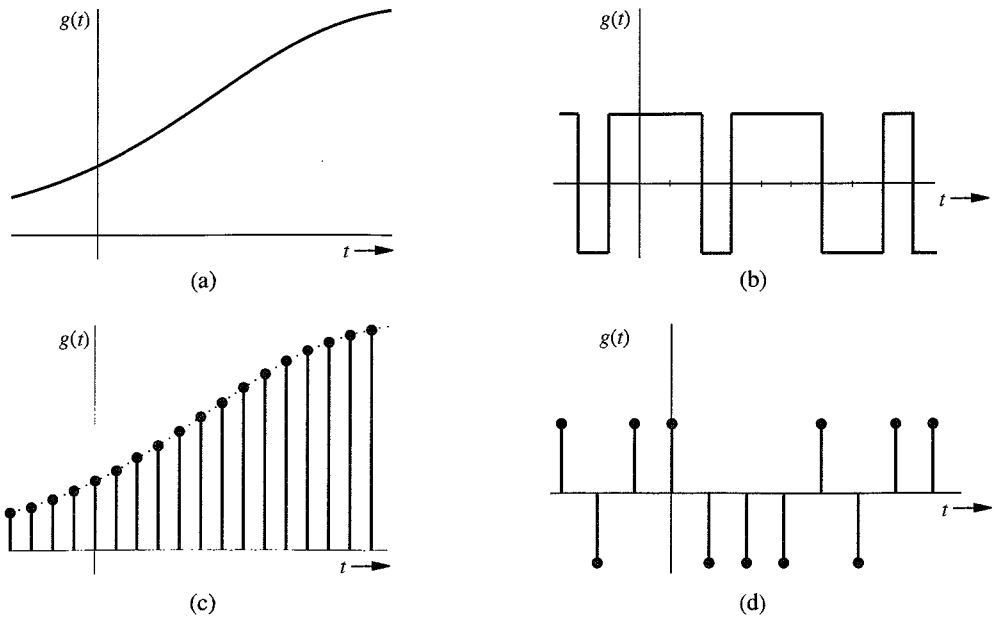
### 2.2.3 Periodic and Aperiodic Signals

A signal  $g(t)$  is said to be **periodic** if for some positive constant  $T_0$ ,

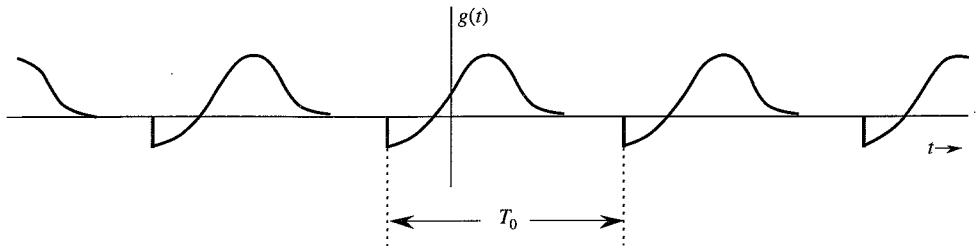
$$g(t) = g(t + T_0) \quad \text{for all } t \quad (2.7)$$

The *smallest* value of  $T_0$  that satisfies the periodicity condition (2.7) is the **period** of  $g(t)$ . The signal in Fig. 2.3b is a periodic signal with period 2. A signal is **aperiodic** if it is not periodic. The signal in Fig. 2.3a is aperiodic.

By definition, a periodic signal  $g(t)$  remains unchanged when time-shifted by one period. This means that a periodic signal must start at  $t = -\infty$  because if it starts at some finite instant, say,  $t = 0$ , the time-shifted signal  $g(t + T_0)$  will start at  $t = -T_0$  and  $g(t + T_0)$  would not be the same as  $g(t)$ . Therefore, a *periodic signal, by definition, must start at  $-\infty$  and continue forever*, as shown in Fig. 2.6. Observe that a periodic signal shifted by an integral multiple of



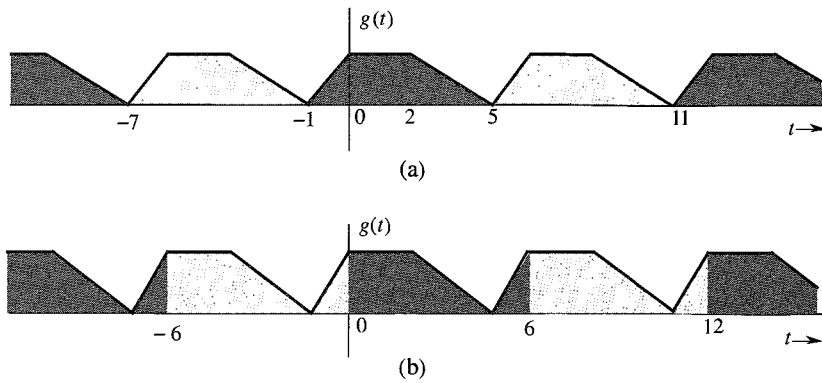
**Figure 2.5** Examples of signals. (a) Analog, continuous time. (b) Digital, continuous time. (c) Analog, discrete time. (d) Digital, discrete time.



**Figure 2.6** Periodic signal of period  $T_0$ .

$T_0$  remains unchanged. Therefore,  $g(t)$  may be considered a periodic signal with period  $mT_0$ , where  $m$  is any integer. However, by definition, the period is the smallest interval that satisfies periodicity condition (2.7). Therefore,  $T_0$  is the period.

*The second important property of a periodic signal  $g(t)$  is that  $g(t)$  can be generated by periodic extension of any segment of  $g(t)$  of duration  $T_0$  (the period).* This means that we can generate  $g(t)$  from any segment of  $g(t)$  with a duration of one period by placing this segment and the reproduction thereof end to end ad infinitum on either side. Figure 2.7 shows a periodic signal  $g(t)$  of period  $T_0 = 6$ . The shaded portion of Fig. 2.7a shows a segment of  $g(t)$  starting at  $t = -1$  and having a duration of one period (6 seconds). This segment, when repeated forever in either direction, results in the periodic signal  $g(t)$ . Figure 2.7b shows another shaded segment of  $g(t)$  of duration  $T_0$  starting at  $t = 0$ . Again we see that this segment,



**Figure 2.7** Generation of a periodic signal by periodic extension of its segment of one-period duration.

when repeated forever on either side, results in  $g(t)$ . The reader can verify that this is possible with any segment of  $g(t)$  starting at any instant as long as the segment duration is one period.

## 2.2.4 Energy and Power Signals

A signal with finite energy is an **energy signal**, and a signal with finite power is a **power signal**. In other words, a signal  $g(t)$  is an energy signal if

$$\int_{-\infty}^{\infty} |g(t)|^2 dt < \infty \quad (2.8)$$

Similarly, a signal with a finite and nonzero power (mean square value) is a power signal. In other words, a signal is a power signal if

$$0 < \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} |g(t)|^2 dt < \infty \quad (2.9)$$

Signals in Fig. 2.3a and b are examples of energy and power signals, respectively. Observe that power is the time average of the energy. Since the averaging is over an infinitely large interval, a signal with finite energy has zero power, and a signal with finite power has infinite energy. Therefore, a signal cannot both be an energy and a power signal. If it is one, it cannot be the other. On the other hand, there are signals that are neither energy nor power signals. The ramp signal is such an example.

### Comments

Every signal that can be generated in a lab has a finite energy. In other words, *every signal observed in real life is an energy signal. A power signal, on the other hand, must necessarily have an infinite duration.* Otherwise its power, which is its average energy (averaged over an infinitely large interval) will not approach a (nonzero) limit. Obviously it is impossible to generate a true power signal in practice because such a signal has infinite duration and infinite energy.

Also because of periodic repetition, periodic signals for which the area under  $|g(t)|^2$  over one period is finite are power signals; however, not all power signals are periodic.

## 2.2.5 Deterministic and Random Signals

A signal whose physical description is known completely, in either a mathematical form or a graphical form, is a **deterministic signal**. If a signal is known only in terms of probabilistic description, such as mean value, mean squared value, and so on, rather than its complete mathematical or graphical description, is a **random signal**. Most of the noise signals encountered in practice are random signals. All message signals are random signals because, as will be shown later, a signal, to convey information, must have some uncertainty (randomness) about it. The treatment of random signals will be discussed in Chapter 11.

## 2.3 SOME USEFUL SIGNAL OPERATIONS

We discuss here three useful signal operations: shifting, scaling, and inversion. Since the independent variable in our signal description is time, these operations are discussed as time shifting, time scaling, and time inversion (or folding). However, this discussion is valid for functions having independent variables other than time (e.g., frequency or distance).

### 2.3.1 Time Shifting

Consider a signal  $g(t)$  (Fig. 2.8a) and the same signal delayed by  $T$  seconds (Fig. 2.8b), which we shall denote by  $\phi(t)$ . Whatever happens in  $g(t)$  (Fig. 2.8a) at some instant  $t$  also happens in  $\phi(t)$  (Fig. 2.8b)  $T$  seconds later at the instant  $t + T$ . Therefore,

$$\phi(t + T) = g(t) \quad (2.10)$$

and

$$\phi(t) = g(t - T) \quad (2.11)$$

Therefore, to time-shift a signal by  $T$ , we replace  $t$  with  $t - T$ . Thus,  $g(t - T)$  represents  $g(t)$  time-shifted by  $T$  seconds. If  $T$  is positive, the shift is to the right (delay). If  $T$  is negative, the shift is to the left (advance). Thus,  $g(t - 2)$  is  $g(t)$  delayed (right-shifted) by 2 seconds, and  $g(t + 2)$  is  $g(t)$  advanced (left-shifted) by 2 seconds.

### 2.3.2 Time Scaling

The compression or expansion of a signal in time is known as **time scaling**. Consider the signal  $g(t)$  of Fig. 2.9a. The signal  $\phi(t)$  in Fig. 2.9b is  $g(t)$  compressed in time by a factor of 2. Therefore, whatever happens in  $g(t)$  at some instant  $t$  also happens to  $\phi(t)$  at the instant  $t/2$ , so that



$$\phi\left(\frac{t}{2}\right) = g(t) \tag{2.12}$$

and

$$\phi(t) = g(2t) \tag{2.13}$$

Observe that because  $g(t) = 0$  at  $t = T_1$  and  $T_2$ , the same thing must happen in  $\phi(t)$  at half

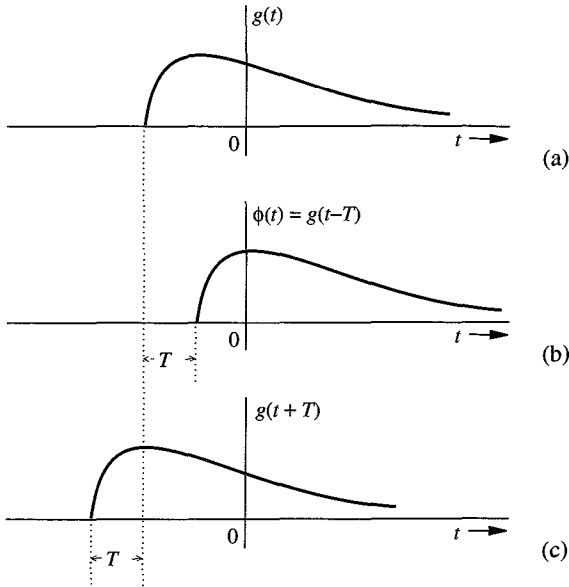


Figure 2.8 Time shifting a signal.

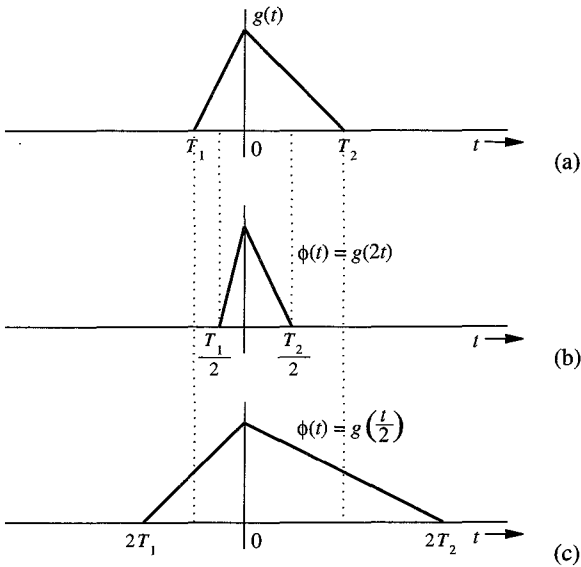


Figure 2.9 Time scaling a signal.

these values. Therefore,  $\phi(t) = 0$  at  $t = T_1/2$  and  $T_2/2$ , as shown in Fig. 2.9b. If  $g(t)$  were recorded on a tape and played back at twice the normal recording speed, we would obtain  $g(2t)$ . In general, if  $g(t)$  is compressed in time by a factor  $a$  ( $a > 1$ ), the resulting signal  $\phi(t)$  is given by

$$\phi(t) = g(at) \quad (2.14)$$

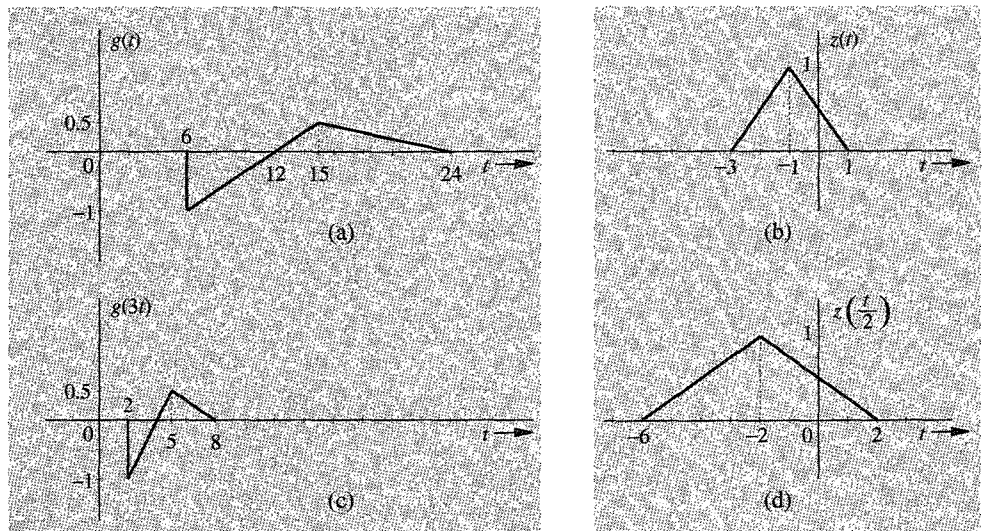
Using a similar argument, we can show that  $g(t)$  expanded (slowed down) in time by a factor  $a$  ( $a > 1$ ) is given by

$$\phi(t) = g\left(\frac{t}{a}\right) \quad (2.15)$$

Figure 2.9c shows  $g(t/2)$ , which is  $g(t)$  expanded in time by a factor of 2. Note that the signal remains anchored at  $t = 0$  during scaling operation (expanding or compressing). In other words, the signal at  $t = 0$  remains unchanged. This is because  $g(t) = g(at) = g(0)$  at  $t = 0$ .

In summary, to time-scale a signal by a factor  $a$ , we replace  $t$  with  $at$ . If  $a > 1$ , the scaling is compression, and if  $a < 1$ , the scaling is expansion.

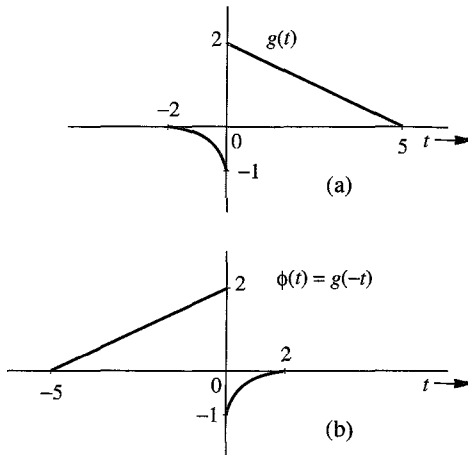
**EXAMPLE 2.3** Figure 2.10a and b shows the signals  $g(t)$  and  $z(t)$ , respectively. Sketch: (a)  $g(3t)$ ; (b)  $z(t/2)$ .



**Figure 2.10** Examples of time compression and time expansion of signals.

(a)  $g(3t)$  is  $g(t)$  compressed by a factor of 3. This means that the values of  $g(t)$  at  $t = 6, 12, 15,$  and  $24$  occur in  $g(3t)$  at the instants  $t = 2, 4, 5,$  and  $8$ , respectively, as shown in Fig. 2.10c.

(b)  $z(t/2)$  is  $z(t)$  expanded (slowed down) by a factor of 2. The values of  $z(t)$  at  $t = 1, -1,$  and  $-3$  occur in  $z(t/2)$  at instants  $2, -2,$  and  $-6$ , respectively, as shown in Fig. 2.10d.



**Figure 2.11** Time inversion (reflection) of a signal.

### 2.3.3 Time Inversion (Time Reversal)

Time inversion may be considered a special case of time scaling with  $a = -1$  in Eq. (2.14). Consider the signal  $g(t)$  in Fig. 2.11a. We can view  $g(t)$  as a rigid wire frame hinged at the vertical axis. To invert  $g(t)$ , we rotate this frame  $180^\circ$  about the vertical axis. This time inversion or folding [the mirror image of  $g(t)$  about the vertical axis] gives us the signal  $\phi(t)$  (Fig. 2.11b). Observe that whatever happens in Fig. 2.11a at some instant  $t$  also happens in Fig. 2.11b at the instant  $-t$ . Therefore,

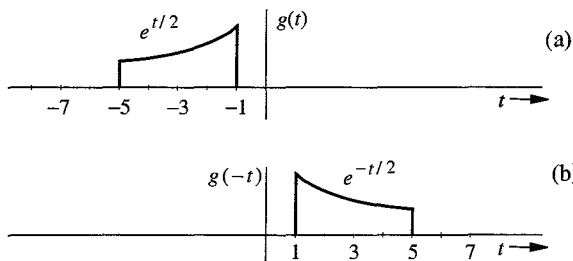
$$\phi(-t) = g(t)$$

and

$$\phi(t) = g(-t) \quad (2.16)$$

Therefore, to time-invert a signal we replace  $t$  with  $-t$ . Thus, the time inversion of signal  $g(t)$  yields  $g(-t)$ . Consequently, the mirror image of  $g(t)$  about the vertical axis is  $g(-t)$ . Recall also that the mirror image of  $g(t)$  about the horizontal axis is  $-g(t)$ .

**EXAMPLE 2.4** For the signal  $g(t)$  shown in Fig. 2.12a, sketch  $g(-t)$ .



**Figure 2.12** Example of time inversion.

The instants  $-1$  and  $-5$  in  $g(t)$  are mapped into instants  $1$  and  $5$  in  $g(-t)$ . If  $g(t) = e^{t/2}$ , then  $g(-t) = e^{-t/2}$ . The signal  $g(-t)$  is shown in Fig. 2.12b.

## 2.4 UNIT IMPULSE FUNCTION

The unit impulse function  $\delta(t)$  is one of the most important functions in the study of signals and systems. This function was first defined by P. A. M. Dirac as

$$\begin{aligned} \delta(t) &= 0 & t &\neq 0 \\ \int_{-\infty}^{\infty} \delta(t) dt &= 1 \end{aligned} \quad (2.17)$$

We can visualize an impulse as a tall, narrow rectangular pulse of unit area, as shown in Fig. 2.13b. The width of this rectangular pulse is some very small value  $\epsilon$ . Its height is a very large value  $1/\epsilon$  in the limit as  $\epsilon \rightarrow 0$ . The unit impulse therefore can be regarded as a rectangular pulse with a width that has become infinitesimally small, a height that has become infinitely large, and an overall area that has been maintained at unity.\* Thus,  $\delta(t) = 0$  everywhere except at  $t = 0$ , where it is undefined. For this reason a unit impulse is represented by the spearlike symbol in Fig. 2.13a.

### Multiplication of a Function by an Impulse

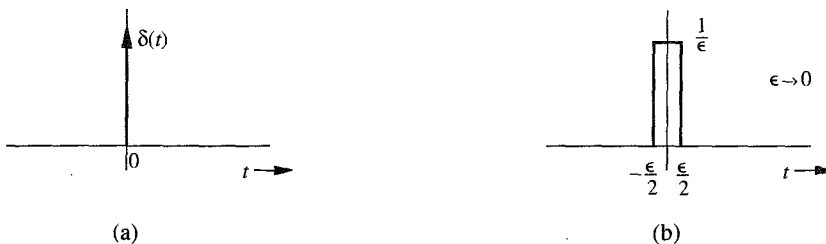
Let us now consider what happens when we multiply the unit impulse  $\delta(t)$  by a function  $\phi(t)$  that is known to be continuous at  $t = 0$ . Since the impulse exists only at  $t = 0$ , and the value of  $\phi(t)$  at  $t = 0$  is  $\phi(0)$ , we obtain

$$\phi(t)\delta(t) = \phi(0)\delta(t) \quad (2.18a)$$

Similarly, if  $\phi(t)$  is multiplied by an impulse  $\delta(t - T)$  (an impulse located at  $t = T$ ), then

$$\phi(t)\delta(t - T) = \phi(T)\delta(t - T) \quad (2.18b)$$

provided  $\phi(t)$  is continuous at  $t = T$ .



**Figure 2.13** Unit impulse and its approximation.

\* The impulse function can also be approximated by other pulses, such as an exponential pulse, a triangular pulse, or a Gaussian pulse.

## Sampling Property of the Unit Impulse Function

From Eq. (2.18a) it follows that

$$\begin{aligned}\int_{-\infty}^{\infty} \phi(t)\delta(t) dt &= \phi(0) \int_{-\infty}^{\infty} \delta(t) dt \\ &= \phi(0)\end{aligned}\quad (2.19a)$$

provided  $\phi(t)$  is continuous at  $t = 0$ . This result means that *the area under the product of a function with an impulse  $\delta(t)$  is equal to the value of that function at the instant where the unit impulse is located*. This property is very important and useful, and is known as the **sampling**, or **sifting**, **property** of the unit impulse.

From Eq. (2.18b) it follows that

$$\int_{-\infty}^{\infty} \phi(t)\delta(t - T) dt = \phi(T) \quad (2.19b)$$

Equation (2.19b) is just another form of sampling or sifting property. In the case of Eq. (2.19b), the impulse  $\delta(t - T)$  is located at  $t = T$ . Therefore, the area under  $\phi(t)\delta(t - T)$  is  $\phi(T)$ , the value of  $\phi(t)$  at the instant where the impulse is located (at  $t = T$ ). In these derivations we have assumed that the function is continuous at the instant where the impulse is located.

## Unit Impulse as a Generalized Function

The definition of the unit impulse function [Eq. (2.17)] leads to a nonunique function.<sup>1</sup> Moreover,  $\delta(t)$  is not even a true function in the ordinary sense. An ordinary function is specified by its values for all time  $t$ . The impulse function is zero everywhere except at  $t = 0$ , and at this only interesting part of its range it is undefined. In a more rigorous approach, the impulse function is defined not as an ordinary function but as a **generalized function**, where  $\delta(t)$  is defined by Eqs. (2.19). We say nothing about what the impulse function is or what it looks like. Instead, it is defined in terms of the effect it has on a test function  $\phi(t)$ . We define a unit impulse as a function for which the area under its product with a function  $\phi(t)$  is equal to the value of the function  $\phi(t)$  at the instant where the impulse is located. Recall that the sampling property [Eqs. (2.19)] is the consequence of the classical (Dirac) definition of impulse in Eq. (2.17). In contrast, *the sampling property [Eqs. (2.19)] defines the impulse function in the generalized function approach*.

## Unit Step Function $u(t)$

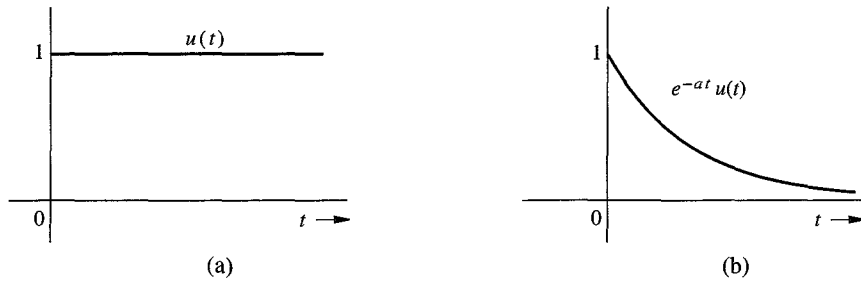
Another familiar and useful function is the **unit step function**  $u(t)$ , defined by (Fig. 2.14a)

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

If we want a signal to start at  $t = 0$  (so that it has a value of zero for  $t < 0$ ), we only need to multiply the signal with  $u(t)$ . A signal that does not start before  $t = 0$  is called a **causal signal**. In other words,  $g(t)$  is a causal signal if

$$g(t) = 0 \quad t < 0$$

The signal  $e^{-at}$  represents an exponential that starts at  $t = -\infty$ . If we want this signal to start at  $t = 0$  (the causal form), it can be described as  $e^{-at}u(t)$  (Fig. 2.14b). From Fig. 2.13b, we



**Figure 2.14** (a) Unit step function  $u(t)$ . (b) Causal exponential  $e^{-at}u(t)$ .

observe that the area from  $-\infty$  to  $t$  under the limiting form of  $\delta(t)$  is zero if  $t < 0$  and unity if  $t \geq 0$ . Consequently,

$$\int_{-\infty}^t \delta(\tau) d\tau = \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases} = u(t) \quad (2.20a)$$

From this result it follows that

$$\frac{du}{dt} = \delta(t) \quad (2.20b)$$

## 2.5 SIGNALS AND VECTORS

There is a perfect analogy between signals and vectors. The analogy is so strong that the term “analogy” understates the reality. Signals are not just *like* vectors. Signals *are* vectors. A vector can be represented as a sum of its components in a variety of ways, depending on the choice of coordinate system. A signal can also be represented as a sum of its components in a variety of ways. Let us begin with some basic vector concepts and then apply those concepts to signals.

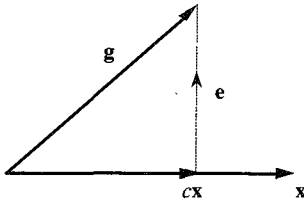
### 2.5.1 Component of a Vector

A vector is specified by its magnitude and its direction. We shall denote all vectors by boldface type. For example,  $\mathbf{x}$  is a certain vector with magnitude or length  $|\mathbf{x}|$ . Consider two vectors  $\mathbf{g}$  and  $\mathbf{x}$ , as shown in Fig. 2.15. Let the component of  $\mathbf{g}$  along  $\mathbf{x}$  be  $c\mathbf{x}$ . Geometrically the component of  $\mathbf{g}$  along  $\mathbf{x}$  is the projection of  $\mathbf{g}$  on  $\mathbf{x}$ , and is obtained by drawing a perpendicular from the tip of  $\mathbf{g}$  on the vector  $\mathbf{x}$ , as shown in Fig. 2.15. What is the mathematical significance of a component of a vector along another vector? As seen from Fig. 2.15, the vector  $\mathbf{g}$  can be expressed in terms of vector  $\mathbf{x}$  as

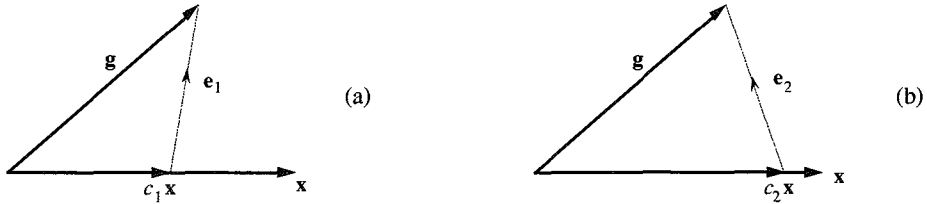
$$\mathbf{g} = c\mathbf{x} + \mathbf{e} \quad (2.21)$$

However, this is not the only way to express  $\mathbf{g}$  in terms of  $\mathbf{x}$ . Figure 2.16 shows two of the infinite other possibilities. From Fig. 2.16a and b, we have

$$\mathbf{g} = c_1\mathbf{x} + \mathbf{e}_1 = c_2\mathbf{x} + \mathbf{e}_2 \quad (2.22)$$



**Figure 2.15** Component (projection) of a vector along another vector.



**Figure 2.16** Approximation of a vector in terms of another vector.

In each of these three representations (Figs. 2.15 and 2.16)  $\mathbf{g}$  is represented in terms of  $\mathbf{x}$  plus another vector, called the **error vector**. If we approximate  $\mathbf{g}$  by  $c\mathbf{x}$  (Fig. 2.15),

$$\mathbf{g} \simeq c\mathbf{x} \quad (2.23)$$

the error in this approximation is the vector  $\mathbf{e} = \mathbf{g} - c\mathbf{x}$ . Similarly, the errors in the approximations in Fig. 2.16a and b are  $\mathbf{e}_1$  and  $\mathbf{e}_2$ . What is unique about the approximation in Fig. 2.15 is that the error vector is the smallest. We can now define mathematically the component of a vector  $\mathbf{g}$  along vector  $\mathbf{x}$  to be  $c\mathbf{x}$ , where  $c$  is chosen to minimize the length of the error vector  $\mathbf{e} = \mathbf{g} - c\mathbf{x}$ . For convenience we define the dot (inner or scalar) product of two vectors  $\mathbf{g}$  and  $\mathbf{x}$  as

$$\mathbf{g} \cdot \mathbf{x} = |\mathbf{g}||\mathbf{x}| \cos \theta \quad (2.24)$$

where  $\theta$  is the angle between vectors  $\mathbf{g}$  and  $\mathbf{x}$ . Using this definition, we can express  $|\mathbf{x}|$ , the length of a vector  $\mathbf{x}$ , as

$$|\mathbf{x}|^2 = \mathbf{x} \cdot \mathbf{x} \quad (2.25)$$

Now, the length of the component of  $\mathbf{g}$  along  $\mathbf{x}$  is  $|\mathbf{g}| \cos \theta$ , but it is also  $c|\mathbf{x}|$ . Therefore,

$$c|\mathbf{x}| = |\mathbf{g}| \cos \theta$$

Multiplying both sides by  $|\mathbf{x}|$  yields

$$c|\mathbf{x}|^2 = |\mathbf{g}||\mathbf{x}| \cos \theta = \mathbf{g} \cdot \mathbf{x}$$

and

$$c = \frac{\mathbf{g} \cdot \mathbf{x}}{\mathbf{x} \cdot \mathbf{x}} = \frac{1}{|\mathbf{x}|^2} \mathbf{g} \cdot \mathbf{x} \quad (2.26)$$

From Fig. 2.15, it is apparent that when  $\mathbf{g}$  and  $\mathbf{x}$  are perpendicular, or orthogonal, then  $\mathbf{g}$  has a zero component along  $\mathbf{x}$ ; consequently,  $c = 0$ . Keeping an eye on Eq. (2.26), we therefore define  $\mathbf{g}$  and  $\mathbf{x}$  to be **orthogonal** if the inner (scalar or dot) product of the two vectors is zero, that is, if

$$\mathbf{g} \cdot \mathbf{x} = 0 \quad (2.27)$$

## 2.5.2 Component of a Signal

The concepts of vector component and orthogonality can be extended to signals. Consider the problem of approximating a real signal  $g(t)$  in terms of another real signal  $x(t)$  over an interval  $[t_1, t_2]$ :

$$g(t) \simeq cx(t) \quad t_1 \leq t \leq t_2 \quad (2.28)$$

The error  $e(t)$  in this approximation is

$$e(t) = \begin{cases} g(t) - cx(t) & t_1 \leq t \leq t_2 \\ 0 & \text{otherwise} \end{cases} \quad (2.29)$$

We now select some criterion for the “best approximation.” We know that the signal energy is one possible measure of a signal size. For best approximation, we need to minimize the error signal, that is, minimize its size, which is its energy  $E_e$  over the interval  $[t_1, t_2]$ , given by

$$\begin{aligned} E_e &= \int_{t_1}^{t_2} e^2(t) dt \\ &= \int_{t_1}^{t_2} [g(t) - cx(t)]^2 dt \end{aligned}$$

Note that the right-hand side is a definite integral with  $t$  as the dummy variable. Hence,  $E_e$  is a function of the parameter  $c$  (not  $t$ ) and  $E_e$  is minimum for some choice of  $c$ . To minimize  $E_e$ , a necessary condition is

$$\frac{dE_e}{dc} = 0 \quad (2.30)$$

or

$$\frac{d}{dc} \left[ \int_{t_1}^{t_2} [g(t) - cx(t)]^2 dt \right] = 0$$

Expanding the squared term inside the integral, we obtain

$$\frac{d}{dc} \left[ \int_{t_1}^{t_2} g^2(t) dt \right] - \frac{d}{dc} \left[ 2c \int_{t_1}^{t_2} g(t)x(t) dt \right] + \frac{d}{dc} \left[ c^2 \int_{t_1}^{t_2} x^2(t) dt \right] = 0$$

from which we obtain

$$-2 \int_{t_1}^{t_2} g(t)x(t) dt + 2c \int_{t_1}^{t_2} x^2(t) dt = 0$$



and

$$c = \frac{\int_{t_1}^{t_2} g(t)x(t) dt}{\int_{t_1}^{t_2} x^2(t) dt} = \frac{1}{E_x} \int_{t_1}^{t_2} g(t)x(t) dt \quad (2.31)$$

We observe a remarkable similarity between the behavior of vectors and signals, as indicated by Eqs. (2.26) and (2.31). It is evident from these two parallel expressions that *the area under the product of two signals corresponds to the inner (scalar or dot) product of two vectors*. In fact, the area under the product of  $g(t)$  and  $x(t)$  is called the **inner product** of  $g(t)$  and  $x(t)$ , and is denoted by  $(f, g)$ . The energy of a signal is the inner product of a signal with itself, and corresponds to the vector length squared (which is the inner product of the vector with itself).

To summarize our discussion, if a signal  $g(t)$  is approximated by another signal  $x(t)$  as

$$g(t) \simeq cx(t)$$

then the optimum value of  $c$  that minimizes the energy of the error signal in this approximation is given by Eq. (2.31).

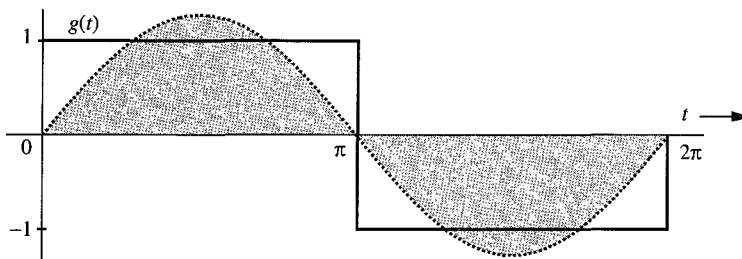
Taking our clue from vectors, we say that a signal  $g(t)$  contains a component  $cx(t)$ , where  $c$  is given by Eq. (2.31). Note that in vector terminology,  $cx(t)$  is the projection of  $g(t)$  on  $x(t)$ . Continuing with the analogy, we say that if the component of a signal  $g(t)$  of the form  $x(t)$  is zero (that is,  $c = 0$ ), the signals  $g(t)$  and  $x(t)$  are orthogonal over the interval  $[t_1, t_2]$ . Therefore, we define the real signals  $g(t)$  and  $x(t)$  to be orthogonal over the interval  $[t_1, t_2]$  if\*

$$\int_{t_1}^{t_2} g(t)x(t) dt = 0 \quad (2.32)$$

**EXAMPLE 2.5** For the square signal  $g(t)$  shown in Fig. 2.17 find the component in  $g(t)$  of the form  $\sin t$ . In other words, approximate  $g(t)$  in terms of  $\sin t$ :

$$g(t) \simeq c \sin t \quad 0 \leq t \leq 2\pi$$

so that the energy of the error signal is minimum.



**Figure 2.17** Approximation of a square signal in terms of a single sinusoid.

\* For complex signals the definition is modified as in Eq. (2.40).

In this case,

$$x(t) = \sin t \quad \text{and} \quad E_x = \int_0^{2\pi} \sin^2 t \, dt = \pi$$

From Eq. (2.31), we find

$$c = \frac{1}{\pi} \int_0^{2\pi} g(t) \sin t \, dt = \frac{1}{\pi} \left[ \int_0^{\pi} \sin t \, dt + \int_{\pi}^{2\pi} -\sin t \, dt \right] = \frac{4}{\pi} \quad (2.33)$$

Therefore,

$$g(t) \simeq \frac{4}{\pi} \sin t \quad (2.34)$$

represents the best approximation of  $g(t)$  by the function  $\sin t$ , which will minimize the error energy. This sinusoidal component of  $g(t)$  is shown shaded in Fig. 2.17. By analogy with vectors, we say that the square function  $g(t)$  shown in Fig. 2.17 has a component of signal  $\sin t$  and that the magnitude of this component is  $4/\pi$ .

### 2.5.3 Orthogonality in Complex Signals

So far we have restricted ourselves to real functions of  $t$ . To generalize the results to complex functions of  $t$ , consider again the problem of approximating a function  $g(t)$  by a function  $x(t)$  over an interval ( $t_1 \leq t \leq t_2$ ):

$$g(t) \simeq cx(t) \quad (2.35)$$

where  $g(t)$  and  $x(t)$  now can be complex functions of  $t$ . Recall that the energy  $E_x$  of the complex signal  $x(t)$  over an interval [ $t_1, t_2$ ] is

$$E_x = \int_{t_1}^{t_2} |x(t)|^2 \, dt$$

In this case, both the coefficient  $c$  and the error

$$e(t) = g(t) - cx(t) \quad (2.36)$$

are complex (in general). For the best approximation, we choose  $c$  such that we minimize  $E_e$ , the energy of the error signal  $e(t)$ , given by

$$E_e = \int_{t_1}^{t_2} |g(t) - cx(t)|^2 \, dt \quad (2.37)$$

Recall also that

$$|u + v|^2 = (u + v)(u^* + v^*) = |u|^2 + |v|^2 + u^*v + uv^* \quad (2.38)$$

Using this result, we can, after some manipulation, express the integral  $E_e$  in Eq. (2.37) as

$$E_e = \int_{t_1}^{t_2} |g(t)|^2 \, dt - \left| \frac{1}{\sqrt{E_x}} \int_{t_1}^{t_2} g(t)x^*(t) \, dt \right|^2 + \left| c\sqrt{E_x} - \frac{1}{\sqrt{E_x}} \int_{t_1}^{t_2} g(t)x^*(t) \, dt \right|^2$$

Since the first two terms on the right-hand side are independent of  $c$ , it is clear that  $E_e$  is minimized by choosing  $c$  such that the third term is zero. This yields

$$c = \frac{1}{E_x} \int_{t_1}^{t_2} g(t)x^*(t) dt \quad (2.39)$$

In light of this result, we need to redefine orthogonality for the complex case as follows: Two complex functions  $x_1(t)$  and  $x_2(t)$  are orthogonal over an interval ( $t \leq t_1 \leq t_2$ ) if

$$\int_{t_1}^{t_2} x_1(t)x_2^*(t) dt = 0 \quad \text{or} \quad \int_{t_1}^{t_2} x_1^*(t)x_2(t) dt = 0 \quad (2.40)$$

Either equality suffices. This is a general definition of orthogonality, which reduces to Eq. (2.32) when the functions are real.

## Energy of the Sum of Orthogonal Signals

We know that the length of the sum of two orthogonal vectors is equal to the sum of the lengths squared of the two vectors. Thus, if vectors  $\mathbf{x}$  and  $\mathbf{y}$  are orthogonal, and if  $\mathbf{z} = \mathbf{x} + \mathbf{y}$ , then

$$|\mathbf{z}|^2 = |\mathbf{x}|^2 + |\mathbf{y}|^2$$

We have a similar result for signals. The energy of the sum of two orthogonal signals is equal to the sum of the energies of the two signals. Thus, if signals  $x(t)$  and  $y(t)$  are orthogonal over an interval  $[t_1, t_2]$ , and if  $z(t) = x(t) + y(t)$ , then

$$E_z = E_x + E_y \quad (2.41)$$

We now prove this result for complex signals of which real signals are a special case. From Eq. (2.38) it follows that

$$\begin{aligned} \int_{t_1}^{t_2} |x(t) + y(t)|^2 dt &= \int_{t_1}^{t_2} |x(t)|^2 dt + \int_{t_1}^{t_2} |y(t)|^2 dt + \int_{t_1}^{t_2} x(t)y^*(t) dt + \int_{t_1}^{t_2} x^*(t)y(t) dt \\ &= \int_{t_1}^{t_2} |x(t)|^2 dt + \int_{t_1}^{t_2} |y(t)|^2 dt \end{aligned} \quad (2.42)$$

The last result follows from the fact that because of orthogonality, the two integrals of the cross products  $x(t)y^*(t)$  and  $x^*(t)y(t)$  are zero. This result can be extended to the sum of any number of mutually orthogonal signals.

## 2.6 SIGNAL COMPARISON: CORRELATION

Section 2.5 has prepared the foundation for signal comparison. Here again, we can benefit by considering the concept of vector comparison. Two vectors  $\mathbf{g}$  and  $\mathbf{x}$  are similar if  $\mathbf{g}$  has a large component along  $\mathbf{x}$ . In other words, if  $c$  in Eq. (2.26) is large, the vectors  $\mathbf{g}$  and  $\mathbf{x}$  are similar. We could consider  $c$  as a quantitative measure of similarity between  $\mathbf{g}$  and  $\mathbf{x}$ . Such a measure, however, would be defective. The amount of similarity between  $\mathbf{g}$  and  $\mathbf{x}$  should be independent of the lengths of  $\mathbf{g}$  and  $\mathbf{x}$ . If we double the length of  $\mathbf{g}$ , for example, the amount of

similarity between  $\mathbf{g}$  and  $\mathbf{x}$  should not change. From Eq. (2.26), however, we see that doubling  $\mathbf{g}$  doubles the value of  $c$  (whereas doubling  $\mathbf{x}$  halves the value of  $c$ ). Our measure is clearly faulty. Similarity between two vectors is indicated by the angle  $\theta$  between the vectors. The smaller the  $\theta$ , the larger is the similarity, and vice versa. The amount of similarity can therefore be conveniently measured by  $\cos \theta$ . The larger the  $\cos \theta$ , larger is the similarity between the two vectors. Thus, a suitable measure would be  $c_n = \cos \theta$ , which is given by

$$c_n = \cos \theta = \frac{\mathbf{g} \cdot \mathbf{x}}{|\mathbf{g}| |\mathbf{x}|} \quad (2.43)$$

We can readily verify that this measure is independent of the lengths of  $\mathbf{g}$  and  $\mathbf{x}$ . This similarity measure  $c_n$  is known as the **correlation coefficient**. Observe that

$$-1 \leq c_n \leq 1 \quad (2.44)$$

Thus, the magnitude of  $c_n$  is never greater than unity. If the two vectors are aligned, the similarity is maximum ( $c_n = 1$ ). Two vectors aligned in opposite directions have maximum dissimilarity ( $c_n = -1$ ). If the two vectors are orthogonal, the similarity is zero.

We use the same argument in defining a similarity index (the correlation coefficient) for signals. We shall consider the signals over the entire time interval from  $-\infty$  to  $\infty$ . To make  $c$  in Eq. (2.31) independent of the energies (sizes) of  $g(t)$  and  $x(t)$ , we must normalize  $c$  by normalizing the two signals to have unit energies. Thus, the appropriate similarity index  $c_n$  analogous to Eq. (2.43) is given by

$$c_n = \frac{1}{\sqrt{E_g E_x}} \int_{-\infty}^{\infty} g(t)x(t) dt \quad (2.45)$$

Observe that multiplying either  $g(t)$  or  $x(t)$  by any constant has no effect on this index. It is independent of the sizes (energies) of  $g(t)$  and  $x(t)$ . Using Schwarz's inequality (proved in Appendix B),\* we can show that the magnitude of  $c_n$  is never greater than 1:

$$-1 \leq c_n \leq 1 \quad (2.46)$$

### Best Friends, Worst Enemies, and Complete Strangers

We can readily verify that if  $g(t) = Kx(t)$ , then  $c_n = 1$  when  $K$  is any positive constant, and  $c_n = -1$  when  $K$  is any negative constant. Also  $c_n = 0$  if  $g(t)$  and  $x(t)$  are orthogonal. Thus, the maximum similarity [when  $g(t) = Kx(t)$ ] is indicated by  $c_n = 1$ , the maximum dissimilarity [when  $g(t) = -Kx(t)$ ] is indicated by  $c_n = -1$ . When the two signals are orthogonal, the similarity is zero. Qualitatively speaking, we may view orthogonal signals as unrelated signals. Note that maximum dissimilarity is different from unrelatedness qualitatively. For example, we have the best friends ( $c_n = 1$ ), the worst enemies ( $c_n = -1$ ), and complete strangers, who do not care whether we exist or not ( $c_n = 0$ ). The worst enemies are not strangers but, in many ways, people who think like us, only in opposite ways.

\* The Schwarz inequality states that for two real energy signals  $g(t)$  and  $x(t)$ ,

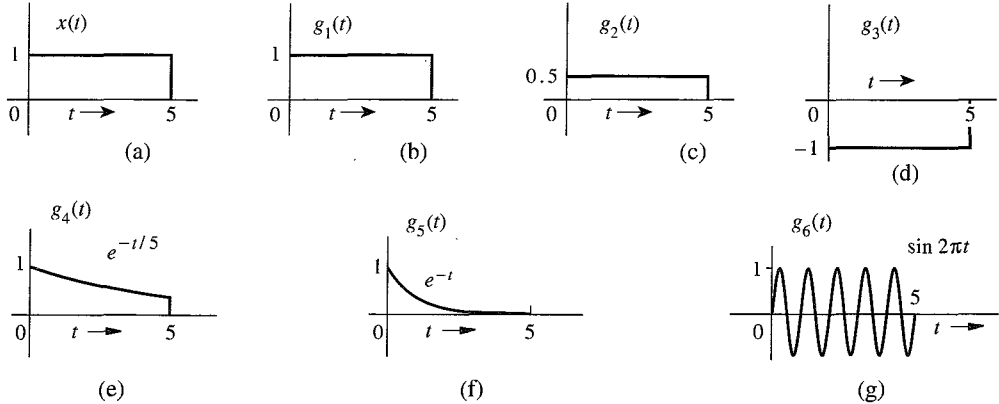
$$\left[ \int_{-\infty}^{\infty} g(t)x(t) dt \right]^2 \leq E_g E_x \quad (2.45n)$$

with equality if and only if  $x(t) = Kg(t)$ , where  $K$  is an arbitrary constant. There is also a similar inequality for complex signals.

We can readily extend this discussion to complex signal comparison. We generalize the definition of  $c_n$  to include complex signals as

$$c_n = \frac{1}{\sqrt{E_g E_x}} \int_{-\infty}^{\infty} g(t)x^*(t) dt \quad (2.47)$$

**EXAMPLE 2.6** Find the correlation coefficient  $c_n$  between the pulse  $x(t)$  and the pulses  $g_i(t)$ ,  $i = 1, 2, 3, 4, 5$ , and 6, shown in Fig. 2.18.



**Figure 2.18** Signals for Example 2.6.

We shall compute  $c_n$  using Eq. (2.45) for each of the six cases. Let us first compute the energies of all the signals,

$$E_x = \int_0^5 x^2(t) dt = \int_0^5 dt = 5$$

In the same way we find  $E_{g_1} = 5$ ,  $E_{g_2} = 1.25$ , and  $E_{g_3} = 5$ . Also to determine  $E_{g_4}$  and  $E_{g_5}$ , we determine the energy  $E$  of  $e^{-at}u(t)$  over the interval  $t = 0$  to  $T$ :

$$E = \int_0^T (e^{-at})^2 dt = \int_0^T e^{-2at} dt = \frac{1}{2a}(1 - e^{-2aT})$$

For  $g_4(t)$ ,  $a = 1/5$  and  $T = 5$ . Therefore,  $E_{g_4} = 2.1617$ . For  $g_5(t)$ ,  $a = 1$  and  $T = \infty$ . Therefore,  $E_{g_5} = 0.5$ . The energy of  $E_{g_6}$  is given by

$$E_{g_6} = \int_0^5 \sin^2 2\pi t dt = 2.5$$

Using Eq. (2.25), the correlation coefficients for the six cases are found as

$$\begin{aligned} (1) \quad & \frac{1}{\sqrt{5 \cdot 5}} \int_0^5 dt = 1 & (2) \quad & \frac{1}{\sqrt{1.25 \cdot 5}} \int_0^5 (0.5) dt = 1 \\ (3) \quad & \frac{1}{\sqrt{5 \cdot 5}} \int_0^5 (-1) dt = -1 & (4) \quad & \frac{1}{\sqrt{2.1617 \cdot 5}} \int_0^5 e^{-t/5} dt = 0.961 \\ (5) \quad & \frac{1}{\sqrt{0.5 \cdot 5}} \int_0^5 e^{-t} dt = 0.628 & (6) \quad & \frac{1}{\sqrt{2.5 \cdot 5}} \int_0^5 \sin 2\pi t dt = 0 \end{aligned}$$

*Comments:* Because  $g_1(t) = x(t)$ , the two signals have the maximum possible similarity, and  $c_n = 1$ . However, the signal  $g_2(t)$  also shows maximum possible similarity with  $c_n = 1$ . This is because we have defined  $c_n$  to measure the similarity of the wave shapes, and it is independent of the amplitude (strength) of the signals compared. The signal  $g_2(t)$  is identical to  $x(t)$  in shape; only the amplitude (strength) is different. Hence,  $c_n = 1$ . The signal  $g_3(t)$ , on the other hand, has the maximum possible dissimilarity with  $x(t)$  because it is equal to  $-x(t)$ . For  $g_4(t)$ ,  $c_n = 0.961$ , implying a high degree of similarity with  $x(t)$ . This is reasonable because  $g_4(t)$  is very similar to  $x(t)$  over the duration of  $x(t)$  (for  $0 \leq t \leq 5$ ). Just by inspection, we notice that the variations or changes in both  $x(t)$  and  $g_4(t)$  are at similar rates. Such is not the case with  $g_5(t)$ , where we notice that variations in  $g_5(t)$  are generally at a higher rate than those in  $x(t)$ . There is still considerable similarity: Both signals always remain positive and show no oscillations. Both signals have zero or negligible strength beyond  $t = 5$ . Thus,  $g_5(t)$  is similar to  $x(t)$ , but not as similar as  $g_4(t)$ . This is why  $c_n = 0.628$  for  $g_5(t)$ . The signal  $g_6(t)$  is orthogonal to  $x(t)$ , so that  $c_n = 0$ . This appears to indicate that the dissimilarity in this case is not as strong as that of  $g_3(t)$ , for which  $c_n = -1$ . This may seem odd because  $g_3(t)$  appears more similar to  $x(t)$  than does  $g_6(t)$ . The dissimilarity between  $x(t)$  and  $g_3(t)$  is of the nature of antipathy (the worst enemy); in a way they are very similar, but in opposite ways. On the other hand, the dissimilarity of  $x(t)$  with  $g_6(t)$  stems from the fact that they are almost of different species or from different planets; it is of the nature of being strangers to each other. Hence, the dissimilarity of  $x(t)$  with  $g_6(t)$  rates lower than that with  $g_3(t)$ .

## 2.6.1 Application to Signal Detection

Correlation between two signals is an extremely important concept, which measures the degree of similarity (agreement or alignment) between the two signals. This concept is widely used for signal processing applications in radar, sonar, digital communication, electronic warfare, and many others.

We explain this concept by an example of radar where a signal pulse is transmitted in order to detect a suspected target. If a target is present, the pulse will be reflected by it. If a target is not present, there will be no reflected pulse, just a noise. By detecting the presence or absence of the reflected pulse we confirm the presence or absence of a target. The crucial problem in this procedure is to detect the heavily attenuated, reflected pulse (of known waveform) buried in the unwanted noise signal. Correlation of the received pulse with the transmitted pulse can be of great help in this situation. A similar situation exists in digital communication, where we are required to detect the presence of one of the two known waveforms in the presence of noise.

We now explain qualitatively how signal detection using the correlation technique is accomplished. Consider the case of binary communication, where two known waveforms are received in a random sequence. Each time we receive a pulse, our task is to determine which of the two (known) waveforms is received. To make the detection easier, we must make the two pulses as dissimilar as possible, which means that we should select one pulse as the negative of the other pulse. This gives the highest dissimilarity ( $c_n = -1$ ). This scheme is sometimes called the **antipodal** scheme. We can also use orthogonal pulses, which result in  $c_n = 0$ . In practice

both these options are used, although the antipodal one is best in terms of distinguishability between the two pulses.

Let us consider the antipodal scheme in which the two pulses are  $p(t)$  and  $-p(t)$ . The correlation coefficient  $c_n$  of these pulses is  $-1$ . Assume that there is no noise nor any other imperfections in the transmission. The receiver consists of a correlator that computes the correlation between  $p(t)$  and the received pulse. If the correlation is 1, we decide that  $p(t)$  is received; if the correlation is  $-1$ , we decide that  $-p(t)$  is received. Because of the maximum possible dissimilarity between the two pulses, detection is easier. In practice, however, there are several imperfections. There is always an unwanted signal (noise) superimposed on the received pulses. Moreover, during transmission, pulses get distorted and dispersed (spread out) in time. Consequently, the correlation coefficient is no longer  $\pm 1$ , but has a smaller magnitude, thus reducing the distinguishability. We use a **threshold detector**, which decides that if the correlation is positive, the received pulse is  $p(t)$ , and if the correlation is negative, the received pulse is  $-p(t)$ .

Suppose, for example, that  $p(t)$  has been transmitted. In the ideal case correlation of this pulse at the receiver would be 1, the maximum possible. Now because of the noise and pulse distortion, the correlation is less than 1. In some extreme situation, channel noise, pulse distortion and overlapping (spreading) from other pulses can make this pulse so dissimilar to  $p(t)$  that the correlation can be negative. In this case, the threshold detector decides that  $-p(t)$  has been received, thus causing a detection error. In the same way, if  $-p(t)$  is transmitted, the detector could decide that  $p(t)$  is transmitted. Our task is to make sure that the transmitted pulses have sufficient energy for the damage caused by noise and other imperfections to remain within a limit so that the error probability is below some acceptable bounds. In the ideal case, the margin provided by the correlation  $c_n$  for distinguishing the two pulses is 2 (from 1 to  $-1$  and vice versa). The noise and channel distortion reduce this margin. That is why it is important to start with as large a margin as possible. This explains why the antipodal scheme has the best performance in terms of guarding against channel noise and pulse distortion. However, as mentioned earlier, because of some other reasons, schemes such as an orthogonal scheme, where  $c_n = 0$ , are also used, even when they provide a smaller margin (from 0 to 1 and vice versa) in distinguishing the pulses. Quantitative discussion of correlation in digital signal detection is discussed in chapters 13 and 14.

In later chapters we shall discuss pulse dispersion and pulse distortion during transmission, as well as the calculation of error probability in the presence of noise.

## 2.6.2 Correlation Functions

Consider the application of correlation to signal detection in a radar, where a signal pulse is transmitted in order to detect a suspected target. If a target is present, the pulse will be reflected by it. If no target is present, there will be no reflected pulse, just a noise. By detecting the presence or absence of the reflected pulse we confirm the presence or absence of a target. By measuring the time delay between the transmitted and the received (reflected) pulses we determine the distance of the target. Let the transmitted and the reflected pulses be denoted by  $g(t)$  and  $z(t)$ , respectively, as shown in Fig. 2.19. If we were to use Eq. (2.45) directly to measure the correlation coefficient  $c_n$ , we would obtain

$$c_n = \frac{1}{\sqrt{E_g E_z}} \int_{-\infty}^{\infty} g(t)z(t) dt = 0 \quad (2.48)$$

Thus, the correlation is zero because the pulses are disjoint (nonoverlapping in time). The integral (2.48) will yield zero value even when the pulses are identical but with relative time shift. To avoid this difficulty, we compare the transmitted pulse  $g(t)$  with the received pulse  $z(t)$  shifted by  $\tau$ . If for some value of  $\tau$ , there is a strong correlation, we not only detect the presence of the pulse but we also detect the relative time shift of  $z(t)$  with respect to  $g(t)$ . For this reason, instead of using the integral on the right hand, we use the modified integral  $\psi_{gz}(\tau)$ , the **cross-correlation** function of two real signals  $g(t)$  and  $z(t)$ , defined by\*

$$\psi_{gz}(\tau) \equiv \int_{-\infty}^{\infty} g(t)z(t + \tau) dt \quad (2.49)$$

Here  $z(t + \tau)$  is the pulse  $z(t)$  left-shifted (advanced) by  $\tau$  seconds. Therefore,  $\psi_{gz}(\tau)$  is an indication of similarity (correlation) of  $g(t)$  with  $z(t)$  advanced (left-shifted) by  $\tau$  seconds.

### Autocorrelation Function

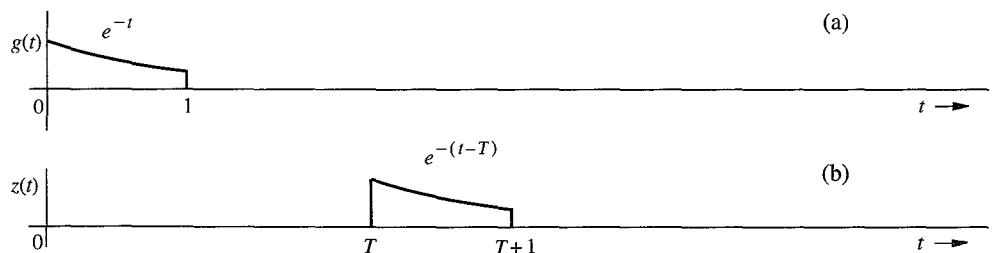
The correlation of a signal with itself is called **autocorrelation**. The autocorrelation function  $\psi_g(\tau)$  of a real signal  $g(t)$  is defined as

$$\psi_g(\tau) \equiv \int_{-\infty}^{\infty} g(t)g(t + \tau) dt \quad (2.50)$$

In Chapter 3, we shall show that the autocorrelation function provides a valuable spectral information about the signal.

## 2.7 SIGNAL REPRESENTATION BY ORTHOGONAL SIGNAL SET

In this section we show a way of representing a signal as a sum of orthogonal signals. Here again we can benefit from the insight gained from a similar problem with vectors. We know that a vector can be represented as the sum of orthogonal vectors, which form the



**Figure 2.19** Physical explanation of the autocorrelation function.

\* For complex signals we define

$$\psi_{gz}(\tau) \equiv \int_{-\infty}^{\infty} g^*(t)z(t + \tau) dt$$



coordinate system of a vector space. The problem with signals is analogous, and the results for signals are parallel to those for vectors. For this reason let us review the case of vector representation.

### 2.7.1 Orthogonal Vector Space

Consider a three-dimensional Cartesian vector space described by three mutually orthogonal vectors  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ , as shown in Fig. 2.20. First, we shall seek to approximate a three-dimensional vector  $\mathbf{g}$  in terms of two mutually orthogonal vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ :

$$\mathbf{g} \simeq c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2$$

The error  $\mathbf{e}$  in this approximation is

$$\mathbf{e} = \mathbf{g} - (c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2)$$

or

$$\mathbf{g} = c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + \mathbf{e}$$

As in the earlier geometrical argument, we see from Fig. 2.20 that the length of  $\mathbf{e}$  is minimum when  $\mathbf{e}$  is perpendicular to the  $\mathbf{x}_1$ - $\mathbf{x}_2$  plane, and  $c_1 \mathbf{x}_1$  and  $c_2 \mathbf{x}_2$  are the projections (components) of  $\mathbf{g}$  on  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , respectively. Therefore, the constants  $c_1$  and  $c_2$  are given by Eq. (2.26).

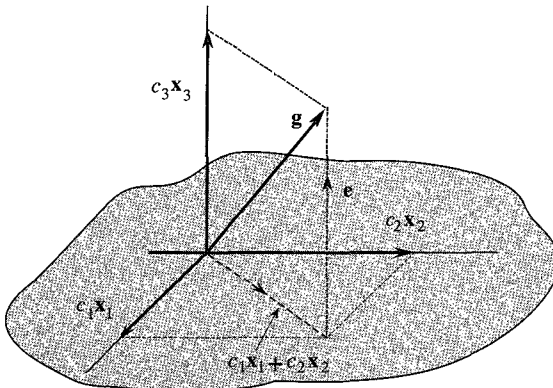
Now let us determine the best approximation to  $\mathbf{g}$  in terms of all three mutually orthogonal vectors  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ :

$$\mathbf{g} \simeq c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + c_3 \mathbf{x}_3 \quad (2.51)$$

Figure 2.20 shows that a unique choice of  $c_1$ ,  $c_2$ , and  $c_3$  exists, for which Eq. (2.51) is no longer an approximation but an equality:

$$\mathbf{g} = c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 + c_3 \mathbf{x}_3$$

In this case,  $c_1 \mathbf{x}_1$ ,  $c_2 \mathbf{x}_2$ , and  $c_3 \mathbf{x}_3$  are the projections (components) of  $\mathbf{g}$  on  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ , respectively. Note that the error in the approximation is zero when  $\mathbf{g}$  is approximated in terms of three mutually orthogonal vectors:  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ . This is because  $\mathbf{g}$  is a three-dimensional



**Figure 2.20** Representation of a vector in three-dimensional space.

vector, and the vectors  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$  represent a *complete set* of orthogonal vectors in three-dimensional space. Completeness here means that it is impossible to find in this space another vector  $\mathbf{x}_4$ , that is orthogonal to all three vectors  $\mathbf{x}_1$ ,  $\mathbf{x}_2$ , and  $\mathbf{x}_3$ . Any vector in this space can then be represented (with zero error) in terms of these three vectors. Such vectors are known as **basis** vectors. If a set of vectors  $\{\mathbf{x}_i\}$  is not complete, the error in the approximation will generally not be zero. Thus, in the three-dimensional case discussed, it is generally not possible to represent a vector  $\mathbf{g}$  in terms of only two basis vectors without an error.

The choice of basis vectors is not unique. In fact, a set of basis vectors corresponds to a particular choice of coordinate system. Thus, a three-dimensional vector  $\mathbf{g}$  may be represented in many different ways, depending on the coordinate system used.

To summarize, if a set of vectors  $\{\mathbf{x}_i\}$  is mutually orthogonal, that is, if

$$\mathbf{x}_m \cdot \mathbf{x}_n = \begin{cases} 0 & m \neq n \\ |\mathbf{x}_m|^2 & m = n \end{cases} \quad (2.52)$$

and if this basis set is complete, a vector  $\mathbf{g}$  in this space can be expressed as

$$\mathbf{g} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + c_3\mathbf{x}_3 \quad (2.53)$$

where the constants  $c_i$  are given by

$$c_i = \frac{\mathbf{g} \cdot \mathbf{x}_i}{\mathbf{x}_i \cdot \mathbf{x}_i} \quad (2.54a)$$

$$= \frac{1}{|\mathbf{x}_i|^2} \mathbf{g} \cdot \mathbf{x}_i \quad i = 1, 2, 3 \quad (2.54b)$$

## 2.7.2 Orthogonal Signal Space

We continue with our signal approximation problem using clues and insights developed for vector approximation. As before, we define the orthogonality of a signal set  $x_1(t)$ ,  $x_2(t)$ ,  $\dots$ ,  $x_N(t)$  over the interval  $[t_1, t_2]$  as

$$\int_{t_1}^{t_2} x_m(t)x_n^*(t) dt = \begin{cases} 0 & m \neq n \\ E_n & m = n \end{cases} \quad (2.55)$$

If the energies  $E_n = 1$  for all  $n$ , then the set is **normalized** and is called an **orthonormal set**. An orthogonal set can always be normalized by dividing  $x_n(t)$  by  $\sqrt{E_n}$  for all  $n$ . Now, consider the problem of approximating a signal  $g(t)$  over the interval  $[t_1, t_2]$  by a set of  $N$  mutually orthogonal signals  $x_1(t)$ ,  $x_2(t)$ ,  $\dots$ ,  $x_N(t)$ :

$$g(t) \simeq c_1x_1(t) + c_2x_2(t) + \dots + c_Nx_N(t) \quad (2.56a)$$

$$= \sum_{n=1}^N c_nx_n(t) \quad t_1 \leq t \leq t_2 \quad (2.56b)$$

It can be shown that  $E_e$ , the energy of the error signal  $e(t)$ , in this approximation is minimized if we choose<sup>2</sup>

$$c_n = \frac{\int_{t_1}^{t_2} g(t)x_n^*(t) dt}{\int_{t_1}^{t_2} x_n^2(t) dt} \quad (2.57a)$$

$$= \frac{1}{E_n} \int_{t_1}^{t_2} g(t)x_n^*(t) dt \quad n = 1, 2, \dots, N \quad (2.57b)$$

Moreover, if the orthogonal set is **complete**, the error energy  $\rightarrow 0$ , and the representation in Eqs.(2.56) is no longer an approximation, but an equality,

$$\begin{aligned} g(t) &= c_1x_1(t) + c_2x_2(t) + \dots + c_nx_n(t) + \dots \\ &= \sum_{n=1}^{\infty} c_nx_n(t) \quad t_1 \leq t \leq t_2 \end{aligned} \quad (2.58)$$

where the coefficients  $c_n$  are given by Eq. (2.57). Because the error signal energy approaches zero, it follows that the energy of  $g(t)$  is now equal to the sum of the energies of its orthogonal components  $c_1x_1(t)$ ,  $c_2x_2(t)$ ,  $c_3x_3(t)$ ,  $\dots$ .

The series on the right-hand side of Eq. (2.58) is called the **generalized Fourier series** of  $g(t)$  with respect to the set  $\{x_n(t)\}$ . When the set  $\{x_n(t)\}$  is such that the error energy  $E_e \rightarrow 0$  as  $N \rightarrow \infty$  for every member of some particular class, we say that the set  $\{x_n(t)\}$  is complete on  $[t_1, t_2]$  for that class of  $g(t)$ , and the set  $\{x_n(t)\}$  is called a set of **basis functions** or **basis signals**. Unless otherwise mentioned, in the future we shall consider only the class of energy signals.

Thus, when the set  $\{x_n(t)\}$  is complete, we have the equality (2.58). One subtle point that must be understood clearly is the meaning of equality in Eq. (2.58). *The equality here is not an equality in the ordinary sense, but in the sense that the error energy, that is, the energy of the difference between the two sides of Eq. (2.58), approaches zero.* If the equality exists in the ordinary sense, the error energy is always zero, but the converse is not necessarily true. The error energy can approach zero even though  $e(t)$ , the difference between the two sides, is nonzero at some isolated instants. This is because even if  $e(t)$  is nonzero at such instants, the area under  $e^2(t)$  is still zero. Thus, the Fourier series on the right-hand side of Eq. (2.58) may differ from  $g(t)$  at a finite number of points. In fact, when  $g(t)$  has a jump discontinuity at  $t = t_0$ , the corresponding Fourier series at  $t_0$  converges to the mean of  $g(t_0^+)$  and  $g(t_0^-)$ .

### Parseval's Theorem

Recall that the energy of the sum of orthogonal signals is equal to the sum of their energies. Therefore, the energy of the right-hand side of Eq. (2.58) is the sum of the energies of the individual orthogonal components. The energy of a component  $c_nx_n(t)$  is  $c_n^2E_n$ . Equating the energies of the two sides of Eq. (2.58) yields

$$\begin{aligned} E_g &= c_1^2E_1 + c_2^2E_2 + c_3^2E_3 + \dots \\ &= \sum_n c_n^2E_n \end{aligned} \quad (2.59)$$

This important result is called **Parseval's theorem**. Recall that the signal energy (the area under the squared value of a signal) is analogous to the square of the length of a vector in the

vector-signal analogy. In vector space we know that the square of the length of a vector is equal to the sum of the squares of the lengths of its orthogonal components. Parseval's theorem [Eq. (2.59)] is the statement of this fact as applied to signals.

### Some Examples of Generalized Fourier Series

The signal representation by generalized Fourier series shows that signals are vectors in every sense. Just as a vector can be represented as a sum of its components in a variety of ways, depending on the choice of a coordinate system, a signal can be represented as a sum of its components in a variety of ways. Just as we have vector coordinate systems formed by mutually orthogonal vectors, such as rectangular, cylindrical, spherical, and so on, we also have signal coordinate systems (basis signals) formed by a variety of sets of mutually orthogonal signals. There exists a large number of orthogonal signal sets which can be used as basis signals for generalized Fourier series. Some well-known signal sets are trigonometric (sinusoid) functions, exponential functions, Walsh functions, Bessel functions, Legendre polynomials, Laguerre functions, Jacobi polynomials, Hermite polynomials, and Chebyshev polynomials. The functions that concern us most in this book are the trigonometric and the exponential sets discussed in the rest of this chapter.

## 2.8 TRIGONOMETRIC FOURIER SERIES

Consider a signal set:

$$\{1, \cos \omega_0 t, \cos 2\omega_0 t, \dots, \cos n\omega_0 t, \dots, \sin \omega_0 t, \sin 2\omega_0 t, \dots, \sin n\omega_0 t, \dots\} \quad (2.60)$$

A sinusoid of frequency  $n\omega_0$  is called the  $n$ th **harmonic** of the sinusoid of frequency  $\omega_0$  when  $n$  is an integer. The sinusoid of frequency  $\omega_0$  serves as an anchor in this set, called the **fundamental**, of which all the remaining terms are harmonics. Note that the constant term 1 is the 0th harmonic in this set because  $\cos(0 \times \omega_0 t) = 1$ . We can show that this set is orthogonal over any interval of duration  $T_0 = 2\pi/\omega_0$ , which is the period of the fundamental. This follows from the equations (proved in Appendix A.1)

$$\int_{T_0} \cos n\omega_0 t \cos m\omega_0 t dt = \begin{cases} 0 & n \neq m \\ \frac{T_0}{2} & m = n \neq 0 \end{cases} \quad (2.61a)$$

$$\int_{T_0} \sin n\omega_0 t \sin m\omega_0 t dt = \begin{cases} 0 & n \neq m \\ \frac{T_0}{2} & n = m \neq 0 \end{cases} \quad (2.61b)$$

and

$$\int_{T_0} \sin n\omega_0 t \cos m\omega_0 t dt = 0 \quad \text{for all } n \text{ and } m \quad (2.61c)$$

The notation  $\int_{T_0}$  means integral over an interval from  $t = t_1$  to  $t = t_1 + T_0$  for any value of  $t_1$ . These equations show that the set (2.60) is orthogonal over any contiguous interval of duration  $T_0$ . This is the **trigonometric set**, which can be shown to be a complete set.<sup>3, 4</sup> Therefore, we can express a signal  $g(t)$  by a trigonometric Fourier series over any interval of duration  $T_0$  seconds as

$$g(t) = a_0 + a_1 \cos \omega_0 t + a_2 \cos 2\omega_0 t + \cdots \\ + b_1 \sin \omega_0 t + b_2 \sin 2\omega_0 t + \cdots \quad t_1 \leq t \leq t_1 + T_0 \quad (2.62a)$$

or

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + b_n \sin n\omega_0 t \quad t_1 \leq t \leq t_1 + T_0 \quad (2.62b)$$

where

$$\omega_0 = \frac{2\pi}{T_0} \quad (2.63)$$

Using Eq. (2.57), we can determine the Fourier coefficients  $a_0$ ,  $a_n$ , and  $b_n$ . Thus,

$$a_n = \frac{\int_{t_1}^{t_1+T_0} g(t) \cos n\omega_0 t \, dt}{\int_{t_1}^{t_1+T_0} \cos^2 n\omega_0 t \, dt} \quad (2.64)$$

The integral in the denominator of Eq. (2.64) as seen from Eq. (2.61a) (with  $m = n$ ) is  $T_0/2$  when  $n \neq 0$ . Moreover, for  $n = 0$ , the denominator is  $T_0$ . Hence,

$$a_0 = \frac{1}{T_0} \int_{t_1}^{t_1+T_0} g(t) \, dt \quad (2.65a)$$

and

$$a_n = \frac{2}{T_0} \int_{t_1}^{t_1+T_0} g(t) \cos n\omega_0 t \, dt \quad n = 1, 2, 3, \dots \quad (2.65b)$$

Using a similar argument, we obtain

$$b_n = \frac{2}{T_0} \int_{t_1}^{t_1+T_0} g(t) \sin n\omega_0 t \, dt \quad n = 1, 2, 3, \dots \quad (2.65c)$$

### Compact Trigonometric Fourier Series

The trigonometric Fourier series in Eq. (2.62) contains sine and cosine terms of the same frequency. We can combine the two terms in a single term of the same frequency using the trigonometric identity

$$a_n \cos n\omega_0 t + b_n \sin n\omega_0 t = C_n \cos (n\omega_0 t + \theta_n) \quad (2.66)$$

where

$$C_n = \sqrt{a_n^2 + b_n^2} \quad (2.67a)$$

$$\theta_n = \tan^{-1} \left( \frac{-b_n}{a_n} \right) \quad (2.67b)$$

For consistency we denote the dc term  $a_0$  by  $C_0$ , that is,

$$C_0 = a_0 \quad (2.67c)$$

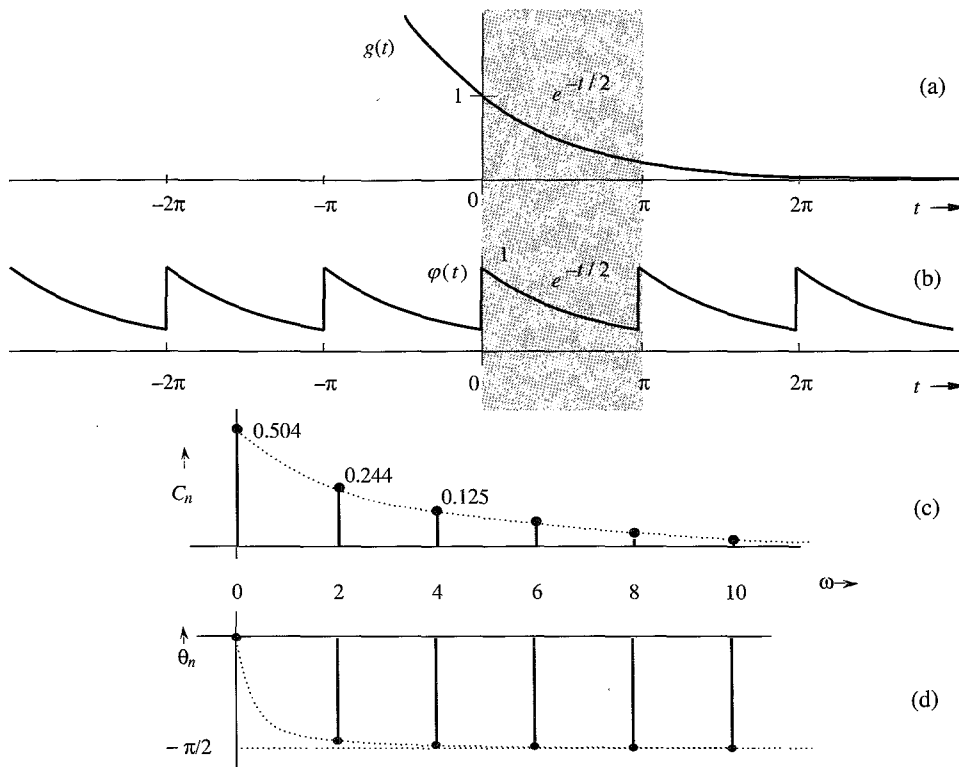
Using the identity (2.66), the trigonometric Fourier series in Eq. (2.62) can be expressed in the **compact form** of the trigonometric Fourier series as

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \quad t_1 \leq t \leq t_1 + T_0 \quad (2.68)$$

where the coefficients  $C_n$  and  $\theta_n$  are computed from  $a_n$  and  $b_n$  using Eqs. (2.67).

Equation (2.65a) shows that  $a_0$  (or  $C_0$ ) is the average value of  $g(t)$  (averaged over one period). This value can often be determined by inspection of  $g(t)$ .

**EXAMPLE 2.7** Find the compact trigonometric Fourier series for the exponential  $e^{-t/2}$  shown in Fig. 2.21a over the interval  $0 \leq t \leq \pi$ .



**Figure 2.21** Periodic signal and its Fourier spectra.

Because we are required to represent  $g(t)$  by the trigonometric Fourier series over the interval  $0 \leq t \leq \pi$ ,  $T_0 = \pi$ , and the fundamental frequency is

$$\omega_0 = \frac{2\pi}{T_0} = 2 \text{ rad/s}$$

Therefore,

$$g(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos 2nt + b_n \sin 2nt \quad 0 \leq t \leq \pi$$

where [from Eq. (2.65a)]

$$a_0 = \frac{1}{\pi} \int_0^{\pi} e^{-t/2} dt = 0.50$$

$$a_n = \frac{2}{\pi} \int_0^{\pi} e^{-t/2} \cos 2nt dt = 0.504 \left( \frac{2}{1 + 16n^2} \right)$$

and

$$b_n = \frac{2}{\pi} \int_0^{\pi} e^{-t/2} \sin 2nt dt = 0.504 \left( \frac{8n}{1 + 16n^2} \right)$$

Therefore,

$$g(t) = 0.504 \left[ 1 + \sum_{n=1}^{\infty} \frac{2}{1 + 16n^2} (\cos 2nt + 4n \sin 2nt) \right] \quad 0 \leq t \leq \pi$$

To find the compact Fourier series, we find its coefficients using Eqs. (2.67) as

$$C_0 = a_0 = 0.504$$

$$C_n = \sqrt{a_n^2 + b_n^2} = 0.504 \sqrt{\frac{4}{(1 + 16n^2)^2} + \frac{64n^2}{(1 + 16n^2)^2}} = 0.504 \left( \frac{2}{\sqrt{1 + 16n^2}} \right)$$

$$\theta_n = \tan^{-1} \left( \frac{-b_n}{a_n} \right) = \tan^{-1}(-4n) = -\tan^{-1} 4n \quad (2.69)$$

The amplitudes and phases of the dc and the first seven harmonics are computed from Eq. (2.69) and displayed in Table 2.1. Using these numerical values, we can express  $g(t)$  in the compact trigonometric Fourier series as

$$g(t) = 0.504 + 0.504 \sum_{n=1}^{\infty} \frac{2}{\sqrt{1 + 16n^2}} \cos(2nt - \tan^{-1} 4n) \quad 0 \leq t \leq \pi \quad (2.70a)$$

$$= 0.504 + 0.244 \cos(2t - 75.96^\circ) + 0.125 \cos(4t - 82.87^\circ) \\ + 0.084 \cos(6t - 85.24^\circ) + 0.063 \cos(8t - 86.42^\circ) + \dots \quad 0 \leq t \leq \pi \quad (2.70b)$$

**Table 2.1**

n	0	1	2	3	4	5	6	7
$C_n$	0.504	0.244	0.125	0.084	0.063	0.0504	0.042	0.036
$\theta_n$	0	-75.96	-82.87	-85.24	-86.42	-87.14	-87.61	-87.95

## Periodicity of the Trigonometric Fourier Series

We have shown how an arbitrary signal  $g(t)$  may be expressed as a trigonometric Fourier series over any interval of  $T_0$  seconds. The Fourier series is equal to  $g(t)$  over this interval alone. Outside this interval the series is not necessarily equal to  $g(t)$ . It would be interesting to find out what happens to the Fourier series outside this interval. We now show that the trigonometric Fourier series is a periodic function of period  $T_0$  (the period of the fundamental).

Let us denote the trigonometric Fourier series on the right-hand side of Eq. (2.68) by  $\varphi(t)$ . Therefore,

$$\varphi(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \quad \text{for all } t$$

and

$$\begin{aligned} \varphi(t + T_0) &= C_0 + \sum_{n=1}^{\infty} C_n \cos[n\omega_0(t + T_0) + \theta_n] \\ &= C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + 2n\pi + \theta_n) \\ &= C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n) \\ &= \varphi(t) \quad \text{for all } t \end{aligned} \tag{2.71}$$

This shows that the trigonometric Fourier series is a periodic function of period  $T_0$  (the period of its fundamental). For instance,  $\varphi(t)$ , the Fourier series on the right-hand side of Eq. (2.70), is a periodic function in which the segment of  $g(t)$  in Fig. 2.21a over the interval  $0 \leq t \leq \pi$  repeats periodically every  $\pi$  seconds, as shown in Fig. 2.21b.\* Thus, when we represent a signal  $g(t)$  by the trigonometric Fourier series over a certain interval of duration  $T_0$ , the function  $g(t)$  and its Fourier series  $\varphi(t)$  need only be equal over that interval of  $T_0$  seconds. Outside this interval, the Fourier series repeats periodically with period  $T_0$ . Now if the function  $g(t)$  were itself to be periodic with period  $T_0$ , then a Fourier series representing  $g(t)$  over an interval  $T_0$  will also represent  $g(t)$  for all  $t$  (not just over the interval  $T_0$ ). Moreover, such a periodic signal  $g(t)$  can be generated by a periodic repetition of any of its segments of duration  $T_0$  (see Sec. 2.2.3, Fig. 2.7). Therefore, the trigonometric Fourier series representing a segment of  $g(t)$  of duration  $T_0$  starting at any instant represents  $g(t)$  for all  $t$ . This means in computing the coefficients  $a_0$ ,  $a_n$ , and  $b_n$ , we may use any value for  $t_1$  in Eqs. (2.65). In other words, we may perform this integration over any interval of  $T_0$ . Thus, the Fourier coefficients of a series representing a periodic signal  $g(t)$  (for all  $t$ ) can be expressed as

$$a_0 = \frac{1}{T_0} \int_{T_0} g(t) dt \tag{2.72a}$$

$$a_n = \frac{2}{T_0} \int_{T_0} g(t) \cos n\omega_0 t dt \quad n = 1, 2, 3, \dots \tag{2.72b}$$

and

$$b_n = \frac{2}{T_0} \int_{T_0} g(t) \sin n\omega_0 t dt \quad n = 1, 2, 3, \dots \tag{2.72c}$$

where  $\int_{T_0}$  means that the integration is performed over any interval of  $T_0$  seconds.

\* In reality, the series convergence at the points of discontinuity shows about 9% overshoot (Gibbs phenomenon).<sup>2</sup>



## Fourier Spectrum

The compact trigonometric Fourier series in Eq. (2.68) indicates that a periodic signal  $g(t)$  can be expressed as a sum of sinusoids of frequencies 0 (dc),  $\omega_0$ ,  $2\omega_0$ ,  $\dots$ ,  $n\omega_0$ ,  $\dots$ , whose amplitudes are  $C_0$ ,  $C_1$ ,  $C_2$ ,  $\dots$ ,  $C_n$ ,  $\dots$  and whose phases are 0,  $\theta_1$ ,  $\theta_2$ ,  $\dots$ ,  $\theta_n$ ,  $\dots$ . We can readily plot amplitude  $C_n$  vs.  $\omega$  (**amplitude spectrum**) and  $\theta_n$  vs.  $\omega$  (**phase spectrum**). These two plots together are the **frequency spectra** of  $g(t)$ .

Figure 2.21c and d show the amplitude and phase spectra for the periodic signal  $\varphi(t)$  in Fig. 2.21b. These spectra tell us at a glance the frequency composition of  $\varphi(t)$ , that is, the amplitudes and phases of various sinusoidal components of  $\varphi(t)$ . Knowing the frequency spectra, we can reconstruct or synthesize  $\varphi(t)$ , as shown on the right-hand side of Eq. (2.70). Therefore, the frequency spectra in Fig. 2.21c and d provide an alternative description—the **frequency-domain description** of  $\varphi(t)$ . The **time-domain description** of  $\varphi(t)$  is shown in Fig. 2.21b. *A signal, therefore, has a dual identity: the time-domain identity  $\varphi(t)$  and the frequency-domain identity (Fourier spectra). The two identities complement each other. Taken together, they provide a better understanding of a signal.*

## Series Convergence at Jump Discontinuities

When there is a jump discontinuity in a periodic signal  $g(t)$ , its Fourier series at the point of discontinuity converges to an average of the left-hand and right-hand limits of  $g(t)$  at the instant of discontinuity\*. In Fig. 2.21b, for instance, the periodic signal  $\varphi(t)$  is discontinuous at  $t = 0$  with  $\varphi(0^+) = 1$  and  $\varphi(0^-) = e^{-\pi/2} = 0.208$ . The corresponding Fourier series converges to a value of  $(1 + 0.208)/2 = 0.604$  at  $t = 0$ . This is easily verified from Eq. (2.70b) by setting  $t = 0$ .

## Existence of the Fourier Series: Dirichlet Conditions

There are two basic conditions for the existence of the Fourier series.

1. For the series to exist, the coefficients  $a_0$ ,  $a_n$ , and  $b_n$  in Eqs. (2.65) must be finite. From Eqs. (2.65) it follows that the existence of these coefficients is guaranteed if  $g(t)$  is absolutely integrable over one period; that is,

$$\int_{T_0} |g(t)| dt < \infty \quad (2.73)$$

This is known as the **weak Dirichlet condition**. *If a function  $g(t)$  satisfies the weak Dirichlet condition, the existence of a Fourier series is guaranteed, but the series may not converge at every point.* For example, if a function  $g(t)$  is infinite at some point, then obviously the series representing the function will be nonconvergent at that point. Similarly, if a function has an infinite number of maxima and minima in one period, then the function contains an appreciable amount of components of frequencies approaching infinity. Thus, the higher coefficients in the series do not decay rapidly, so that the series will not converge rapidly or uniformly. Thus, for a convergent Fourier series, in addition to condition (2.73), we require that:

2. The function  $g(t)$  have only a finite number of maxima and minima in one period, and it may have only a finite number of finite discontinuities in one period.

These two conditions are known as the **strong Dirichlet conditions**. We note here that any periodic waveform that can be generated in a laboratory satisfies strong Dirichlet

\* This behavior of the Fourier series is dictated by its error energy minimization property, discussed in Sec. 2.7.

conditions, and hence possesses a convergent Fourier series. Thus, a physical possibility of a periodic waveform is a valid and sufficient condition for the existence of a convergent series.

**EXAMPLE 2.8** Find the compact trigonometric Fourier series for the periodic square wave  $w(t)$  shown in Fig. 2.22a, and sketch its amplitude and phase spectra.

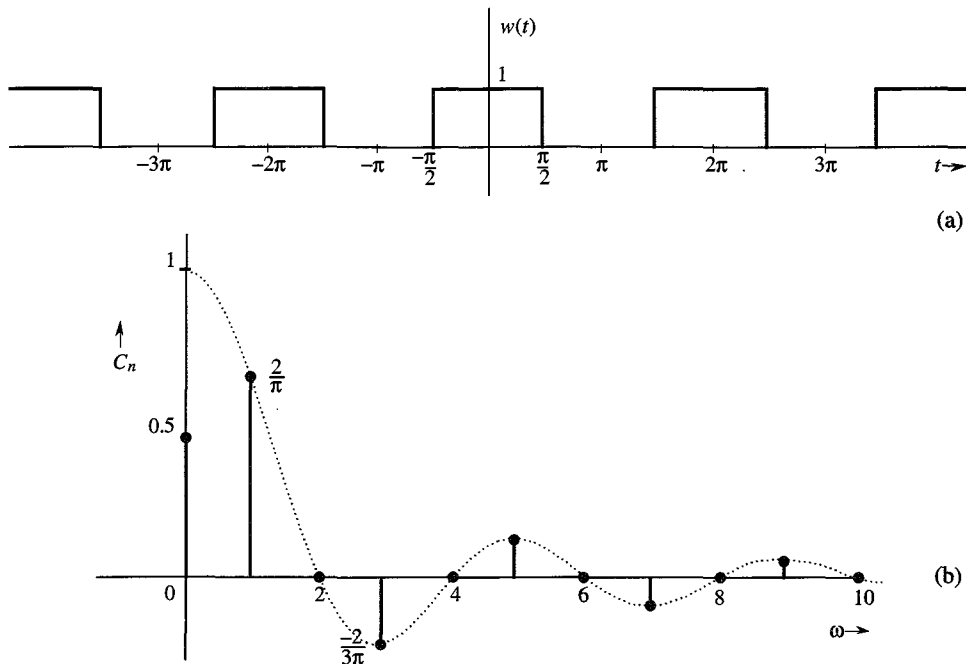


Figure 2.22 Square pulse periodic signal and its Fourier spectra.

The Fourier series is

$$w(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega_0 t + b_n \sin n\omega_0 t$$

where

$$a_0 = \frac{1}{T_0} \int_{T_0} w(t) dt$$

In the preceding equation, we may integrate  $w(t)$  over any interval of duration  $T_0$ . Figure 2.22a shows that the best choice for a region of integration is from  $-T_0/2$  to  $T_0/2$ . Because  $w(t) = 1$  only over  $(-T_0/4, T_0/4)$  and  $w(t) = 0$  over the remaining segment,

$$a_0 = \frac{1}{T_0} \int_{-T_0/4}^{T_0/4} dt = \frac{1}{2} \quad (2.74a)$$

We could have found  $a_0$ , the average value of  $w(t)$ , to be  $1/2$  merely by inspection of  $w(t)$  in Fig. 2.22a. Also,

$$\begin{aligned}
 a_n &= \frac{2}{T_0} \int_{-T_0/4}^{T_0/4} \cos n\omega_0 t \, dt = \frac{2}{n\pi} \sin \left( \frac{n\pi}{2} \right) \\
 &= \begin{cases} 0 & n \text{ even} \\ \frac{2}{\pi n} & n = 1, 5, 9, 13, \dots \\ -\frac{2}{\pi n} & n = 3, 7, 11, 15, \dots \end{cases} \quad (2.74b)
 \end{aligned}$$

$$b_n = \frac{2}{T_0} \int_{-T_0/4}^{T_0/4} \sin nt \, dt = 0 \quad (2.74c)$$

In these derivations we used the fact that  $\omega_0 T_0 = 2\pi$ . Therefore,

$$w(t) = \frac{1}{2} + \frac{2}{\pi} \left( \cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t - \frac{1}{7} \cos 7\omega_0 t + \dots \right) \quad (2.75)$$

Observe that  $b_n = 0$  and all the sine terms are zero. Only the cosine terms appear in the trigonometric series. The series is therefore already in compact form, except that the amplitudes of alternating harmonics are negative. Now by definition, amplitudes  $C_n$  are positive [see Eq. (2.67a)]. The negative sign can be accommodated by a phase of  $\pi$  radians. This can be seen from the trigonometric identity\*

$$-\cos x = \cos(x - \pi)$$

Using this fact, we can express the series in Eq. (2.75) as

$$\begin{aligned}
 w(t) &= \frac{1}{2} + \frac{2}{\pi} \left[ \cos \omega_0 t + \frac{1}{3} \cos(3\omega_0 t - \pi) + \frac{1}{5} \cos 5\omega_0 t \right. \\
 &\quad \left. + \frac{1}{7} \cos(7\omega_0 t - \pi) + \frac{1}{9} \cos 9\omega_0 t + \dots \right]
 \end{aligned}$$

This is the desired form of the compact trigonometric Fourier series. The amplitudes are

$$\begin{aligned}
 C_0 &= \frac{1}{2} \\
 C_n &= \begin{cases} 0 & n \text{ even} \\ \frac{2}{\pi n} & n \text{ odd} \end{cases} \\
 \theta_n &= \begin{cases} 0 & \text{for all } n \neq 3, 7, 11, 15, \dots \\ -\pi & n = 3, 7, 11, 15, \dots \end{cases}
 \end{aligned}$$

We could plot amplitude and phase spectra using these values. We can, however, simplify our task in this special case if we allow amplitude  $C_n$  to take on negative values. If this is allowed, we do not need a phase of  $-\pi$  to account for the sign. This means the phases of all components are zero, and we can discard the phase spectrum and manage with only the amplitude spectrum, as shown in Fig. 2.22b. Observe that there is no loss of information in doing so and that the amplitude spectrum in Fig. 2.22b has the complete information about

\* Because  $\cos(x \pm \pi) = -\cos x$ , we could have chosen the phase  $\pi$  or  $-\pi$ . In fact,  $\cos(x \pm N\pi) = -\cos x$  for any odd integral value of  $N$ . Therefore, the phase can be chosen as  $\pm N\pi$ , where  $N$  is any convenient odd integer.

the Fourier series in Eq. (2.75). Therefore, whenever all sine terms vanish ( $b_n = 0$ ), it is convenient to allow  $C_n$  to take on negative values. This permits the spectral information to be conveyed by a single spectrum—the amplitude spectrum. Because  $C_n$  can be positive as well as negative, the spectrum is called the **amplitude spectrum** rather than the magnitude spectrum.

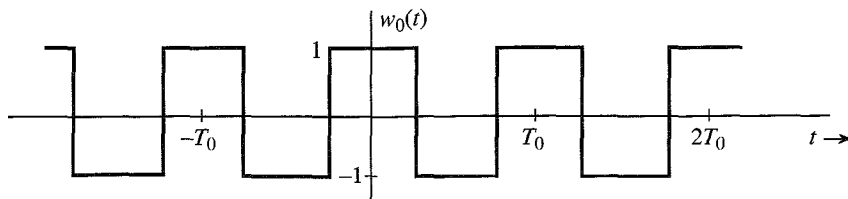


Figure 2.23 Bipolar square pulse periodic signal.

Another useful function that is related to the periodic square wave is the bipolar square wave  $w_0(t)$  shown in Fig. 2.23a. We encounter this signal in switching applications. Note that  $w_0(t)$  is basically  $w(t)$  minus its dc component. It is easy to see that

$$w_0(t) = 2[w(t) - 0.5]$$

Hence, from Eq. (2.75) it follows that

$$w_0(t) = \frac{4}{\pi} \left( \cos \omega_0 t - \frac{1}{3} \cos 3\omega_0 t + \frac{1}{5} \cos 5\omega_0 t - \frac{1}{7} \cos 7\omega_0 t + \dots \right) \quad (2.76)$$

Comparison of this equation with Eq. (2.75) shows that the Fourier components of  $w_0(t)$  are identical to those of  $w(t)$  [Eq. (2.75)] in every respect except for doubling the amplitudes and loss of dc.

**EXAMPLE 2.9** Find the trigonometric Fourier series and sketch the corresponding spectra for the periodic impulse train  $\delta_{T_0}(t)$  shown in Fig. 2.24a.

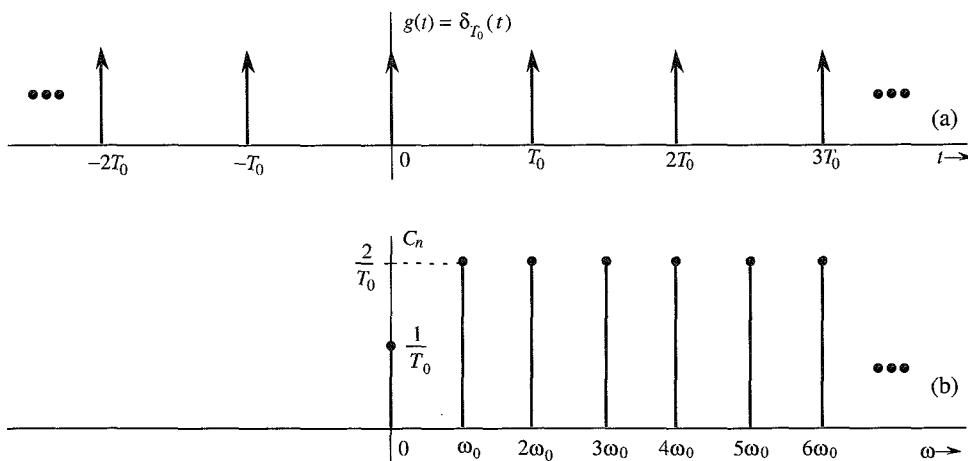


Figure 2.24 Impulse train and its Fourier spectrum.

The trigonometric Fourier series for  $\delta_{T_0}(t)$  is given by

$$\delta_{T_0}(t) = C_0 + \sum C_n \cos(n\omega_0 t + \theta_n) \quad \omega_0 = \frac{2\pi}{T_0}$$

We first compute  $a_0$ ,  $a_n$ , and  $b_n$ :

$$a_0 = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} \delta(t) dt = \frac{1}{T_0}$$

$$a_n = \frac{2}{T_0} \int_{-T_0/2}^{T_0/2} \delta(t) \cos n\omega_0 t dt = \frac{2}{T_0}$$

This result follows from the sampling property (2.19) of the impulse function. Similarly, using the sampling property of the impulse, we obtain

$$b_n = \frac{2}{T_0} \int_{-T_0/2}^{T_0/2} \delta(t) \sin n\omega_0 t dt = 0$$

Therefore,  $C_0 = 1/T_0$ ,  $C_n = 2/T_0$ , and  $\theta_n = 0$ . Thus,

$$\delta_{T_0}(t) = \frac{1}{T_0} \left( 1 + 2 \sum_{n=1}^{\infty} \cos n\omega_0 t \right) \quad (2.77)$$

Figure 2.24b shows the amplitude spectrum. The phase spectrum is zero.

### Effect of Symmetry

The Fourier series for the signal  $g(t)$  in Fig. 2.21a (Example 2.7) consists of sine and cosine terms, but the series for the signal  $w(t)$  in Fig. 2.22a (Example 2.8) consists of cosine terms only. In some cases the Fourier series consists of sine terms only. This is no accident. It can be shown that the Fourier series of any even periodic function  $g(t)$  consists of cosine terms only and the series of any odd periodic function  $g(t)$  consists of sine terms only (see Prob. 2.8-3).

## 2.9 EXPONENTIAL FOURIER SERIES

It is shown in Appendix A.2 that the set of exponentials  $e^{jn\omega_0 t}$  ( $n = 0, \pm 1, \pm 2, \dots$ ) is orthogonal over any interval of duration  $T_0 = 2\pi/\omega_0$ ; that is,

$$\int_{T_0} e^{jm\omega_0 t} (e^{jn\omega_0 t})^* dt = \int_{T_0} e^{j(m-n)\omega_0 t} dt = \begin{cases} 0 & m \neq n \\ T_0 & m = n \end{cases} \quad (2.78)$$

Moreover, this set is a complete set.<sup>3, 4</sup> From Eqs. (2.58) and (2.57) it follows that a signal  $g(t)$  can be expressed over an interval of duration  $T_0$  seconds as an exponential Fourier series

$$g(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \quad (2.79)$$

where [see Eq. (2.57)]

$$D_n = \frac{1}{T_0} \int_{T_0} g(t) e^{-jn\omega_0 t} dt \quad (2.80)$$

The exponential Fourier series is basically another form of the trigonometric Fourier series. Each sinusoid of frequency  $\omega$  can be expressed as the sum of the two exponentials  $e^{j\omega t}$  and  $e^{-j\omega t}$ . This results in the exponential Fourier series consisting of components of the form  $e^{jn\omega_0 t}$  with  $n$  varying from  $-\infty$  to  $\infty$ . The exponential Fourier series in Eq. (2.79) is periodic with period  $T_0$ .

In order to see its close connection with the trigonometric series, we shall rederive the exponential Fourier series from the trigonometric Fourier series. A sinusoid in the trigonometric series can be expressed as a sum of two exponentials using Euler's formula:

$$\begin{aligned} C_n \cos(n\omega_0 t + \theta_n) &= \frac{C_n}{2} [e^{j(n\omega_0 t + \theta_n)} + e^{-j(n\omega_0 t + \theta_n)}] \\ &= \left(\frac{C_n}{2} e^{j\theta_n}\right) e^{jn\omega_0 t} + \left(\frac{C_n}{2} e^{-j\theta_n}\right) e^{-jn\omega_0 t} \\ &= D_n e^{jn\omega_0 t} + D_{-n} e^{-jn\omega_0 t} \end{aligned} \quad (2.81)$$

where

$$\begin{aligned} D_n &= \frac{1}{2} C_n e^{j\theta_n} \\ D_{-n} &= \frac{1}{2} C_n e^{-j\theta_n} \end{aligned} \quad (2.82)$$

The compact trigonometric Fourier series of a periodic signal  $g(t)$  is given by

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

The use of Eq. (2.81) in the preceding equation (and letting  $C_0 = D_0$ ) yields

$$\begin{aligned} g(t) &= D_0 + \sum_{n=1}^{\infty} D_n e^{jn\omega_0 t} + D_{-n} e^{-jn\omega_0 t} \\ &= D_0 + \sum_{n=-\infty}^{\infty} (n \neq 0) D_n e^{jn\omega_0 t} \end{aligned}$$

which is precisely equivalent to Eq. (2.79) derived earlier. Equations (2.82) show the close connection between the coefficients of the trigonometric and the exponential Fourier series.

Observe the compactness of expressions (2.79) and (2.80) and compare them to expressions corresponding to trigonometric Fourier series. These two equations demonstrate very clearly the principal virtue of exponential Fourier series. First, the form of the series is more compact. Second, the mathematical expression for deriving the coefficients of the series is also compact. It is much more convenient to handle the exponential series than the trigonometric one. In the system analysis also, the exponential form proves more convenient than the trigonometric form. For these reasons we shall use exponential (rather than trigonometric) representation of signals in the rest of the book.

---

**EXAMPLE 2.10** Find the exponential Fourier series for the signal in Fig. 2.21b (Example 2.7).

In this case,  $T_0 = \pi$ ,  $\omega_0 = 2\pi/T_0 = 2$ , and

$$\varphi(t) = \sum_{n=-\infty}^{\infty} D_n e^{j2nt}$$

where

$$\begin{aligned} D_n &= \frac{1}{T_0} \int_{T_0} \varphi(t) e^{-j2nt} dt \\ &= \frac{1}{\pi} \int_0^{\pi} e^{-t/2} e^{-j2nt} dt \\ &= \frac{1}{\pi} \int_0^{\pi} e^{-(\frac{1}{2} + j2n)t} dt \\ &= \frac{-1}{\pi (\frac{1}{2} + j2n)} e^{-(\frac{1}{2} + j2n)t} \Big|_0^{\pi} \\ &= \frac{0.504}{1 + j4n} \end{aligned} \quad (2.83)$$

and

$$\varphi(t) = 0.504 \sum_{n=-\infty}^{\infty} \frac{1}{1 + j4n} e^{j2nt} \quad (2.84a)$$

$$\begin{aligned} &= 0.504 \left[ 1 + \frac{1}{1 + j4} e^{j2t} + \frac{1}{1 + j8} e^{j4t} + \frac{1}{1 + j12} e^{j6t} + \dots \right. \\ &\quad \left. + \frac{1}{1 - j4} e^{-j2t} + \frac{1}{1 - j8} e^{-j4t} + \frac{1}{1 - j12} e^{-j6t} + \dots \right] \end{aligned} \quad (2.84b)$$

Observe that the coefficients  $D_n$  are complex. Moreover,  $D_n$  and  $D_{-n}$  are conjugates, as expected [see Eqs. (2.82)].

## 2.9.1 Exponential Fourier Spectra

In exponential spectra, we plot coefficients  $D_n$  as a function of  $\omega$ . But since  $D_n$  is complex in general, we need two plots: the real and the imaginary parts of  $D_n$  or the amplitude (magnitude) and the angle of  $D_n$ . We prefer the latter because of its close connection to the amplitudes and phases of corresponding components of the trigonometric Fourier series. We therefore plot  $|D_n|$  vs.  $\omega$  and  $\angle D_n$  vs.  $\omega$ . This requires that the coefficients  $D_n$  be expressed in polar form as  $|D_n|e^{j\angle D_n}$ .

A comparison of Eqs. (2.65a) and (2.80) (for  $n = 0$ ) shows that  $D_0 = a_0 = C_0$ . Equations (2.82) show that for a real periodic signal the twin coefficients  $D_n$  and  $D_{-n}$  are conjugates, and

$$|D_n| = |D_{-n}| = \frac{1}{2} C_n \quad (2.85a)$$

$$\angle D_n = \theta_n \quad \text{and} \quad \angle D_{-n} = -\theta_n \quad (2.85b)$$

Thus,

$$D_n = |D_n|e^{j\theta_n} \quad \text{and} \quad D_{-n} = |D_n|e^{-j\theta_n} \quad (2.86)$$

Note that  $|D_n|$  are the amplitudes (magnitudes) and  $\angle D_n$  are the angles of various exponential components. From Eqs. (2.85) it follows that the amplitude spectrum ( $|D_n|$  vs.  $\omega$ ) is an even function of  $\omega$  and the angle spectrum ( $\angle D_n$  vs.  $\omega$ ) is an odd function of  $\omega$  when  $g(t)$  is a real signal.

For the series in Example 2.10 [Eq. (2.84b)], for instance,

$$D_0 = 0.504$$

$$D_1 = \frac{0.504}{1 + j4} = 0.122e^{-j75.96^\circ} \implies |D_1| = 0.122 \quad \angle D_1 = -75.96^\circ$$

$$D_{-1} = \frac{0.504}{1 - j4} = 0.122e^{j75.96^\circ} \implies |D_{-1}| = 0.122 \quad \angle D_{-1} = 75.96^\circ$$

and

$$D_2 = \frac{0.504}{1 + j8} = 0.0625e^{-j82.87^\circ} \implies |D_2| = 0.0625 \quad \angle D_2 = -82.87^\circ$$

$$D_{-2} = \frac{0.504}{1 - j8} = 0.0625e^{j82.87^\circ} \implies |D_{-2}| = 0.0625 \quad \angle D_{-2} = 82.87^\circ$$

and so on. Note that  $D_n$  and  $D_{-n}$  are conjugates, as expected [see Eqs. (2.85)].

Figure 2.25 shows the frequency spectra (amplitude and angle) of the exponential Fourier series for the periodic signal  $\varphi(t)$  in Fig. 2.21b.

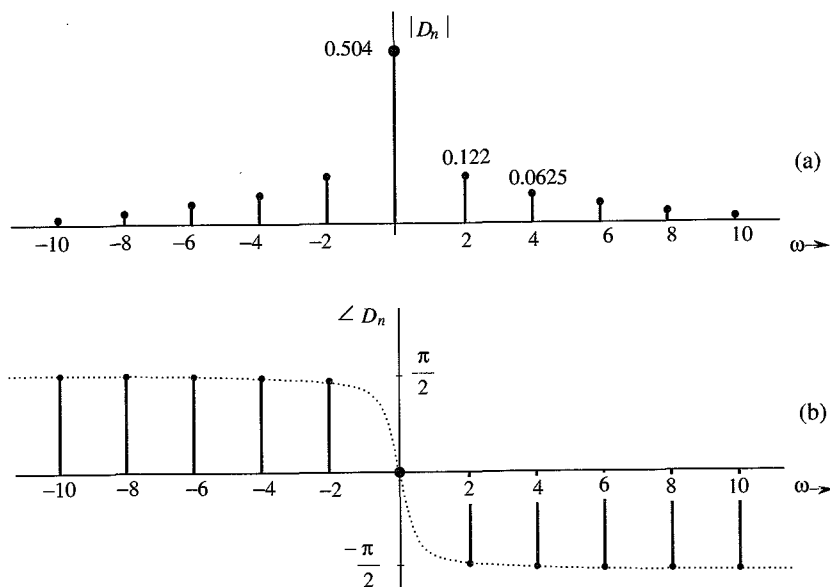


Figure 2.25 Exponential Fourier spectra for the signal in Fig. 2.21a.



We notice some interesting features of these spectra. First, the spectra exist for positive as well as negative values of  $\omega$  (the frequency). Second, the amplitude spectrum is an even function of  $\omega$  and the angle spectrum is an odd function of  $\omega$ . Finally, we see a close connection between these spectra and the spectra of the corresponding trigonometric Fourier series for  $\varphi(t)$  (Fig. 2.21c and d).

### What Is a Negative Frequency?

The existence of the spectrum at negative frequencies is somewhat disturbing because by definition, the frequency (number of repetitions per second) is a positive quantity. How do we interpret a negative frequency? Using a trigonometric identity, the sinusoid of a negative frequency  $-\omega_0$  can be expressed as

$$\cos(-\omega_0 t + \theta) = \cos(\omega_0 t - \theta)$$

This clearly shows that the frequency of a sinusoid  $\cos(\omega_0 t + \theta)$  is  $|\omega_0|$ , which is a positive quantity. The same conclusion is reached by observing that

$$e^{\pm j\omega_0 t} = \cos \omega_0 t \pm j \sin \omega_0 t$$

Thus, the frequency of exponentials  $e^{\pm j\omega_0 t}$  is indeed  $|\omega_0|$ . How do we then interpret the spectral plots for negative values of  $\omega$ ? A healthier way of looking at the situation is to say that *exponential spectra are a graphical representation of coefficients  $D_n$  as a function of  $\omega$* . Existence of the spectrum at  $\omega = -n\omega_0$  is merely an indication of the fact that an exponential component  $e^{-jn\omega_0 t}$  exists in the series. We know that a sinusoid of frequency  $n\omega_0$  can be expressed in terms of a pair of exponentials  $e^{jn\omega_0 t}$  and  $e^{-jn\omega_0 t}$  [see Eq. (2.81)].

Equations (2.85) show the close connection between trigonometric spectra ( $C_n$  and  $\theta_n$ ) and exponential spectra ( $|D_n|$  and  $\angle D_n$ ). The dc components  $D_0$  and  $C_0$  are identical in both spectra. Moreover, the exponential amplitude spectrum  $|D_n|$  is half the trigonometric amplitude spectrum  $C_n$  for  $n \geq 1$ . The exponential angle spectrum  $\angle D_n$  is identical to the trigonometric phase spectrum  $\theta_n$  for  $n \geq 0$ . We can therefore produce the exponential spectra merely by the inspection of trigonometric spectra, and vice versa.

---

**EXAMPLE 2.11** Find the exponential Fourier series for the periodic square wave  $w(t)$  shown in Fig. 2.22a.

We have

$$w(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t}$$

where

$$\begin{aligned} D_n &= \frac{1}{T_0} \int_{T_0} w(t) e^{-jn\omega_0 t} dt \\ &= \frac{1}{T_0} \int_{-T_0/4}^{T_0/4} e^{-jn\omega_0 t} dt \\ &= \frac{1}{-jn\omega_0 T_0} (e^{-jn\omega_0 T_0/4} - e^{jn\omega_0 T_0/4}) \\ &= \frac{2}{n\omega_0 T_0} \sin\left(\frac{n\omega_0 T_0}{4}\right) = \frac{1}{n\pi} \sin\left(\frac{n\pi}{2}\right) \end{aligned}$$

In this case  $D_n$  is real. Consequently, we can do without the phase or angle plot if we plot  $D_n$  vs  $\omega$  instead of the amplitude spectrum ( $|D_n|$  vs.  $\omega$ ), as shown in Fig. 2.26. Compare this spectrum with the trigonometric spectrum in Fig. 2.22b. Observe that  $D_0 = C_0$  and  $|D_n| = |D_{-n}| = C_n/2$ , as expected.

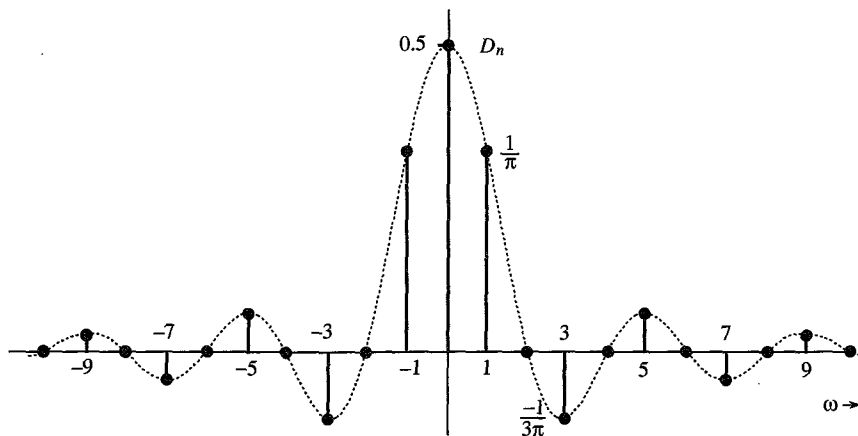


Figure 2.26 Exponential Fourier spectrum of the square pulse periodic signal.

**EXAMPLE 2.12** Find the exponential Fourier series and sketch the corresponding spectra for the impulse train  $\delta_{T_0}(t)$  shown in Fig. 2.27.

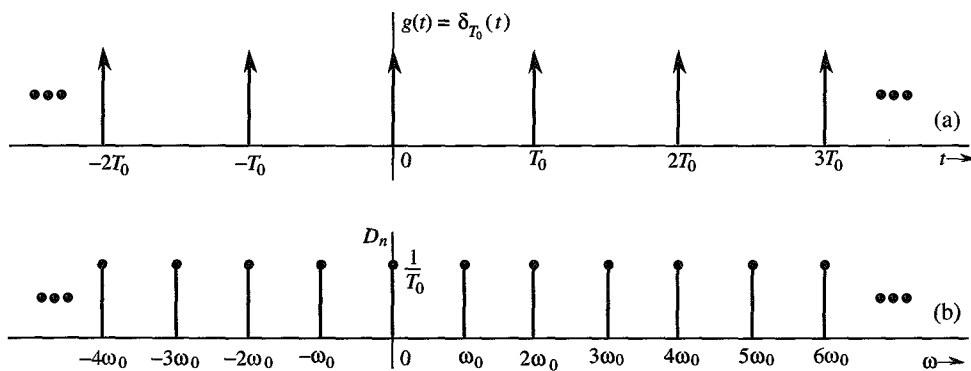


Figure 2.27 Impulse train and its exponential Fourier spectra.

The exponential Fourier series is given by

$$\delta_{T_0}(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \quad \omega_0 = \frac{2\pi}{T_0} \quad (2.87)$$

where

$$D_n = \frac{1}{T_0} \int_{T_0} \delta_{T_0}(t) e^{-jn\omega_0 t} dt$$

Choosing the interval of integration  $(-T_0/2, T_0/2)$  and recognizing that over this interval  $\delta_{T_0}(t) = \delta(t)$ ,

$$D_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} \delta(t) e^{-jn\omega_0 t} dt$$

In this integral, the impulse is located at  $t = 0$ . From the sampling property of the impulse function, the integral on the right-hand side is the value of  $e^{-jn\omega_0 t}$  at  $t = 0$  (where the impulse is located). Therefore,

$$D_n = \frac{1}{T_0} \quad (2.88)$$

and

$$\delta_{T_0}(t) = \frac{1}{T_0} \sum_{n=-\infty}^{\infty} e^{jn\omega_0 t} \quad \omega_0 = \frac{2\pi}{T_0} \quad (2.89)$$

Equation (2.89) shows that the exponential spectrum is uniform ( $D_n = 1/T_0$ ) for all the frequencies, as shown in Fig. 2.27. The spectrum, being real, requires only the amplitude plot. All phases are zero. Compare this spectrum to the trigonometric spectrum shown in Fig. 2.24b. The dc components are identical and the exponential spectrum amplitudes are half those in the trigonometric spectrum for all  $\omega > 0$ .

### Parseval's Theorem

A periodic signal  $g(t)$  is a power signal, and every term in its Fourier series is also a power signal. The power  $P_g$  of  $g(t)$  is equal to the power of its Fourier series. Because the Fourier series consists of terms that are mutually orthogonal over one period, the power of the Fourier series is equal to the sum of the powers of its Fourier components. This follows from Parseval's theorem. We have already demonstrated this result in Example 2.2 for the trigonometric Fourier series. It is also valid for the exponential Fourier series. Thus, for the trigonometric Fourier series

$$g(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

the power of  $g(t)$  is given by

$$P_g = C_0^2 + \frac{1}{2} \sum_{n=1}^{\infty} C_n^2 \quad (2.90)$$

For the exponential Fourier series

$$g(t) = D_0 + \sum_{\substack{n=-\infty \\ (n \neq 0)}}^{\infty} D_n e^{jn\omega_0 t}$$

the power is given by (see Prob. 2.1-7)

$$P_g = \sum_{n=-\infty}^{\infty} |D_n|^2 \quad (2.91a)$$

For a real  $g(t)$ ,  $|D_{-n}| = |D_n|$ . Therefore,

$$P_g = D_0^2 + 2 \sum_{n=1}^{\infty} |D_n|^2 \quad (2.91b)$$

*Comments:* Parseval's theorem occurs in many different forms, such as in Eqs. (2.59), (2.90), and (2.91). Yet another form is found in Eq. (3.64). Although these forms appear different, they all state the same principle, that is, the square of the length of a vector equals the sum of the squares of its orthogonal components. The form (2.59) applies to energy signals, the form (2.90) applies to periodic signals represented by the trigonometric Fourier series, and the form (2.91) applies to periodic signals represented by the exponential Fourier series.

## 2.10 NUMERICAL COMPUTATION OF $D_n$

We can compute  $D_n$  numerically using the **discrete Fourier transform (DFT)**, which uses the samples of a periodic signal  $g(t)$  over one period. The sampling interval is  $T_s$  seconds. Hence, there are  $N_0 = T_0/T_s$  number of samples in one period  $T_0$ . To find the relationship between  $D_n$  and the samples of  $g(t)$ , consider Eq. (2.80),

$$\begin{aligned} D_n &= \frac{1}{T_0} \int_{T_0} g(t) e^{-jn\omega_0 t} dt \\ &= \lim_{T_s \rightarrow 0} \frac{1}{T_0} \sum_{k=0}^{N_0-1} g(kT_s) e^{-jn\omega_0 kT_s} T_s \\ &= \lim_{T_s \rightarrow 0} \frac{1}{N_0} \sum_{k=0}^{N_0-1} g(kT_s) e^{-jn\Omega_0 k} \end{aligned} \quad (2.92)$$

where  $g(kT_s)$  is the  $k$ th sample of  $g(t)$  and

$$\Omega_0 = \omega_0 T_s, \quad N_0 = \frac{T_0}{T_s} \quad (2.93)$$

In practice, it is impossible to make  $T_s \rightarrow 0$  in computing the right-hand side of Eq. (2.92). We can make it small, but not zero because it will increase the data without limit. Thus, we shall ignore the limit on  $T_s$  in Eq. (2.92) with the implicit understanding that  $T_s$  is reasonably small. This results in some computational error, which is inevitable in any numerical evaluation of an integral. The error resulting from nonzero  $T_s$  is called the **aliasing error**, which will be discussed in more details in Chapter 6. Thus, we can express Eq. (2.92) as

$$D_n = \frac{1}{N_0} \sum_{k=0}^{N_0-1} g(kT_s) e^{-jn\Omega_0 k} \quad (2.94)$$

This equation shows that  $D_{n+N_0} = D_n$ . Hence, Eq. (2.94) yields the Fourier spectrum  $D_n$  repeating periodically with period  $N_0$ . This will result in overlapping of various components. To reduce the effect of such overlapping, we need to increase  $N_0$  as much as practicable. We shall see later [Sec. (6.1)] that the overlapping appears as if the spectrum above the  $(N_0/2)$ th harmonics had folded back at this frequency  $(N_0\omega_0/2)$ . Hence, to minimize the effect of this spectral folding, we should make sure that  $D_n$  for  $n \geq N_0/2$  is negligible. The DFT (or FFT) gives the coefficients  $D_n$  for  $n \geq 0$  up to  $n = N_0/2$ . Beyond  $n = N_0/2$ , the coefficients represent the values for negative  $n$  because of the periodicity property  $D_{n+N_0} = D_n$ . For instance, when  $N_0 = 32$ ,  $D_{17} = D_{-15}$ ,  $D_{18} = D_{-14}$ ,  $\dots$ ,  $D_{31} = D_{-1}$ . The cycle repeats again from  $n = 32$  on.

We can use the efficient **fast Fourier transform (FFT)** to compute the right-hand side of Eq. (2.94). We shall use MATLAB to implement the FFT algorithm. For this purpose, we need samples of  $g(t)$  over one period starting at  $t = 0$ . In this algorithm, it is also preferable (although not necessary) that  $N_0$  be a power of 2, that is,  $N_0 = 2^m$ , where  $m$  is an integer.

### Computer Example C2.1

Compute and plot the trigonometric and exponential Fourier spectra for the periodic signal in Fig. 2.21b (Example 2.7).

The samples of  $g(t)$  start at  $t = 0$ , and the last ( $N_0$ )th sample is at  $t = T_0 - T_s$ . (The last sample is not at  $t = T_0$  because the sample at  $t = 0$  is identical to the sample at  $t = T_0$ , and the next cycle begins at  $t = T_0$ .) At the points of discontinuity, the sample value is taken as the average of the values of the function on two sides of the discontinuity. Thus, in the present case, the first sample (at  $t = 0$ ) is not 1, but  $(e^{-\pi/2} + 1)/2 = 0.604$ . To determine  $N_0$ , we require  $D_n$  for  $n \geq N_0/2$  to be relatively small. Because  $g(t)$  has a jump discontinuity,  $D_n$  decays rather slowly as  $1/n$ . Hence, a choice of  $N_0 = 200$  is acceptable because the  $(N_0/2)$ th (100th) harmonic is about 0.01 (about 1%) of the fundamental. However, we also require  $N_0$  to be a power of 2. Hence, we shall take  $N_0 = 256 = 2^8$ .

We write and save a MATLAB file (or program) `c21.m` to compute and plot the Fourier coefficients.

```
% (c21.m)
%M is the number of coefficients to be computed
T0=pi;N0=256;Ts=T0/N0;M=10;
t=0:Ts:Ts*(N0-1); t=t';
g=exp(-t/2);g(1)=0.604;
% fft(g) is the FFT [the sum on the right-hand side of Eq. (2.94)]
Dn=fft(g)/N0
[Dnangle,Dnmag]=cart2pol(real(Dn),imag(Dn));
k=0:length(Dn)-1;k=k';
subplot(211),stem(k,Dnmag)
subplot(212),stem(k,Dnangle)
```

To compute trigonometric Fourier series coefficients, we recall program `c21.m` along with commands to convert  $D_n$  into  $C_n$  and  $\theta_n$ .

```
c21;clg
C0=Dnmag(1); Cn=2*Dnmag(2:M);
```

```

Amplitudes=[C0;Cn]
Angles=Dnangles(1:M);
Angles=Angles*(180/pi);
disp('Amplitudes Angles')
[Amplitudes Angles]
% To Plot the Fourier coefficients
k=0:length(Amplitudes)-1; k=k';
subplot(211),stem(k,Amplitudes)
subplot(212),stem(k,Angles)
ans =
    Amplitudes    Angles
    0.5043         0
    0.2446       -75.9622
    0.1251       -82.8719
    0.0837       -85.2317
    0.0629       -86.4175
    0.0503       -87.1299
    0.0419       -87.6048
    0.0359       -87.9437
    0.0314       -88.1977
    0.0279       -88.3949

```

## REFERENCES

1. A. Papoulis, *The Fourier Integral and Its Applications*, McGraw-Hill, New York, 1962.
2. B. P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA, 1998.
3. P. L. Walker, *The Theory of Fourier Series and Integrals*, Wiley-Interscience, New York, 1986.
4. R. V. Churchill, and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed., McGraw-Hill, New York, 1978.

## PROBLEMS

- 2.1-1 Find the energies of the signals shown in Fig. P2.1-1. Comment on the effect on energy of sign change, time shifting or doubling of the signal. What is the effect on the energy if the signal is multiplied by  $k$ ?

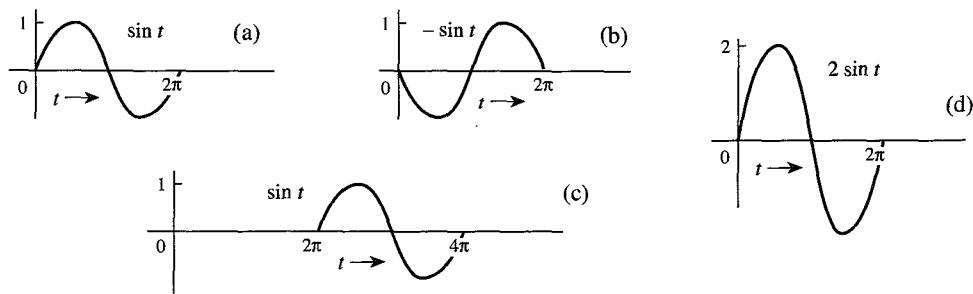


Figure P2.1-1

- 2.1-2 (a)** Find  $E_x$  and  $E_y$ , the energies of the signals  $x(t)$  and  $y(t)$  shown in Fig. P2.1-2a. Sketch the signals  $x(t) + y(t)$  and  $x(t) - y(t)$  and show that the energies of either of these two signals are equal to  $E_x + E_y$ . Repeat the procedure for the signal pair of Fig. P2.1-2b.
- (b)** Repeat the procedure for the signal pair of Fig. P2.1-2c. Are the energies of the signals  $x(t) + y(t)$  and  $x(t) - y(t)$  identical in this case?

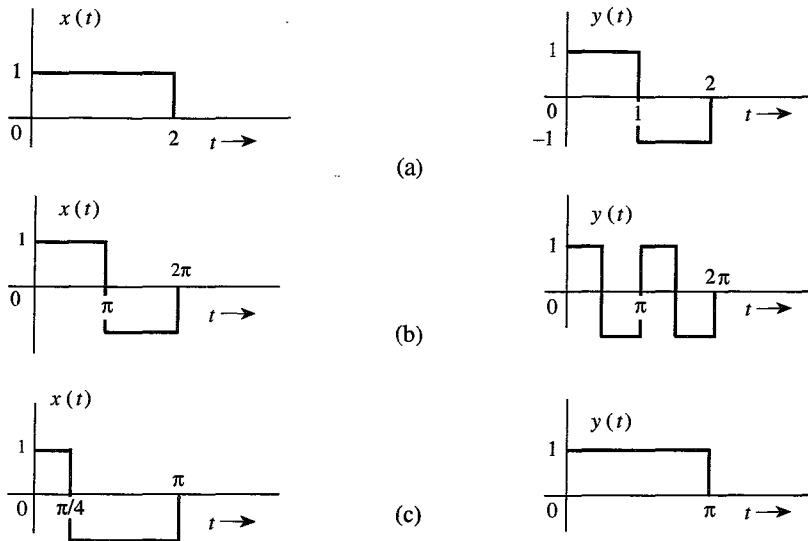


Figure P2.1-2

- 2.1-3** Redo Example 2.2a to find the power of a sinusoid  $C \cos(\omega_0 t + \theta)$  by averaging the signal energy over one period  $2\pi/\omega_0$  (rather than averaging over the infinitely large interval).
- 2.1-4** Show that if  $\omega_1 = \omega_2$ , the power of  $g(t) = C_1 \cos(\omega_1 t + \theta_1) + C_2 \cos(\omega_2 t + \theta_2)$  is  $[C_1^2 + C_2^2 + 2C_1 C_2 \cos(\theta_1 - \theta_2)]/2$ , which is not equal to  $(C_1^2 + C_2^2)/2$ .
- 2.1-5** Find the power of the periodic signal  $g(t)$  shown in Fig. P2.1-5. Find also the powers and the rms values of: (a)  $-g(t)$ ; (b)  $2g(t)$ ; (c)  $cg(t)$ . Comment.

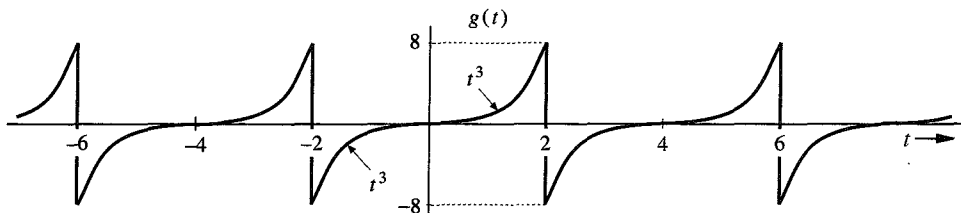


Figure P2.1-5

- 2.1-6** Find the power and the rms value for the signals in: (a) Fig. 2.21b; (b) Fig. 2.22a; (c) Fig. 2.23; (d) Fig. P2.8-4a; (e) Fig. P2.8-4c.
- 2.1-7** Show that the power of a signal  $g(t)$  given by

$$g(t) = \sum_{k=m}^n D_k e^{j\omega_k t} \quad \omega_i \neq \omega_k \text{ for all } i \neq k$$

is (Parseval's theorem)

$$P_g = \sum_{k=m}^n |D_k|^2$$

2.1-8 Determine the power and the rms value for each of the following signals:

- (a)  $10 \cos\left(100t + \frac{\pi}{3}\right)$       (b)  $10 \cos\left(100t + \frac{\pi}{3}\right) + 16 \sin\left(150t + \frac{\pi}{5}\right)$   
 (c)  $(10 + 2 \sin 3t) \cos 10t$       (d)  $10 \cos 5t \cos 10t$   
 (e)  $10 \sin 5t \cos 10t$       (f)  $e^{j\omega t} \cos \omega_0 t$

2.2-1 Show that an exponential  $e^{-at}$  starting at  $-\infty$  is neither an energy nor a power signal for any real value of  $a$ . However, if  $a$  is imaginary, it is a power signal with power  $P_g = 1$  regardless of the value of  $a$ .

2.3-1 In Fig. P2.3-1, the signal  $g_1(t) = g(-t)$ . Express signals  $g_2(t)$ ,  $g_3(t)$ ,  $g_4(t)$ , and  $g_5(t)$  in terms of signals  $g(t)$ ,  $g_1(t)$ , and their time-shifted, time-scaled, or time-inverted versions. For instance  $g_2(t) = g(t - T) + g_1(t - T)$  for some suitable value of  $T$ . Similarly, both  $g_3(t)$  and  $g_4(t)$  can be expressed as  $g(t - T) + g(t + T)$  for some suitable value of  $T$ .  $g_5(t)$  can be expressed as  $g(t)$  time-shifted, time-scaled, and then multiplied by a constant. (These operations may be performed in any order).

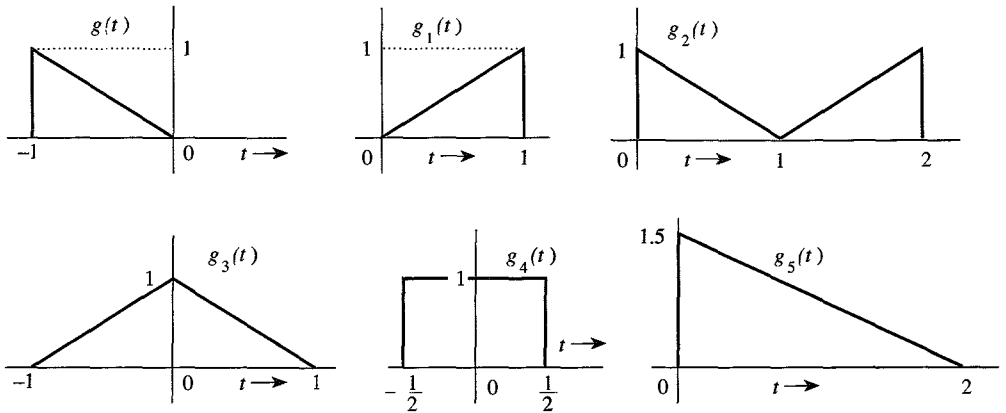


Figure P2.3-1

2.3-2 For the signal  $g(t)$  shown in Fig. P2.3-2, sketch the signals: (a)  $g(-t)$ ; (b)  $g(t + 6)$ ; (c)  $g(3t)$ ; (d)  $g(6 - t)$ .

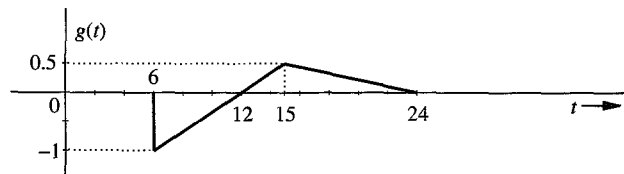


Figure P2.3-2



- 2.3-3** For the signal  $g(t)$  shown in Fig. P2.3-3, sketch: (a)  $g(t - 4)$ ; (b)  $g(t/1.5)$ ; (c)  $g(2t - 4)$  (d)  $g(2 - t)$ . *Hint:* Recall that replacing  $t$  with  $t - T$  delays the signal by  $T$ . Thus,  $g(2t - 4)$  is  $g(2t)$  with  $t$  replaced by  $t - 2$ . Similarly,  $g(2 - t)$  is  $g(-t)$  with  $t$  replaced by  $t - 2$ .

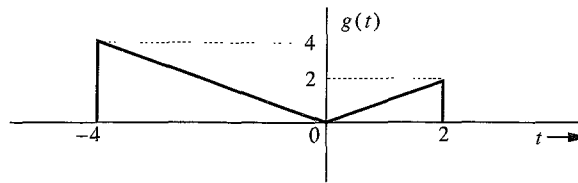


Figure P2.3-3

- 2.3-4** For an energy signal  $g(t)$  with energy  $E_g$ , show that the energy of any one of the signals  $-g(t)$ ,  $g(-t)$ , and  $g(t - T)$  is  $E_g$ . Show also that the energy of  $g(at)$  as well as  $g(at - b)$  is  $E_g/a$ . This shows that time inversion and time shifting do not affect signal energy. On the other hand, time compression of a signal by a factor  $a$  reduces the energy by the factor  $a$ . What is the effect on signal energy if the signal is: (a) time-expanded by a factor  $a$  ( $a > 1$ ); (b) multiplied by a constant  $a$ ?

- 2.4-1** Simplify the following expressions:

$$\begin{array}{ll} \text{(a)} \quad \left( \frac{\sin t}{t^2 + 2} \right) \delta(t) & \text{(b)} \quad \left( \frac{j\omega + 2}{\omega^2 + 9} \right) \delta(\omega) \\ \text{(c)} \quad [e^{-t} \cos(3t - 60^\circ)] \delta(t) & \text{(d)} \quad \left[ \frac{\sin \frac{\pi}{2}(t - 2)}{t^2 + 4} \right] \delta(t - 1) \\ \text{(e)} \quad \left( \frac{1}{j\omega + 2} \right) \delta(\omega + 3) & \text{(f)} \quad \left( \frac{\sin k\omega}{\omega} \right) \delta(\omega) \end{array}$$

*Hint:* Use Eq. (2.18). For part (f) use L'Hôpital's rule.

- 2.4-2** Evaluate the following integrals:

$$\begin{array}{ll} \text{(a)} \quad \int_{-\infty}^{\infty} g(\tau) \delta(t - \tau) d\tau & \text{(b)} \quad \int_{-\infty}^{\infty} \delta(\tau) g(t - \tau) d\tau \\ \text{(c)} \quad \int_{-\infty}^{\infty} \delta(t) e^{-j\omega t} dt & \text{(d)} \quad \int_{-\infty}^{\infty} \delta(t - 2) \sin \pi t dt \\ \text{(e)} \quad \int_{-\infty}^{\infty} \delta(t + 3) e^{-t} dt & \text{(f)} \quad \int_{-\infty}^{\infty} (t^3 + 4) \delta(1 - t) dt \\ \text{(g)} \quad \int_{-\infty}^{\infty} g(2 - t) \delta(3 - t) dt & \text{(h)} \quad \int_{-\infty}^{\infty} e^{(x-1)} \cos \frac{\pi}{2}(x - 5) \delta(x - 3) dx \end{array}$$

*Hint:*  $\delta(x)$  is located at  $x = 0$ . For example,  $\delta(1 - t)$  is located at  $1 - t = 0$ ; that is, at  $t = 1$ , and so on.

- 2.4-3** Prove that

$$\delta(at) = \frac{1}{|a|} \delta(t)$$

Hence, show that

$$\delta(\omega) = \frac{1}{2\pi} \delta(f) \quad \text{where } \omega = 2\pi f$$

*Hint:* Show that

$$\int_{-\infty}^{\infty} \phi(t) \delta(at) dt = \frac{1}{|a|} \phi(0)$$

**2.5-1** Derive Eq. (2.26) in an alternate way by observing that  $\mathbf{e} = (\mathbf{g} - c\mathbf{x})$ , and

$$|\mathbf{e}|^2 = (\mathbf{g} - c\mathbf{x}) \cdot (\mathbf{g} - c\mathbf{x}) = |\mathbf{g}|^2 + c^2 |\mathbf{x}|^2 - 2c\mathbf{g} \cdot \mathbf{x}$$

**2.5-2** For the signals  $g(t)$  and  $x(t)$  shown in Fig. P2.5-2, find the component of the form  $x(t)$  contained in  $g(t)$ . In other words, find the optimum value of  $c$  in the approximation  $g(t) \approx cx(t)$  so that the error signal energy is minimum. What is the error signal energy?

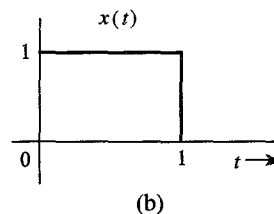
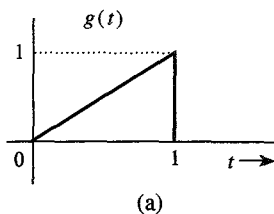


Figure P2.5-2

**2.5-3** For the signals  $g(t)$  and  $x(t)$  shown in Fig. P2.5-2, find the component of the form  $g(t)$  contained in  $x(t)$ . In other words, find the optimum value of  $c$  in the approximation  $x(t) \approx cg(t)$  so that the error signal energy is minimum. What is the error signal energy?

**2.5-4** Repeat Prob. 2.5-2 if  $x(t)$  is the sinusoid pulse shown in Fig. P2.5-4.

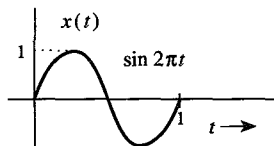


Figure P2.5-4

**2.5-5** Energies of the two energy signals  $x(t)$  and  $y(t)$  are  $E_x$  and  $E_y$ , respectively.

- If  $x(t)$  and  $y(t)$  are orthogonal, then show that the energy of the signal  $x(t) + y(t)$  is identical to the energy of the signal  $x(t) - y(t)$ , and is given by  $E_x + E_y$ .
- If  $x(t)$  and  $y(t)$  are orthogonal, find the energies of signals  $c_1x(t) + c_2y(t)$  and  $c_1x(t) - c_2y(t)$ .
- We define  $E_{xy}$ , the cross energy of the two energy signals  $x(t)$  and  $y(t)$ , as

$$E_{xy} = \int_{-\infty}^{\infty} x(t)y^*(t) dt$$

If  $z(t) = x(t) \pm y(t)$ , then show that

$$E_z = E_x + E_y \pm (E_{xy} + E_{yx})$$

**2.5-6** Let  $x_1(t)$  and  $x_2(t)$  be two unit energy signals orthogonal over an interval from  $t = t_1$  to  $t_2$ . We can represent  $x_1(t)$  and  $x_2(t)$  by two unit length, orthogonal vectors  $(\mathbf{x}_1, \mathbf{x}_2)$ . Consider a signal  $g(t)$  where

$$g(t) = c_1x_1(t) + c_2x_2(t) \quad t_1 \leq t \leq t_2$$

This signal can be represented as a vector  $\mathbf{g}$  by a point  $(c_1, c_2)$  in the  $x_1$ - $x_2$  plane.

(a) Determine the vector representation of the following six signals in this two-dimensional vector space:

- (i)  $g_1(t) = 2x_1(t) - x_2(t)$
- (ii)  $g_2(t) = -x_1(t) + 2x_2(t)$
- (iii)  $g_3(t) = -x_2(t)$
- (iv)  $g_4(t) = x_1(t) + 2x_2(t)$
- (v)  $g_5(t) = 2x_1(t) + x_2(t)$
- (vi)  $g_6(t) = 3x_1(t)$

(b) Point out pairs of mutually orthogonal vectors among these six vectors. Verify that the pairs of signals corresponding to these orthogonal vectors are also orthogonal.

**2.6-1** Find the correlation coefficient  $c_n$  of signal  $x(t)$  and each of the four pulses  $g_1(t)$ ,  $g_2(t)$ ,  $g_3(t)$ , and  $g_4(t)$  shown in Fig. P2.6-1. Which pair of pulses would you select for a binary communication in order to provide maximum margin against the noise along the transmission path?

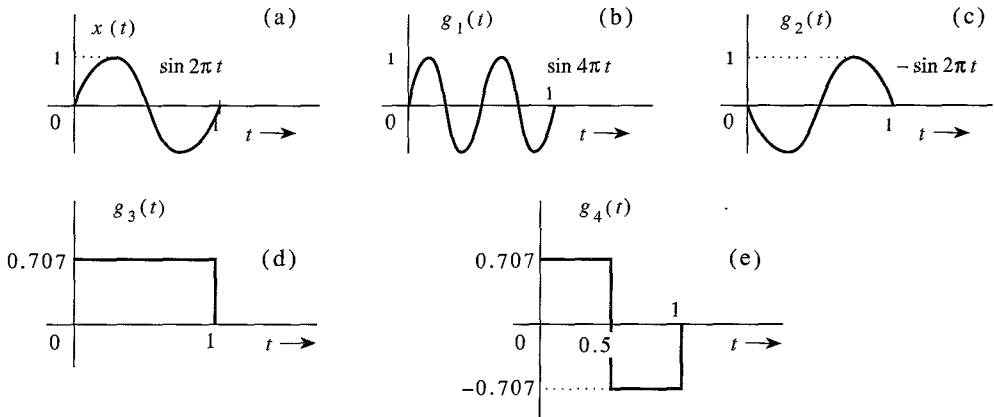


Figure P2.6-1

**2.8-1** (a) Sketch the signal  $g(t) = t^2$  and find the trigonometric Fourier series to represent  $g(t)$  over the interval  $(-1, 1)$ . Sketch the Fourier series  $\varphi(t)$  for all values of  $t$ .

(b) Verify Parseval's theorem [Eq. (2.90)] for this case, given that

$$\sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}$$

2.8-2 (a) Sketch the signal  $g(t) = t$  and find the trigonometric Fourier series to represent  $g(t)$  over the interval  $(-\pi, \pi)$ . Sketch the Fourier series  $\varphi(t)$  for all values of  $t$ .

(b) Verify Parseval's theorem [Eq. (2.90)] for this case, given that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

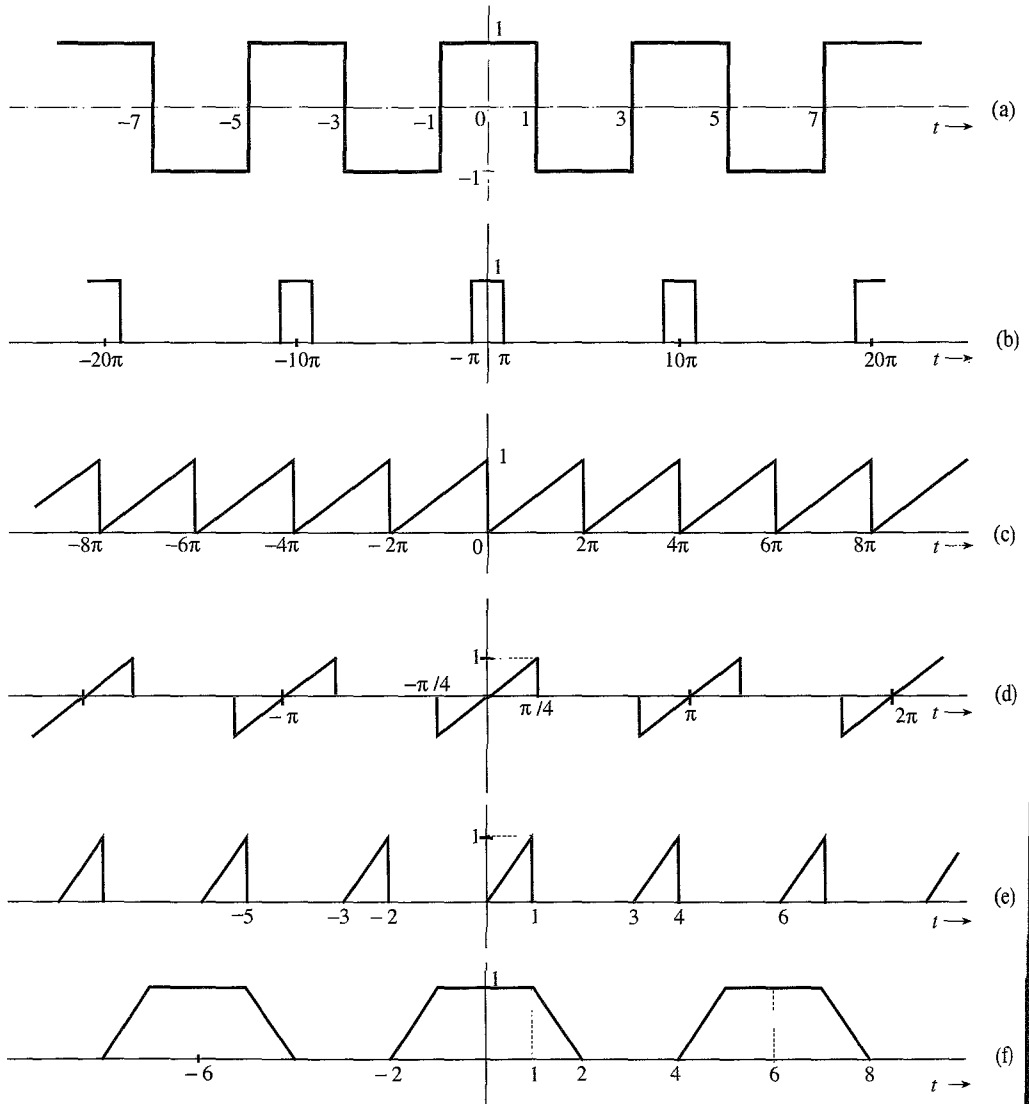


Figure P2.8-4

**2.8-3** If a periodic signal satisfies certain symmetry conditions, the evaluation of the Fourier series components is somewhat simplified. Show that:

- (a) If  $g(t) = g(-t)$  (even symmetry), then all the sine terms in the Fourier series vanish ( $b_n = 0$ ).  
 (b) If  $g(t) = -g(-t)$  (odd symmetry), then the dc and all the cosine terms in the Fourier series vanish ( $a_0 = a_n = 0$ ).

Further, show that in each case the Fourier coefficients can be evaluated by integrating the periodic signal over the half-cycle only. This is because the entire information of one cycle is implicit in a half-cycle due to symmetry. *Hint:* If  $g_e(t)$  and  $g_o(t)$  are even and odd functions, respectively, of  $t$ , then (assuming no impulse or its derivative at the origin)

$$\int_{-a}^a g_e(t) dt = 2 \int_0^a g_e(t) dt \quad \text{and} \quad \int_{-a}^a g_o(t) dt = 0$$

Also the product of an even and an odd function is an odd function, the product of two odd functions is an even function, and the product of two even functions is an even function.

**2.8-4** For each of the periodic signals shown in Fig. P2.8-4, find the compact trigonometric Fourier series and sketch the amplitude and phase spectra. If either the sine or the cosine terms are absent in the Fourier series, explain why.

**2.8-5** (a) Show that an arbitrary function  $g(t)$  can be expressed as a sum of an even function  $g_e(t)$  and an odd function  $g_o(t)$ :

$$g(t) = g_e(t) + g_o(t)$$

*Hint:*

$$g(t) = \underbrace{\frac{1}{2}[g(t) + g(-t)]}_{g_e(t)} + \underbrace{\frac{1}{2}[g(t) - g(-t)]}_{g_o(t)}$$

(b) Determine the odd and even components of the functions: (i)  $u(t)$ ; (ii)  $e^{-at}u(t)$ ; (iii)  $e^{jt}$ .

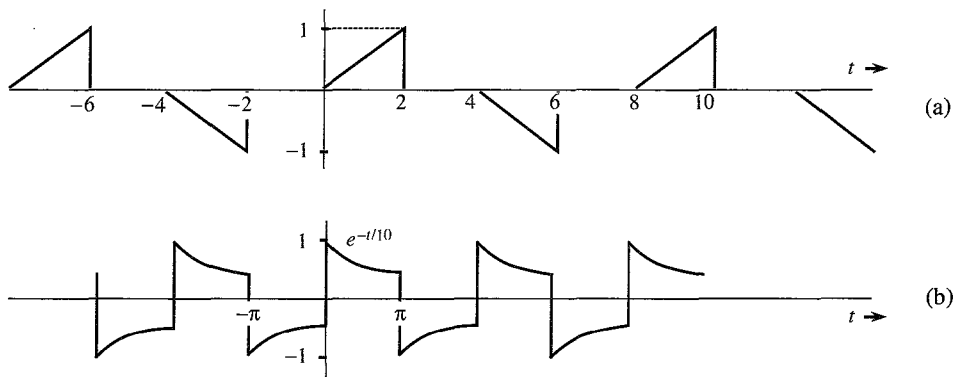


Figure P2.8-6

- 2.8-6** If the two halves of one period of a periodic signal are of identical shape except that one is the negative of the other, the periodic signal is said to have a **half-wave symmetry**. If a periodic signal  $g(t)$  with a period  $T_0$  satisfies the half-wave symmetry condition, then

$$g\left(t - \frac{T_0}{2}\right) = -g(t)$$

In this case, show that all the even-numbered harmonics vanish, and that the odd-numbered harmonic coefficients are given by

$$a_n = \frac{4}{T_0} \int_0^{T_0/2} g(t) \cos n\omega_0 t \, dt \quad \text{and} \quad b_n = \frac{4}{T_0} \int_0^{T_0/2} g(t) \sin n\omega_0 t \, dt$$

Using these results, find the Fourier series for the periodic signals in Fig. P2.8-6.

- 2.9-1** For each of the periodic signals in Fig. P2.8-4, find exponential Fourier series and sketch the corresponding spectra.

- 2.9-2** A periodic signal  $g(t)$  is expressed by the following Fourier series:

$$g(t) = 3 \cos t + \cos\left(5t - \frac{2\pi}{3}\right) + 2 \cos\left(8t + \frac{2\pi}{3}\right)$$

- (a) Sketch the amplitude and phase spectra for the trigonometric series.  
 (b) By inspection of spectra in part (a), sketch the exponential Fourier series spectra.  
 (c) By inspection of spectra in part (b), write the exponential Fourier series for  $g(t)$ .

- 2.9-3** Figure P2.9-3 shows the trigonometric Fourier spectra of a periodic signal  $g(t)$ .

- (a) By inspection of Fig. P2.9-3, find the trigonometric Fourier series representing  $g(t)$ .  
 (b) By inspection of Fig. P2.9-3, sketch the exponential Fourier spectra of  $g(t)$ .  
 (c) By inspection of the exponential Fourier spectra obtained in part (b), find the exponential Fourier series for  $g(t)$ .  
 (d) Show that the series found in parts (a) and (c) are equivalent.

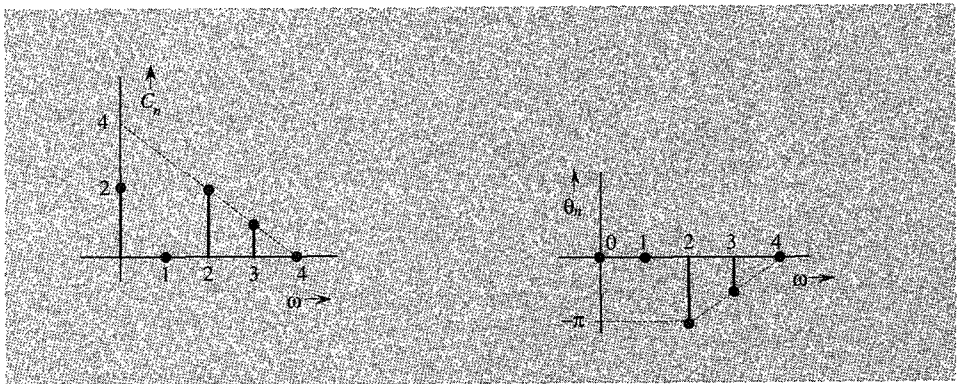
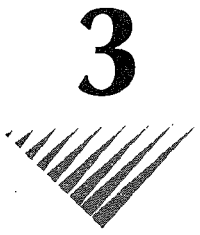


Figure P2.9-3

- 2.9-4** Show that the coefficients of the exponential Fourier series of an even periodic signal are real and those of an odd periodic signal are imaginary.



# 3 ANALYSIS AND TRANSMISSION OF SIGNALS

**E**lectrical engineers instinctively think of signals in terms of their frequency spectra and think of systems in terms of their frequency responses. Even teenagers know about audio signals having a bandwidth of 20 kHz and good-quality loud speakers responding up to 20 kHz. This is basically thinking in the frequency domain. In the last chapter we discussed spectral representation of periodic signals (Fourier series). In this chapter we extend this spectral representation to aperiodic signals.

## 3.1 APERIODIC SIGNAL REPRESENTATION BY FOURIER INTEGRAL

Applying a limiting process, we now show that an aperiodic signal can be expressed as a continuous sum (integral) of everlasting exponentials. To represent an aperiodic signal  $g(t)$ , such as the one shown in Fig. 3.1a by everlasting exponential signals, let us construct a new periodic signal  $g_{T_0}(t)$  formed by repeating the signal  $g(t)$  every  $T_0$  seconds, as shown in Fig. 3.1b. The period  $T_0$  is made long enough to avoid overlap between the repeating pulses. The periodic signal  $g_{T_0}(t)$  can be represented by an exponential Fourier series. If we let  $T_0 \rightarrow \infty$ , the pulses in the periodic signal repeat after an infinite interval, and therefore

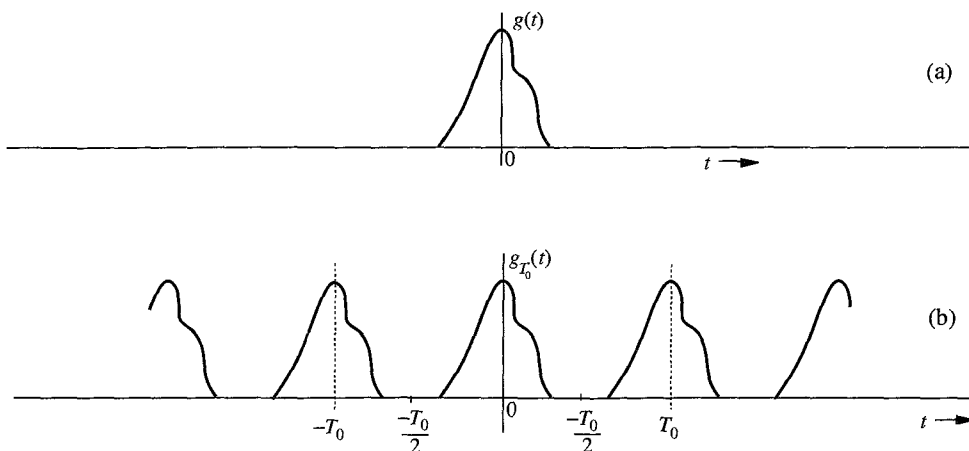
$$\lim_{T_0 \rightarrow \infty} g_{T_0}(t) = g(t)$$

Thus, the Fourier series representing  $g_{T_0}(t)$  will also represent  $g(t)$  in the limit  $T_0 \rightarrow \infty$ . The exponential Fourier series for  $g_{T_0}(t)$  is given by

$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \quad (3.1)$$

in which

$$D_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} g_{T_0}(t) e^{-jn\omega_0 t} dt \quad (3.2a)$$



**Figure 3.1** Construction of a periodic signal by periodic extension of  $g(t)$ .

and

$$\omega_0 = \frac{2\pi}{T_0} \quad (3.2b)$$

Observe that integrating  $g_{T_0}(t)$  over  $(-T_0/2, T_0/2)$  is the same as integrating  $g(t)$  over  $(-\infty, \infty)$ . Therefore, Eq. (3.2a) can be expressed as

$$D_n = \frac{1}{T_0} \int_{-\infty}^{\infty} g(t) e^{-jn\omega_0 t} dt \quad (3.2c)$$

It is interesting to see how the nature of the spectrum changes as  $T_0$  increases. To understand this fascinating behavior, let us define  $G(\omega)$ , a continuous function of  $\omega$ , as

$$G(\omega) = \int_{-\infty}^{\infty} g(t) e^{-j\omega t} dt \quad (3.3)$$

A glance at Eqs. (3.2c) and (3.3) shows that

$$D_n = \frac{1}{T_0} G(n\omega_0) \quad (3.4)$$

This shows that the Fourier coefficients  $D_n$  are  $(1/T_0)$  times the samples of  $G(\omega)$  uniformly spaced at intervals of  $\omega_0$  rad/s, as shown in Fig. 3.2a\*. Therefore,  $(1/T_0)G(\omega)$  is the envelope for the coefficients  $D_n$ . We now let  $T_0 \rightarrow \infty$  by doubling  $T_0$  repeatedly. Doubling  $T_0$  halves the fundamental frequency  $\omega_0$ , so that there are now twice as many components (samples) in the spectrum. However, by doubling  $T_0$ , the envelope  $(1/T_0)G(\omega)$  is halved, as shown in Fig. 3.2b. If we continue this process of doubling  $T_0$  repeatedly, the spectrum progressively becomes denser while its magnitude becomes smaller. Note, however, that the relative shape of the envelope remains the same [proportional to  $G(\omega)$  in Eq. (3.3)]. In the limit as  $T_0 \rightarrow \infty$ ,  $\omega_0 \rightarrow 0$  and  $D_n \rightarrow 0$ . This means that the spectrum is so dense that the spectral components

\* For the sake of simplicity we assume  $D_n$  and therefore  $G(\omega)$  in Fig. 3.2 to be real. The argument, however, is also valid for complex  $D_n$  [or  $G(\omega)$ ].



are spaced at zero (infinitesimal) interval. At the same time, the amplitude of each component is zero (infinitesimal). We have *nothing of everything, yet we have something!* This sounds like *Alice in Wonderland*, but as we shall see, these are the classic characteristics of a very familiar phenomenon.\*

Substitution of Eq. (3.4) in Eq. (3.1) yields

$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} \frac{G(n\omega_0)}{T_0} e^{jn\omega_0 t} \tag{3.5}$$

As  $T_0 \rightarrow \infty$ ,  $\omega_0$  becomes infinitesimal ( $\omega_0 \rightarrow 0$ ). Because of this, we shall replace  $\omega_0$  by a more appropriate notation,  $\Delta\omega$ . In terms of this new notation, Eq. (3.2b) becomes

$$\Delta\omega = \frac{2\pi}{T_0}$$

and Eq. (3.5) becomes

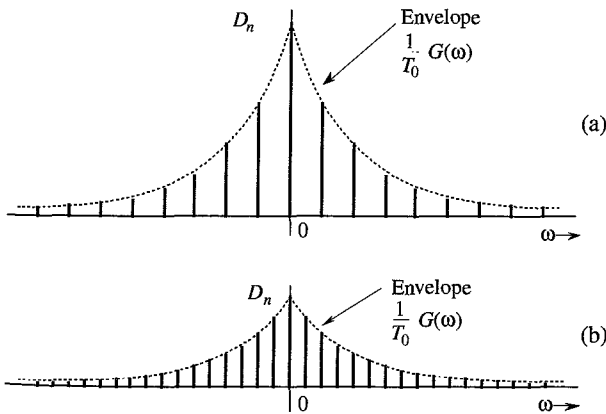
$$g_{T_0}(t) = \sum_{n=-\infty}^{\infty} \left[ \frac{G(n\Delta\omega)\Delta\omega}{2\pi} \right] e^{(jn\Delta\omega)t} \tag{3.6a}$$

Equation (3.6a) shows that  $g_{T_0}(t)$  can be expressed as a sum of everlasting exponentials of frequencies  $0, \pm\Delta\omega, \pm2\Delta\omega, \pm3\Delta\omega, \dots$  (the Fourier series). The amount of the component of frequency  $n\Delta\omega$  is  $[G(n\Delta\omega)\Delta\omega]/2\pi$ . In the limit as  $T_0 \rightarrow \infty$ ,  $\Delta\omega \rightarrow 0$  and  $g_{T_0}(t) \rightarrow g(t)$ . Therefore,

$$g(t) = \lim_{T_0 \rightarrow \infty} g_{T_0}(t) = \lim_{\Delta\omega \rightarrow 0} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} G(n\Delta\omega) e^{(jn\Delta\omega)t} \Delta\omega \tag{3.6b}$$

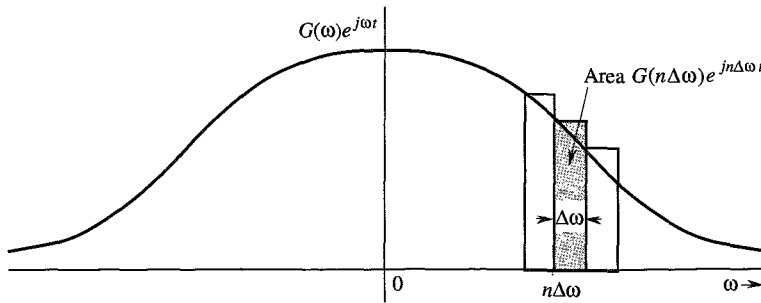
The sum on the right-hand side of Eq. (3.6b) can be viewed as the area under the function  $G(\omega)e^{j\omega t}$ , as shown in Fig. 3.3. Therefore,

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{j\omega t} d\omega \tag{3.7}$$



**Figure 3.2** Change in the Fourier spectrum when the period  $T_0$  in Fig. 3.1 is doubled.

\* You may consider this as an irrefutable proof of the proposition that 0% ownership of everything is better than 100% ownership of nothing!



**Figure 3.3** The Fourier series becomes the Fourier integral in the limit as  $T_0 \rightarrow \infty$ .

The integral on the right-hand side is called the **Fourier integral**. We have now succeeded in representing an aperiodic signal  $g(t)$  by a Fourier integral\* (rather than a Fourier series). This integral is basically a Fourier series (in the limit) with fundamental frequency  $\Delta\omega \rightarrow 0$ , as seen from Eq. (3.6). The amount of the exponential  $e^{jn\Delta\omega t}$  is  $G(n\Delta\omega)\Delta\omega/2\pi$ . Thus, the function  $G(\omega)$  given by Eq. (3.3) acts as a spectral function.

We call  $G(\omega)$  the **direct** Fourier transform of  $g(t)$ , and  $g(t)$  the **inverse** Fourier transform of  $G(\omega)$ . The same information is conveyed by the statement that  $g(t)$  and  $G(\omega)$  are a Fourier transform pair. Symbolically, this is expressed as

$$G(\omega) = \mathcal{F}[g(t)] \quad \text{and} \quad g(t) = \mathcal{F}^{-1}[G(\omega)]$$

or

$$g(t) \iff G(\omega)$$

To recapitulate,

$$G(\omega) = \int_{-\infty}^{\infty} g(t)e^{-j\omega t} dt \quad (3.8a)$$

and

$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega)e^{j\omega t} d\omega \quad (3.8b)$$

It is helpful to keep in mind that the Fourier integral in Eq. (3.8b) is of the nature of a Fourier series with fundamental frequency  $\Delta\omega$  approaching zero [Eq. (3.6b)]. Therefore, most of the discussion and properties of Fourier series apply to the Fourier transform as well. We can plot the spectrum  $G(\omega)$  as a function of  $\omega$ . Since  $G(\omega)$  is complex, we have both amplitude and angle (or phase) spectra:

$$G(\omega) = |G(\omega)|e^{j\theta_g(\omega)}$$

in which  $|G(\omega)|$  is the amplitude and  $\theta_g(\omega)$  is the angle (or phase) of  $G(\omega)$ . From Eq. (3.8a),

$$G(-\omega) = \int_{-\infty}^{\infty} g(t)e^{j\omega t} dt$$

\*This should not be considered as a rigorous proof of Eq. (3.7). The situation is not as simple as we have made it appear.<sup>1</sup>

### Conjugate Symmetry Property

From this equation and Eq. (3.8a), it follows that if  $g(t)$  is a real function of  $t$ , then  $G(\omega)$  and  $G(-\omega)$  are complex conjugates, that is,

$$G(-\omega) = G^*(\omega) \quad (3.9)$$

Therefore,

$$|G(-\omega)| = |G(\omega)| \quad (3.10a)$$

$$\theta_g(-\omega) = -\theta_g(\omega) \quad (3.10b)$$

Thus, for real  $g(t)$ , the amplitude spectrum  $|G(\omega)|$  is an even function, and the phase spectrum  $\theta_g(\omega)$  is an odd function of  $\omega$ . This property (the **conjugate symmetry property**) is valid only for real  $g(t)$ . These results were derived earlier for the Fourier spectrum of a periodic signal [Eqs. (2.85)] and should come as no surprise. *The transform  $G(\omega)$  is the frequency-domain specification of  $g(t)$ .*

**EXAMPLE 3.1** Find the Fourier transform of  $e^{-at}u(t)$ .

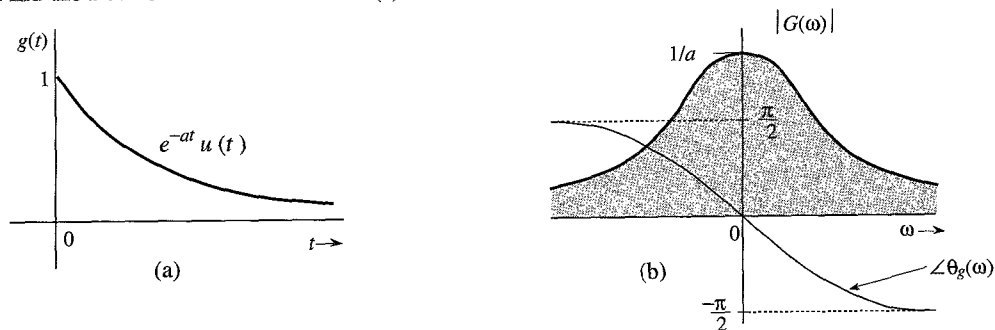


Figure 3.4  $e^{-at}u(t)$  and its Fourier spectra.

By definition [Eq. (3.8a)],

$$G(\omega) = \int_{-\infty}^{\infty} e^{-at}u(t)e^{-j\omega t} dt = \int_0^{\infty} e^{-(a+j\omega)t} dt = \frac{-1}{a+j\omega} e^{-(a+j\omega)t} \Big|_0^{\infty}$$

But  $|e^{-j\omega t}| = 1$ . Therefore, as  $t \rightarrow \infty$ ,  $e^{-(a+j\omega)t} = e^{-at}e^{-j\omega t} = 0$  if  $a > 0$ . Therefore,

$$G(\omega) = \frac{1}{a+j\omega} \quad a > 0 \quad (3.11a)$$

Expressing  $a + j\omega$  in the polar form as  $\sqrt{a^2 + \omega^2} e^{j \tan^{-1}(\frac{\omega}{a})}$ , we obtain

$$G(\omega) = \frac{1}{\sqrt{a^2 + \omega^2}} e^{-j \tan^{-1}(\frac{\omega}{a})} \quad (3.11b)$$

Therefore,

$$|G(\omega)| = \frac{1}{\sqrt{a^2 + \omega^2}} \quad \text{and} \quad \theta_g(\omega) = -\tan^{-1}\left(\frac{\omega}{a}\right)$$

The amplitude spectrum  $|G(\omega)|$  and the phase spectrum  $\theta_g(\omega)$  are shown in Fig. 3.4b. Observe that  $|G(\omega)|$  is an even function of  $\omega$ , and  $\theta_g(\omega)$  is an odd function of  $\omega$ , as expected.

### Existence of the Fourier Transform

In Example 3.1 we observed that when  $a < 0$ , the Fourier integral for  $e^{-at}u(t)$  does not converge. Hence, the Fourier transform for  $e^{-at}u(t)$  does not exist if  $a < 0$  (growing exponential). Clearly, not all signals are Fourier transformable. The existence of the Fourier transform is assured for any  $g(t)$  satisfying the Dirichlet conditions mentioned in Sec. 2.8. The first of these conditions is\*

$$\int_{-\infty}^{\infty} |g(t)| dt < \infty \quad (3.12)$$

To show this, recall that  $|e^{-j\omega t}| = 1$ . Hence, from Eq. (3.8a) we obtain

$$|G(\omega)| \leq \int_{-\infty}^{\infty} |g(t)| dt$$

This shows that the existence of the Fourier transform is assured if condition (3.12) is satisfied. Otherwise, there is no guarantee. We have seen in Example 3.1 that for an exponentially growing signal (which violates this condition) the Fourier transform does not exist. Although this condition is sufficient, it is not necessary for the existence of the Fourier transform of a signal. For example, the signal  $(\sin at)/t$ , violates condition (3.12), but does have a Fourier transform. Any signal that can be generated in practice satisfies the Dirichlet conditions and therefore has a Fourier transform. Thus, the physical existence of a signal is a sufficient condition for the existence of its transform.

### Linearity of the Fourier Transform

The Fourier transform is linear; that is, if

$$g_1(t) \iff G_1(\omega) \quad \text{and} \quad g_2(t) \iff G_2(\omega)$$

then

$$a_1g_1(t) + a_2g_2(t) \iff a_1G_1(\omega) + a_2G_2(\omega) \quad (3.13)$$

The proof is trivial and follows directly from Eq. (3.8a). This result can be extended to any finite number of terms.

## 3.1.1 Physical Appreciation of the Fourier Transform

In understanding any aspect of the Fourier transform, we should remember that Fourier representation is a way of expressing a signal in terms of everlasting sinusoids, or exponentials.

\* The remaining Dirichlet conditions are as follows: In any finite interval,  $g(t)$  may have only a finite number of maxima and minima and a finite number of finite discontinuities. When these conditions are satisfied, the Fourier integral on the right-hand side of Eq. (3.8b) converges to  $g(t)$  at all points where  $g(t)$  is continuous and converges to the average of the right-hand and left-hand limits of  $g(t)$  at points where  $g(t)$  is discontinuous.

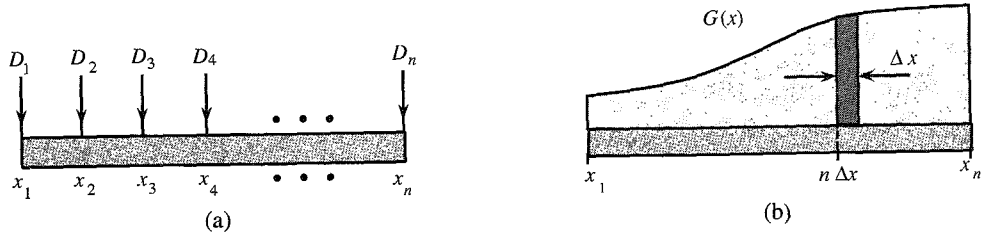


Figure 3.5 Analogy for Fourier transform.

The Fourier spectrum of a signal indicates the relative amplitudes and phases of the sinusoids that are required to synthesize that signal. A periodic signal Fourier spectrum has finite amplitudes and exists at discrete frequencies ( $\omega_0$  and its multiples). Such a spectrum is easy to visualize, but the spectrum of an aperiodic signal is not easy to visualize because it has a continuous spectrum that exists at every frequency. The continuous spectrum concept can be appreciated by considering an analogous, more tangible phenomenon. One familiar example of a continuous distribution is the loading of a beam. Consider a beam loaded with weights  $D_1, D_2, D_3, \dots, D_n$  units at the uniformly spaced points  $x_1, x_2, \dots, x_n$ , as shown in Fig. 3.5a. The total load  $W_T$  on the beam is given by the sum of these loads at each of the  $n$  points:

$$W_T = \sum_{i=1}^n D_i$$

Consider now the case of a continuously loaded beam, as shown in Fig. 3.5b. In this case, although there appears to be a load at every point, the load at any one point is zero. This does not mean that there is no load on the beam. A meaningful measure of load in this situation is not the load at a point, but rather the loading density per unit length at that point. Let  $G(x)$  be the loading density per unit length of beam. This means that the load over a beam length  $\Delta x$  ( $\Delta x \rightarrow 0$ ) at some point  $x$  is  $G(x)\Delta x$ . To find the total load on the beam, we divide the beam into segments of interval  $\Delta x$  ( $\Delta x \rightarrow 0$ ). The load over the  $n$ th such segment of length  $\Delta x$  is  $[G(n\Delta x)] \Delta x$ . The total load  $W_T$  is given by

$$\begin{aligned} W_T &= \lim_{\Delta x \rightarrow 0} \sum_{x_1}^{x_n} G(n\Delta x) \Delta x \\ &= \int_{x_1}^{x_n} G(x) dx \end{aligned}$$

In the case of discrete loading (Fig. 3.5a), the load exists only at the  $n$  discrete points. At other points there is no load. On the other hand, in the continuously loaded case, the load exists at every point, but at any specific point  $x$  the load is zero. The load over a small interval  $\Delta x$ , however, is  $[G(n\Delta x)] \Delta x$  (Fig. 3.5b). Thus, even though the load at a point  $x$  is zero, the relative load at that point is  $G(x)$ .

An exactly analogous situation exists in the case of a signal spectrum. When  $g(t)$  is periodic, the spectrum is discrete, and  $g(t)$  can be expressed as a sum of discrete exponentials with finite amplitudes:

$$g(t) = \sum_n D_n e^{jn\omega_0 t}$$

For an aperiodic signal, the spectrum becomes continuous; that is, the spectrum exists for every value of  $\omega$ , but the amplitude of each component in the spectrum is zero. The meaningful measure here is not the amplitude of a component of some frequency but the spectral density per unit bandwidth. From Eq. (3.6b) it is clear that  $g(t)$  is synthesized by adding exponentials of the form  $e^{jn\Delta\omega t}$ , in which the contribution by any one exponential component is zero. But the contribution by exponentials in an infinitesimal band  $\Delta\omega$  located at  $\omega = n\Delta\omega$  is  $(1/2\pi)G(n\Delta\omega)\Delta\omega$ , and the addition of all these components yields  $g(t)$  in the integral form:

$$g(t) = \lim_{\Delta\omega \rightarrow 0} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} G(n\Delta\omega) e^{(jn\Delta\omega)t} \Delta\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) e^{j\omega t} d\omega$$

The contribution by components within the band  $d\omega$  is  $(1/2\pi)G(\omega) d\omega = G(\omega) df$ , in which  $df$  is the bandwidth in hertz. Clearly  $G(\omega)$  is the **spectral density** per unit bandwidth (in hertz). This also means that even if the amplitude of any one component is zero, the relative amount of a component of frequency  $\omega$  is  $G(\omega)$ . Although  $G(\omega)$  is a spectral density, in practice it is customarily called the **spectrum** of  $g(t)$  rather than the spectral density of  $g(t)$ . Deferring to this convention, we shall call  $G(\omega)$  the Fourier spectrum (or Fourier transform) of  $g(t)$ .

### A Marvelous Balancing Act

An important point to remember here is that  $g(t)$  is represented (or synthesized) by exponentials or sinusoids that are everlasting (not causal). This leads to a rather fascinating picture when we try to visualize the synthesis of a time-limited pulse signal  $g(t)$  (Fig. 3.6) according to the sinusoidal components in its Fourier spectrum. The signal  $g(t)$  exists only over an interval  $(a, b)$  and is zero outside this interval. The spectrum of  $g(t)$  contains an infinite number of exponentials (or sinusoids) which start at  $t = -\infty$  and continue forever. The amplitudes and phases of these components are such that they add up exactly to  $g(t)$  over the finite interval  $(a, b)$  and add up to zero everywhere outside this interval. Juggling with such a perfect and delicate balance of amplitudes and phases of an infinite number of components boggles the human imagination. Yet, the Fourier transform accomplishes it routinely, without much thinking on our part. Indeed, we become so involved in mathematical manipulations that we fail to notice this marvel.

## 3.2 TRANSFORMS OF SOME USEFUL FUNCTIONS

For convenience, we now introduce a compact notation for some useful functions such as gate, triangle, and interpolation functions.

### Unit Gate Function

We define a unit gate function  $\text{rect}(x)$  as a gate pulse of unit height and unit width, centered at the origin, as shown in Fig. 3.7a:

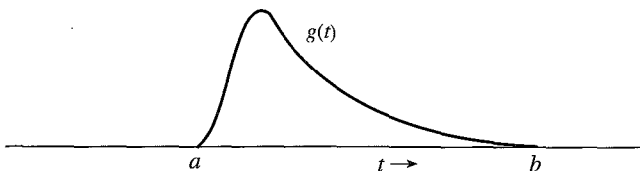


Figure 3.6 A time-limited pulse.

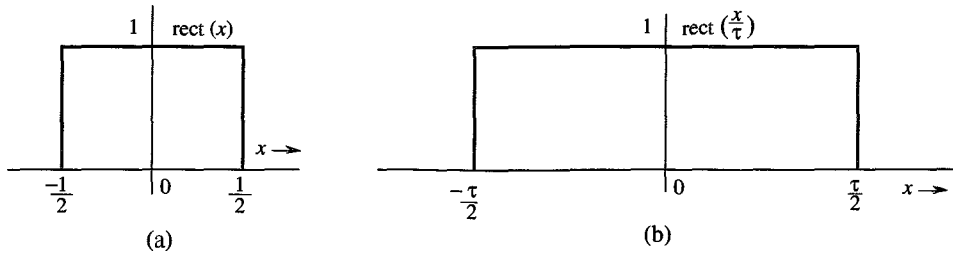


Figure 3.7 Gate pulse.

$$\text{rect}(x) = \begin{cases} 0 & |x| > \frac{1}{2} \\ \frac{1}{2} & |x| = \frac{1}{2} \\ 1 & |x| < \frac{1}{2} \end{cases} \quad (3.14)$$

The gate pulse in Fig. 3.7b is the unit gate pulse  $\text{rect}(x)$  expanded by a factor  $\tau$  and therefore can be expressed as  $\text{rect}(x/\tau)$  (see Sec. 2.3.2). Observe that  $\tau$ , the denominator of the argument of  $\text{rect}(x/\tau)$ , indicates the width of the pulse.

### Unit Triangle Function

We define a unit triangle function  $\Delta(x)$  as a triangular pulse of unit height and unit width, centered at the origin, as shown in Fig. 3.8a:

$$\Delta(x) = \begin{cases} 0 & |x| > \frac{1}{2} \\ 1 - 2|x| & |x| < \frac{1}{2} \end{cases} \quad (3.15)$$

The pulse in Fig. 3.8b is  $\Delta(x/\tau)$ . Observe that here, as for the gate pulse, the denominator  $\tau$  of the argument of  $\Delta(x/\tau)$  indicates the pulse width.

### Interpolation Function $\text{sinc}(x)$

The function  $\sin x/x$  is the “sine over argument” function denoted by  $\text{sinc}(x)$ .\* This function

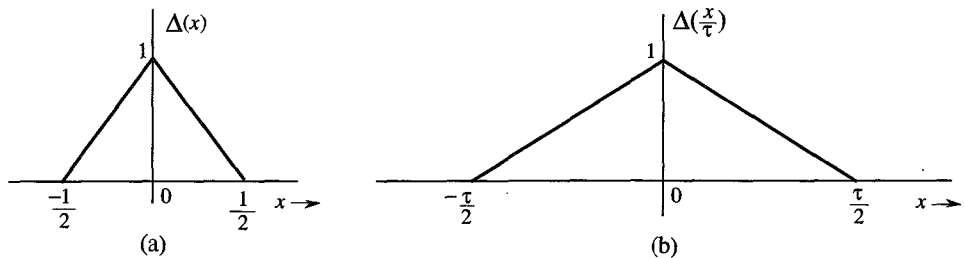


Figure 3.8 Triangle pulse.

\*  $\text{sinc}(x)$  is also denoted by  $\text{Sa}(x)$  in the literature. Some authors define  $\text{sinc}(x)$  as

$$\text{sinc}(x) = \frac{\sin \pi x}{\pi x}$$

plays an important role in signal processing. It is also known as the **filtering** or **interpolating function**. We define

$$\text{sinc}(x) = \frac{\sin x}{x} \quad (3.16)$$

Inspection of Eq. (3.16) shows that

1.  $\text{sinc}(x)$  is an even function of  $x$ .
2.  $\text{sinc}(x) = 0$  when  $\sin x = 0$  except at  $x = 0$ , where it is indeterminate. This means that  $\text{sinc}(x) = 0$  for  $x = \pm\pi, \pm 2\pi, \pm 3\pi, \dots$ .
3. Using L'Hôpital's rule, we find  $\text{sinc}(0) = 1$ .
4.  $\text{sinc}(x)$  is the product of an oscillating signal  $\sin x$  (of period  $2\pi$ ) and a monotonically decreasing function  $1/x$ . Therefore,  $\text{sinc}(x)$  exhibits sinusoidal oscillations of period  $2\pi$ , with amplitude decreasing continuously as  $1/x$ .

Figure 3.9a shows  $\text{sinc}(x)$ . Observe that  $\text{sinc}(x) = 0$  for values of  $x$  that are positive and negative integral multiples of  $\pi$ . Figure 3.9b shows  $\text{sinc}(3\omega/7)$ . The argument  $3\omega/7 = \pi$  when  $\omega = 7\pi/3$ . Therefore, the first zero of this function occurs at  $\omega = 7\pi/3$ .

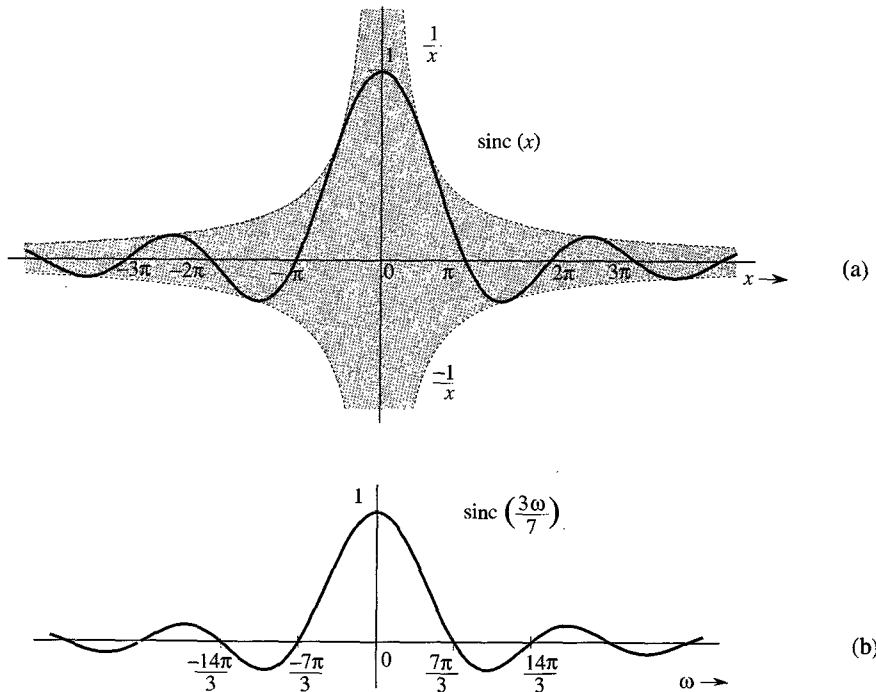


Figure 3.9 Sinc pulse.

**EXAMPLE 3.2** Find the Fourier transform of  $g(t) = \text{rect}(t/\tau)$  (Fig. 3.10a).



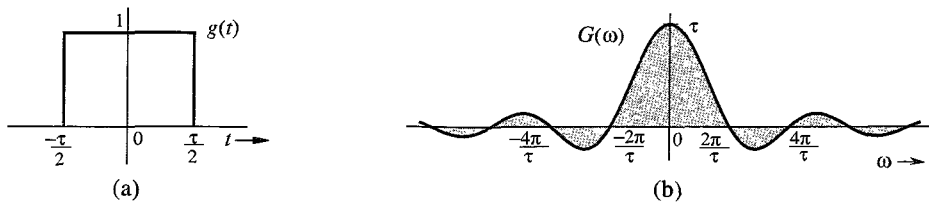


Figure 3.10 Gate pulse and its Fourier spectrum.

We have

$$G(\omega) = \int_{-\infty}^{\infty} \text{rect} \left( \frac{t}{\tau} \right) e^{-j\omega t} dt$$

Since  $\text{rect}(t/\tau) = 1$  for  $|t| < \tau/2$ , and since it is zero for  $|t| > \tau/2$ ,

$$\begin{aligned} G(\omega) &= \int_{-\tau/2}^{\tau/2} e^{-j\omega t} dt \\ &= -\frac{1}{j\omega} (e^{-j\omega\tau/2} - e^{j\omega\tau/2}) = \frac{2 \sin(\omega\tau/2)}{\omega} \\ &= \tau \frac{\sin(\omega\tau/2)}{(\omega\tau/2)} = \tau \text{sinc} \left( \frac{\omega\tau}{2} \right) \end{aligned}$$

Therefore,

$$\text{rect} \left( \frac{t}{\tau} \right) \iff \tau \text{sinc} \left( \frac{\omega\tau}{2} \right) \quad (3.17)$$

Recall that  $\text{sinc}(x) = 0$  when  $x = \pm n\pi$ . Hence,  $\text{sinc}(\omega\tau/2) = 0$  when  $\omega\tau/2 = \pm n\pi$ ; that is, when  $\omega = \pm 2n\pi/\tau$  ( $n = 1, 2, 3, \dots$ ), as shown in Fig. 3.10b. Observe that in this case  $G(\omega)$  happens to be real. Hence, we may convey the spectral information by a single plot of  $G(\omega)$  shown in Fig. 3.10b.

### Bandwidth of $\text{rect} \left( \frac{t}{\tau} \right)$

The spectrum  $G(\omega)$  in Fig. 3.10 peaks at  $\omega = 0$  and decays at higher frequencies. Therefore,  $\text{rect}(t/\tau)$  is a low-pass signal with most of the signal energy in lower frequency components. **Signal bandwidth** is the difference between the highest (significant) frequency and the lowest (significant) frequency in the signal spectrum. Strictly speaking, because the spectrum extends from 0 to  $\infty$ , the bandwidth is  $\infty$  in the present case. However, much of the spectrum is concentrated within the first lobe (from  $\omega = 0$  to  $\omega = 2\pi/\tau$ ), and we may consider  $\omega = 2\pi/\tau$  to be the highest (significant) frequency in the spectrum. Therefore, a rough estimate of the bandwidth\* of a rectangular pulse of width  $\tau$  seconds is  $2\pi/\tau$  rad/s, or  $1/\tau$  Hz. Note the reciprocal relationship of the pulse width with its bandwidth. We shall observe later that this result is true in general.

\* To compute the bandwidth, we must consider the spectrum only for positive values of  $\omega$ . The trigonometric spectrum exists only for positive frequencies. The negative frequencies occur because we use exponential spectra for mathematical convenience. Each sinusoid  $\cos \omega_n t$  appears as a sum of two exponential components  $e^{j\omega_n t}$  and  $e^{-j\omega_n t}$  with frequencies  $\omega_n$  and  $-\omega_n$ , respectively. But in reality, there is only one component of frequency  $\omega_n$ .

**EXAMPLE 3.3** Find the Fourier transform of the unit impulse  $\delta(t)$ .

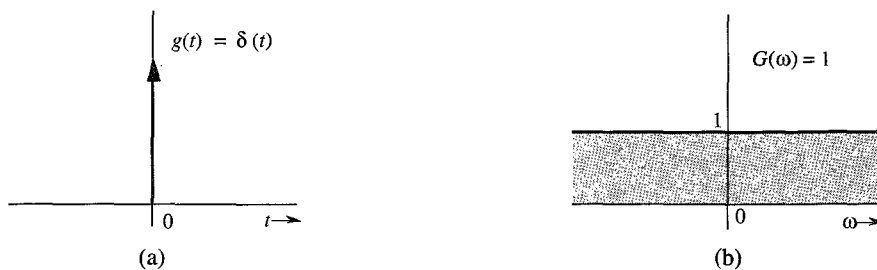
Using the sampling property of the impulse [Eq. (2.19a)], we obtain

$$\mathcal{F}[\delta(t)] = \int_{-\infty}^{\infty} \delta(t) e^{-j\omega t} dt = 1 \quad (3.18a)$$

or

$$\delta(t) \iff 1 \quad (3.18b)$$

Figure 3.11 shows  $\delta(t)$  and its spectrum.



**Figure 3.11** Unit impulse and its Fourier spectrum.

**EXAMPLE 3.4** Find the inverse Fourier transform of  $\delta(\omega)$ .

From Eq. (3.8b) and the sampling property of the impulse function,

$$\mathcal{F}^{-1}[\delta(\omega)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \delta(\omega) e^{j\omega t} d\omega = \frac{1}{2\pi}$$

Therefore,

$$\frac{1}{2\pi} \iff \delta(\omega) \quad (3.19a)$$

or

$$1 \iff 2\pi\delta(\omega) \quad (3.19b)$$

This shows that the spectrum of a constant signal  $g(t) = 1$  is an impulse  $2\pi\delta(\omega)$ , as shown in Fig. 3.12.

The result [Eq. (3.19b)] also could have been anticipated on qualitative grounds. Recall that the Fourier transform of  $g(t)$  is a spectral representation of  $g(t)$  in terms of everlasting exponential components of the form  $e^{j\omega t}$ . Now to represent a constant signal  $g(t) = 1$ , we need a single everlasting exponential\*  $e^{j\omega t}$  with  $\omega = 0$ . This results in a

\* The constant multiplier  $2\pi$  in the spectrum [ $G(\omega) = 2\pi\delta(\omega)$ ] may be a bit puzzling. Since  $1 = e^{j\omega t}$  with  $\omega = 0$ , it appears that the Fourier transform of  $g(t) = 1$  should be an impulse of strength unity rather than  $2\pi$ . Recall, however, that in the Fourier transform  $g(t)$  is synthesized not by exponentials of amplitude  $G(n\Delta\omega)\Delta\omega$ , but of amplitude  $1/2\pi$  times  $G(n\Delta\omega)\Delta\omega$ , as seen from Eq. (3.6b). Had we used variable  $f$  (in hertz) instead of  $\omega$ , the spectrum would have been a unit impulse.

spectrum at a single frequency  $\omega = 0$ . Another way of looking at the situation is that  $g(t) = 1$  is a dc signal which has a single frequency  $\omega = 0$  (dc).

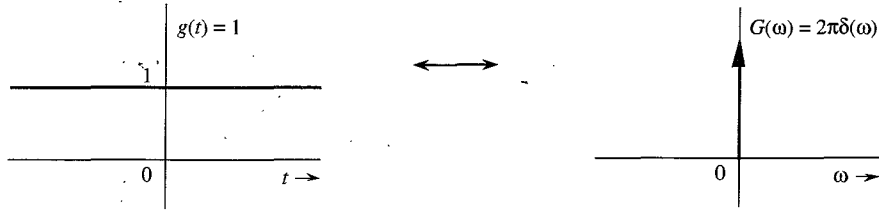


Figure 3.12 Constant (dc) signal and its Fourier spectrum.

If an impulse at  $\omega = 0$  is a spectrum of a dc signal, what does an impulse at  $\omega = \omega_0$  represent? We shall answer this question in the next example.

**EXAMPLE 3.5** Find the inverse Fourier transform of  $\delta(\omega - \omega_0)$ .

Using the sampling property of the impulse function, we obtain

$$\mathcal{F}^{-1}[\delta(\omega - \omega_0)] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \delta(\omega - \omega_0) e^{j\omega t} d\omega = \frac{1}{2\pi} e^{j\omega_0 t}$$

Therefore,

$$\frac{1}{2\pi} e^{j\omega_0 t} \iff \delta(\omega - \omega_0)$$

or

$$e^{j\omega_0 t} \iff 2\pi \delta(\omega - \omega_0) \quad (3.20a)$$

This result shows that the spectrum of an everlasting exponential  $e^{j\omega_0 t}$  is a single impulse at  $\omega = \omega_0$ . We reach the same conclusion by qualitative reasoning. To represent the everlasting exponential  $e^{j\omega_0 t}$ , we need a single everlasting exponential  $e^{j\omega t}$  with  $\omega = \omega_0$ . Therefore, the spectrum consists of a single component at frequency  $\omega = \omega_0$ .

From Eq. (3.20a) it follows that

$$e^{-j\omega_0 t} \iff 2\pi \delta(\omega + \omega_0) \quad (3.20b)$$

**EXAMPLE 3.6** Find the Fourier transforms of the everlasting sinusoid  $\cos \omega_0 t$ .

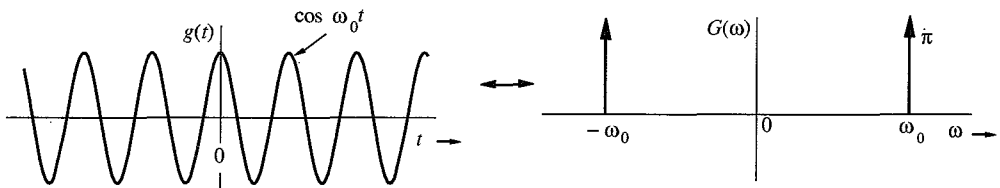


Figure 3.13 Cosine signal and its Fourier spectrum.

Recall the Euler formula

$$\cos \omega_0 t = \frac{1}{2} (e^{j\omega_0 t} + e^{-j\omega_0 t})$$

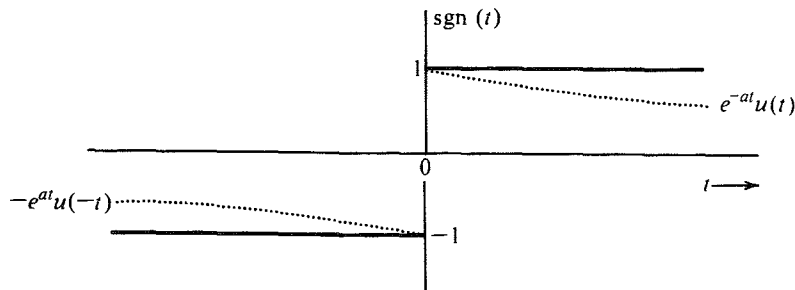
Adding Eqs. (3.20a) and (3.20b), and using the above formula, we obtain

$$\cos \omega_0 t \iff \pi[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)] \quad (3.21)$$

The spectrum of  $\cos \omega_0 t$  consists of two impulses at  $\omega_0$  and  $-\omega_0$ , as shown in Fig. 3.13. The result also follows from qualitative reasoning. An everlasting sinusoid  $\cos \omega_0 t$  can be synthesized by two everlasting exponentials,  $e^{j\omega_0 t}$  and  $e^{-j\omega_0 t}$ . Therefore, the Fourier spectrum consists of only two components of frequencies  $\omega_0$  and  $-\omega_0$ .

**EXAMPLE 3.7** Find the Fourier transform of the sign function  $\text{sgn } t$  (pronounced signum  $t$ ), shown in Fig. 3.14. Its value is  $+1$  or  $-1$ , depending on whether  $t$  is positive or negative:

$$\text{sgn } t = \begin{cases} 1 & t > 0 \\ -1 & t < 0 \end{cases} \quad (3.22)$$



**Figure 3.14** Sign function.

The transform of  $\text{sgn } t$  can be obtained by considering  $\text{sgn } t$  as a sum of two exponentials, as shown in Fig. 3.14, in the limit as  $a \rightarrow 0$ :

$$\text{sgn } t = \lim_{a \rightarrow 0} [e^{-at}u(t) - e^{at}u(-t)]$$

Therefore,

$$\begin{aligned} \mathcal{F}[\text{sgn } t] &= \lim_{a \rightarrow 0} \{ \mathcal{F}[e^{-at}u(t)] - \mathcal{F}[e^{at}u(-t)] \} \\ &= \lim_{a \rightarrow 0} \left( \frac{1}{a + j\omega} - \frac{1}{a - j\omega} \right) \quad (\text{see pairs 1 and 2 in Table 3.1}) \\ &= \lim_{a \rightarrow 0} \left( \frac{-2j\omega}{a^2 + \omega^2} \right) = \frac{2}{j\omega} \end{aligned} \quad (3.23)$$

### 3.3 SOME PROPERTIES OF THE FOURIER TRANSFORM

We now study some of the important properties of the Fourier transform and their implications as well as their applications. Before embarking on this study, it is important to point out a pervasive aspect of the Fourier transform—the **time-frequency duality**.

Table 3.1

Short Table of Fourier Transforms

	$g(t)$	$G(\omega)$	
1	$e^{-at}u(t)$	$\frac{1}{a + j\omega}$	$a > 0$
2	$e^{at}u(-t)$	$\frac{1}{a - j\omega}$	$a > 0$
3	$e^{-a t }$	$\frac{2a}{a^2 + \omega^2}$	$a > 0$
4	$te^{-at}u(t)$	$\frac{1}{(a + j\omega)^2}$	$a > 0$
5	$t^n e^{-at}u(t)$	$\frac{n!}{(a + j\omega)^{n+1}}$	$a > 0$
6	$\delta(t)$	1	
7	1	$2\pi\delta(\omega)$	
8	$e^{j\omega_0 t}$	$2\pi\delta(\omega - \omega_0)$	
9	$\cos \omega_0 t$	$\pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]$	
10	$\sin \omega_0 t$	$j\pi[\delta(\omega + \omega_0) - \delta(\omega - \omega_0)]$	
11	$u(t)$	$\pi\delta(\omega) + \frac{1}{j\omega}$	
12	$\text{sgn } t$	$\frac{2}{j\omega}$	
13	$\cos \omega_0 t u(t)$	$\frac{\pi}{2}[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)] + \frac{j\omega}{\omega_0^2 - \omega^2}$	
14	$\sin \omega_0 t u(t)$	$\frac{\pi}{2j}[\delta(\omega - \omega_0) - \delta(\omega + \omega_0)] + \frac{\omega_0}{\omega_0^2 - \omega^2}$	
15	$e^{-at} \sin \omega_0 t u(t)$	$\frac{\omega_0}{(a + j\omega)^2 + \omega_0^2}$	$a > 0$
16	$e^{-at} \cos \omega_0 t u(t)$	$\frac{a + j\omega}{(a + j\omega)^2 + \omega_0^2}$	$a > 0$
17	$\text{rect}\left(\frac{t}{\tau}\right)$	$\tau \text{sinc}\left(\frac{\omega\tau}{2}\right)$	
18	$\frac{W}{\pi} \text{sinc}(Wt)$	$\text{rect}\left(\frac{\omega}{2W}\right)$	
19	$\Delta\left(\frac{t}{\tau}\right)$	$\frac{\tau}{2} \text{sinc}^2\left(\frac{\omega\tau}{4}\right)$	
20	$\frac{W}{2\pi} \text{sinc}^2\left(\frac{Wt}{2}\right)$	$\Delta\left(\frac{\omega}{2W}\right)$	
21	$\sum_{n=-\infty}^{\infty} \delta(t - nT)$	$\omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)$	$\omega_0 = \frac{2\pi}{T}$
22	$e^{-t^2/2\sigma^2}$	$\sigma\sqrt{2\pi}e^{-\sigma^2\omega^2/2}$	

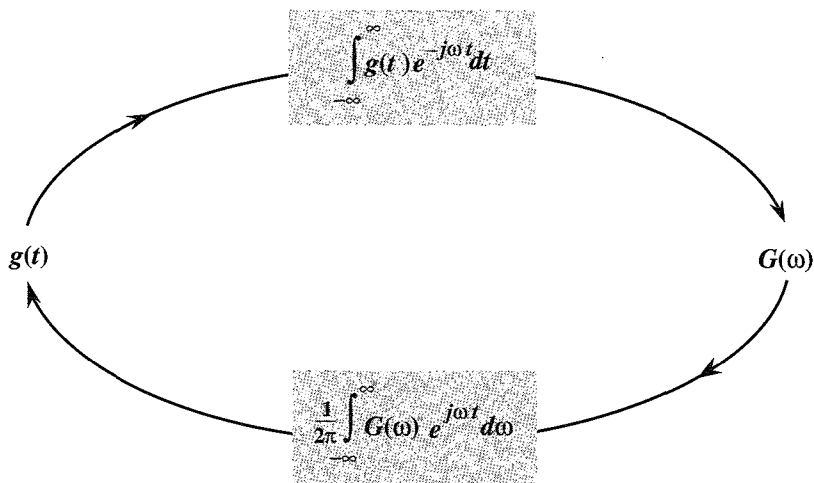


Figure 3.15 Near symmetry between direct and inverse Fourier transforms.

### 3.3.1 Symmetry of Direct and Inverse Transform Operations—Time-Frequency Duality

Equations (3.8) show an interesting fact: the direct and the inverse transform operations are remarkably similar. These operations, required to go from  $g(t)$  to  $G(\omega)$  and then from  $G(\omega)$  to  $g(t)$ , are shown graphically in Fig. 3.15. There are only two minor differences in these operations: the factor  $2\pi$  appears only in the inverse operator, and the exponential indices in the two operations have opposite signs. Otherwise the two operations are symmetrical.\* This observation has far-reaching consequences in the study of Fourier transform. It is the basis of the so-called duality of time and frequency. *The duality principle may be compared with a photograph and its negative. A photograph can be obtained from its negative, and by using an identical procedure, the negative can be obtained from the photograph.* For any result or relationship between  $g(t)$  and  $G(\omega)$ , there exists a dual result or relationship, obtained by interchanging the roles of  $g(t)$  and  $G(\omega)$  in the original result (along with some minor modifications arising because of the factor  $2\pi$  and a sign change). For example, the time-shifting property, to be proved later, states that if  $g(t) \iff G(\omega)$ , then

$$g(t - t_0) \iff G(\omega)e^{-j\omega t_0}$$

The dual of this property (the frequency-shifting property) states that

\* Of the two differences, the former can be eliminated by a change of variable from  $\omega$  to  $f$  (in hertz). In this case,

$$\omega = 2\pi f \quad \text{and} \quad d\omega = 2\pi df$$

Therefore, the direct and the inverse transforms are given by

$$G(2\pi f) = \int_{-\infty}^{\infty} g(t) e^{-j2\pi f t} dt \quad \text{and} \quad g(t) = \int_{-\infty}^{\infty} G(2\pi f) e^{j2\pi f t} df$$

This leaves only one significant difference, that of sign change in the exponential index. Otherwise the two operations are symmetrical.

$$g(t)e^{j\omega_0 t} \iff G(\omega - \omega_0)$$

Observe the role reversal of time and frequency in these two equations (with the minor difference of the sign change in the exponential index). The value of this principle lies in the fact that *whenever we derive any result, we can be sure that it has a dual*. This can give valuable insights about many unsuspected properties or results in signal processing.

The properties of the Fourier transform are useful not only in deriving the direct and the inverse transforms of many functions, but also in obtaining several valuable results in signal processing. The reader should not fail to observe the ever-present duality in this discussion. We begin with the symmetry property, which is one of the consequences of the duality principle discussed.

### 3.3.2 Symmetry Property

This property states that if

$$g(t) \iff G(\omega)$$

then

$$G(t) \iff 2\pi g(-\omega) \quad (3.24)$$

*Proof:* From Eq. (3.8b),

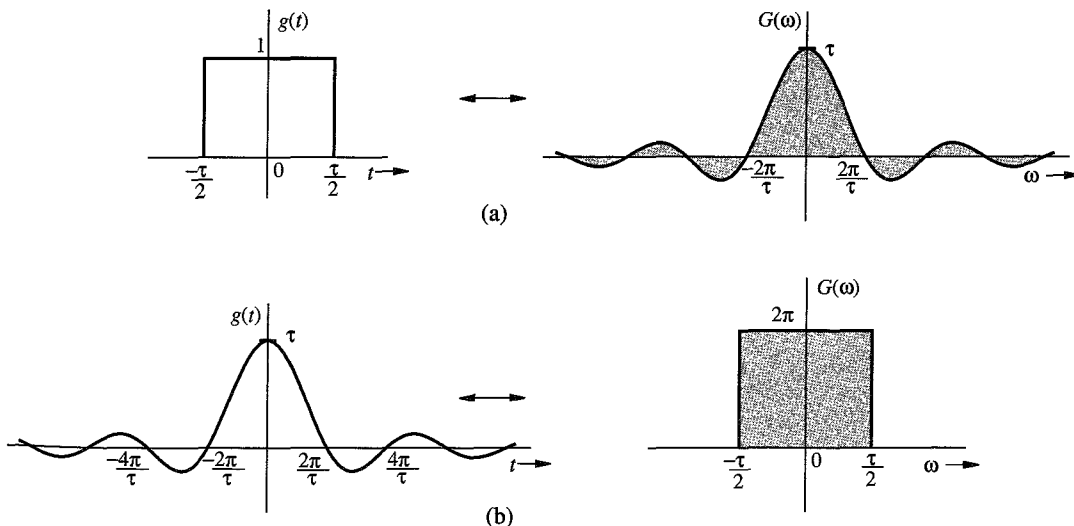
$$g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(x)e^{jxt} dx$$

Hence,

$$2\pi g(-t) = \int_{-\infty}^{\infty} G(x)e^{-jxt} dx$$

Changing  $t$  to  $\omega$  yields Eq. (3.24).

**EXAMPLE 3.8** In this example we shall apply the symmetry property [Eq. (3.24)] to the pair in Fig. 3.16a.



**Figure 3.16** Duality property of the Fourier transform.

From Eq. (3.17) we have

$$\underbrace{\text{rect}\left(\frac{t}{\tau}\right)}_{g(t)} \iff \underbrace{\tau \text{sinc}\left(\frac{\omega\tau}{2}\right)}_{G(\omega)} \quad (3.25)$$

Also  $G(t)$  is the same as  $G(\omega)$  with  $\omega$  replaced by  $t$ , and  $g(-\omega)$  is the same as  $g(t)$  with  $t$  replaced by  $-\omega$ . Therefore, the symmetry property (3.24) yields

$$\underbrace{\tau \text{sinc}\left(\frac{\tau t}{2}\right)}_{G(t)} \iff \underbrace{2\pi \text{rect}\left(\frac{-\omega}{\tau}\right)}_{2\pi g(-\omega)} = 2\pi \text{rect}\left(\frac{\omega}{\tau}\right) \quad (3.26)$$

In Eq. (3.26) we used the fact that  $\text{rect}(-x) = \text{rect}(x)$  because  $\text{rect}$  is an even function. Figure 3.16b shows this pair graphically. Observe the interchange of the roles of  $t$  and  $\omega$  (with the minor adjustment of the factor  $2\pi$ ). This result appears as pair 18 in Table 3.1 (with  $\tau/2 = W$ ).

As an interesting exercise, by applying the symmetry property, the reader should generate a dual of every pair in Table 3.1.

### 3.3.3 Scaling Property

If

$$g(t) \iff G(\omega)$$

then, for any real constant  $a$ ,

$$g(at) \iff \frac{1}{|a|} G\left(\frac{\omega}{a}\right) \quad (3.27)$$

*Proof:* For a positive real constant  $a$ ,

$$\mathcal{F}[g(at)] = \int_{-\infty}^{\infty} g(at)e^{-j\omega t} dt = \frac{1}{a} \int_{-\infty}^{\infty} g(x)e^{(-j\omega/a)x} dx = \frac{1}{a} G\left(\frac{\omega}{a}\right)$$

Similarly, it can be shown that if  $a < 0$ ,

$$g(at) \iff \frac{-1}{a} G\left(\frac{\omega}{a}\right)$$

Hence follows Eq. (3.27).



### Significance of the Scaling Property

The function  $g(at)$  represents the function  $g(t)$  compressed in time by a factor  $a$  (see Sec. 2.3.2). Similarly, a function  $G(\omega/a)$  represents the function  $G(\omega)$  expanded in frequency by the same factor  $a$ . *The scaling property states that time compression of a signal results in its spectral expansion, and time expansion of the signal results in its spectral compression.* Intuitively compression in time by a factor  $a$  means that the signal is varying rapidly by the same factor. To synthesize such a signal, the frequencies of its sinusoidal components must be increased by the factor  $a$ , implying that its frequency spectrum is expanded by the factor  $a$ . Similarly, a signal expanded in time varies more slowly; hence, the frequencies of its components are lowered, implying that its frequency spectrum is compressed. For instance, the signal  $\cos 2\omega_0 t$  is the same as the signal  $\cos \omega_0 t$  time-compressed by a factor of 2. Clearly, the spectrum of the former (impulse at  $\pm 2\omega_0$ ) is an expanded version of the spectrum of the latter (impulse at  $\pm\omega_0$ ). The effect of this scaling is demonstrated in Fig. 3.17.

### Reciprocity of Signal Duration and Its Bandwidth

The scaling property implies that if  $g(t)$  is wider, its spectrum is narrower, and vice versa. Doubling the signal duration halves its bandwidth, and vice versa. This suggests that the bandwidth of a signal is inversely proportional to the signal duration or width (in seconds). We have already verified this fact for the gate pulse, where we found that the bandwidth of a gate pulse of width  $\tau$  seconds is  $1/\tau$  Hz. More discussion of this interesting topic can be found in the literature.<sup>2</sup>

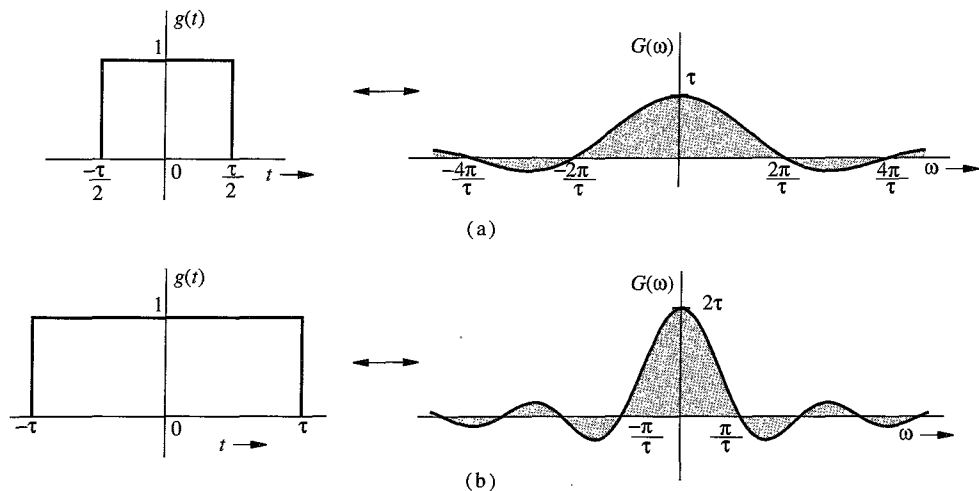


Figure 3.17 Scaling property of the Fourier transform.

**EXAMPLE 3.9** Show that

$$g(-t) \iff G(-\omega) \quad (3.28)$$

Using this result and the fact that  $e^{-at}u(t) \iff 1/a + j\omega$ , find the Fourier transforms of  $e^{at}u(-t)$  and  $e^{-a|t|}$ .

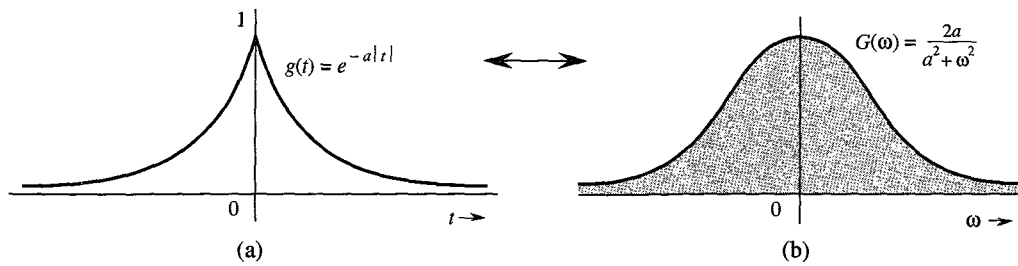


Figure 3.18  $e^{-a|t|}$  and its Fourier spectrum.

Equation (3.28) follows from Eq. (3.27) by letting  $a = -1$ . Application of Eq. (3.28) to pair 1 of Table 3.1 yields

$$e^{at}u(-t) \iff \frac{1}{a - j\omega}$$

Also

$$e^{-a|t|} = e^{-at}u(t) + e^{at}u(-t)$$

Therefore,

$$e^{-a|t|} \iff \frac{1}{a + j\omega} + \frac{1}{a - j\omega} = \frac{2a}{a^2 + \omega^2} \quad (3.29)$$

The signal  $e^{-a|t|}$  and its spectrum are shown in Fig. 3.18.

### 3.3.4 Time-Shifting Property

If

$$g(t) \iff G(\omega)$$

then

$$g(t - t_0) \iff G(\omega)e^{-j\omega t_0} \quad (3.30a)$$

*Proof:* By definition,

$$\mathcal{F}[g(t - t_0)] = \int_{-\infty}^{\infty} g(t - t_0)e^{-j\omega t} dt$$

Letting  $t - t_0 = x$ , we have

$$\begin{aligned} \mathcal{F}[g(t - t_0)] &= \int_{-\infty}^{\infty} g(x)e^{-j\omega(x + t_0)} dx \\ &= e^{-j\omega t_0} \int_{-\infty}^{\infty} g(x)e^{-j\omega x} dx = G(\omega)e^{-j\omega t_0} \end{aligned} \quad (3.30b)$$

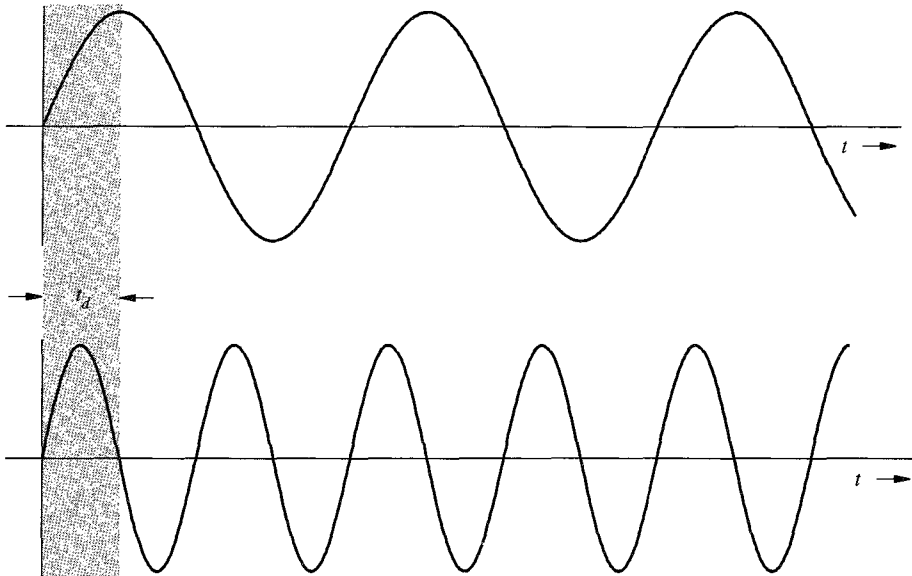
This result shows that *delaying a signal by  $t_0$  seconds does not change its amplitude spectrum. The phase spectrum, however, is changed by  $-\omega t_0$ .*

### Physical Explanation of the Linear Phase

Time delay in a signal causes a linear phase shift in its spectrum. This result can also be derived by heuristic reasoning. Imagine  $g(t)$  being synthesized by its Fourier components, which are sinusoids of certain amplitudes and phases. The delayed signal  $g(t - t_0)$  can be synthesized by the same sinusoidal components, each delayed by  $t_0$  seconds. The amplitudes of the components remain unchanged. Therefore, the amplitude spectrum of  $g(t - t_0)$  is identical to that of  $g(t)$ . The time delay of  $t_0$  in each sinusoid, however, does change the phase of each component. Now, a sinusoid  $\cos \omega t$  delayed by  $t_0$  is given by

$$\cos \omega(t - t_0) = \cos(\omega t - \omega t_0)$$

Therefore, a time delay  $t_0$  in a sinusoid of frequency  $\omega$  manifests as a phase delay of  $\omega t_0$ . This is a linear function of  $\omega$ , meaning that higher frequency components must undergo proportionately higher phase shifts to achieve the same time delay. This effect is shown in Fig. 3.19 with two sinusoids, the frequency of the lower sinusoid being twice that of the upper. The same time delay  $t_0$  amounts to a phase shift of  $\pi/2$  in the upper sinusoid and a phase shift of  $\pi$  in the lower sinusoid. This verifies the fact that *to achieve the same time delay, higher frequency sinusoids must undergo proportionately higher phase shifts.*



**Figure 3.19** Physical explanation of the time-shifting property.

---

**EXAMPLE 3.10** Find the Fourier transform of  $e^{-a|t-t_0|}$ .

■ This function, shown in Fig. 3.20a, is a time-shifted version of  $e^{-a|t|}$  (shown in Fig. 3.18a). From Eqs. (3.29) and (3.30) we have

$$e^{-a|t-t_0|} \iff \frac{2a}{a^2 + \omega^2} e^{-j\omega t_0} \tag{3.31}$$

The spectrum of  $e^{-a|t-t_0|}$  (Fig. 3.20b) is the same as that of  $e^{-a|t|}$  (Fig. 3.18b), except for an added phase shift of  $- \omega t_0$ .

Observe that the time delay  $t_0$  causes a linear phase spectrum  $- \omega t_0$ . This example clearly demonstrates the effect of time shift.

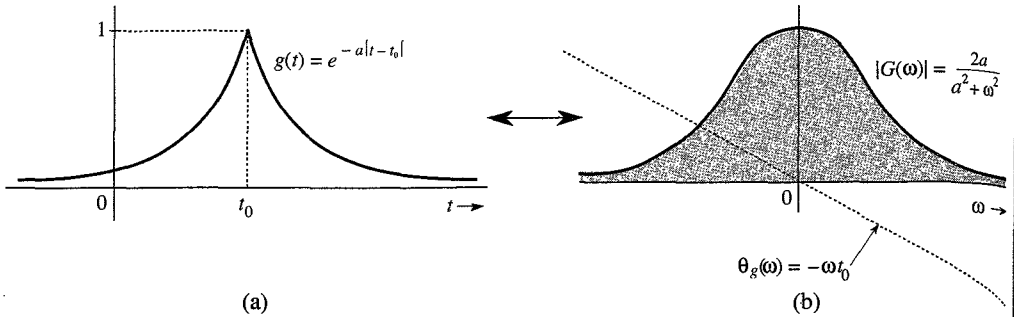


Figure 3.20 Effect of time shifting on the Fourier spectrum of a signal.

**EXAMPLE 3.11** Show that

$$g(t - T) + g(t + T) \iff 2G(\omega) \cos T\omega \tag{3.32}$$

■ This follows directly from Eqs. (3.30).

### 3.3.5 Frequency-Shifting Property

If

$$g(t) \iff G(\omega)$$

then

$$g(t)e^{j\omega_0 t} \iff G(\omega - \omega_0) \tag{3.33}$$

*Proof:* By definition,

$$\mathcal{F}[g(t)e^{j\omega_0 t}] = \int_{-\infty}^{\infty} g(t)e^{j\omega_0 t} e^{-j\omega t} dt = \int_{-\infty}^{\infty} g(t)e^{-j(\omega - \omega_0)t} dt = G(\omega - \omega_0)$$

This property states that multiplication of a signal by a factor  $e^{j\omega_0 t}$  shifts the spectrum of that signal by  $\omega = \omega_0$ . Note the duality between the time-shifting and the frequency-shifting properties.

Changing  $\omega_0$  to  $-\omega_0$  in Eq. (3.33) yields

$$g(t)e^{-j\omega_0 t} \iff G(\omega + \omega_0) \quad (3.34)$$

Because  $e^{j\omega_0 t}$  is not a real function that can be generated, frequency shifting in practice is achieved by multiplying  $g(t)$  by a sinusoid. This can be seen from the fact that

$$g(t) \cos \omega_0 t = \frac{1}{2} [g(t)e^{j\omega_0 t} + g(t)e^{-j\omega_0 t}]$$

From Eqs. (3.33) and (3.34), it follows that

$$g(t) \cos \omega_0 t \iff \frac{1}{2} [G(\omega - \omega_0) + G(\omega + \omega_0)] \quad (3.35)$$

This shows that the multiplication of a signal  $g(t)$  by a sinusoid of frequency  $\omega_0$  shifts the spectrum  $G(\omega)$  by  $\pm\omega_0$ . Multiplication of a sinusoid  $\cos \omega_0 t$  by  $g(t)$  amounts to modulating the sinusoid amplitude. This type of modulation is known as **amplitude modulation**. The sinusoid  $\cos \omega_0 t$  is called the **carrier**, the signal  $g(t)$  is the **modulating signal**, and the signal  $g(t) \cos \omega_0 t$  is the **modulated signal**. Modulation and demodulation will be discussed in detail in Chapter 4.

To sketch a signal  $g(t) \cos \omega_0 t$ , we observe that

$$g(t) \cos \omega_0 t = \begin{cases} g(t) & \text{when } \cos \omega_0 t = 1 \\ -g(t) & \text{when } \cos \omega_0 t = -1 \end{cases}$$

Therefore,  $g(t) \cos \omega_0 t$  touches  $g(t)$  when the sinusoid  $\cos \omega_0 t$  is at its positive peaks and touches  $-g(t)$  when  $\cos \omega_0 t$  is at its negative peaks. This means that  $g(t)$  and  $-g(t)$  act as envelopes for the signal  $g(t) \cos \omega_0 t$  (see Fig. 3.21c). The signal  $-g(t)$  is a mirror image of  $g(t)$  about the horizontal axis. Figure 3.21 shows the signals  $g(t)$ ,  $g(t) \cos \omega_0 t$  and their spectra.

### Shifting the Phase Spectrum of a Modulated Signal

We can shift the phase of each spectral component of a modulated signal by a constant amount  $\theta_0$  merely by using a carrier  $\cos(\omega_0 t + \theta_0)$  instead of  $\cos \omega_0 t$ . If a signal  $g(t)$  is multiplied by  $\cos(\omega_0 t + \theta_0)$ , then using an argument similar to that used to derive Eq. (3.35), we can show that

$$g(t) \cos(\omega_0 t + \theta_0) \iff \frac{1}{2} [G(\omega - \omega_0) e^{j\theta_0} + G(\omega + \omega_0) e^{-j\theta_0}] \quad (3.36)$$

For a special case when  $\theta_0 = -\pi/2$ , Eq. (3.36) becomes

$$g(t) \sin \omega_0 t \iff \frac{1}{2} [G(\omega - \omega_0) e^{-j\pi/2} + G(\omega + \omega_0) e^{j\pi/2}] \quad (3.37)$$

Observe that  $\sin \omega_0 t$  is  $\cos \omega_0 t$  with a phase delay of  $\pi/2$ . Thus, shifting the carrier phase by  $\pi/2$  shifts the phase of every spectral component by  $\pi/2$ . Figure 3.21e and f shows the signal  $g(t) \sin \omega_0 t$  and its spectrum.

---

**EXAMPLE 3.12** Find and sketch the Fourier transform of the modulated signal  $g(t) \cos \omega_0 t$  in which  $g(t)$  is a gate pulse  $\text{rect}(t/T)$ , as shown in Fig. 3.22a.

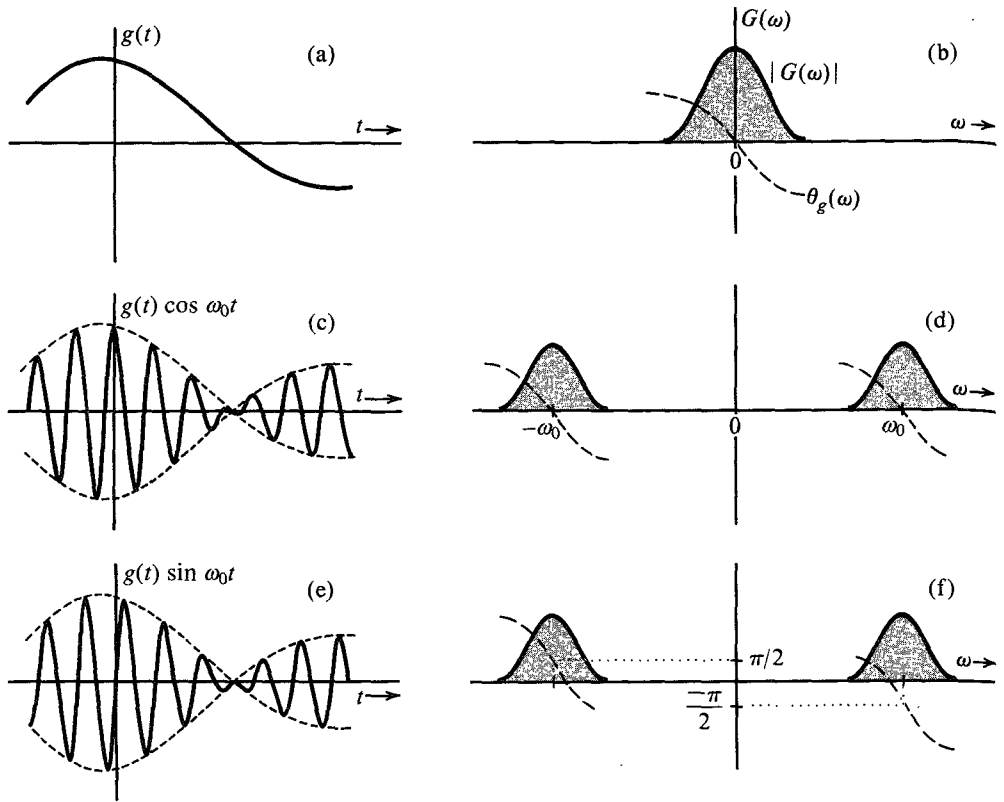


Figure 3.21 Amplitude modulation of a signal causes spectral shifting.

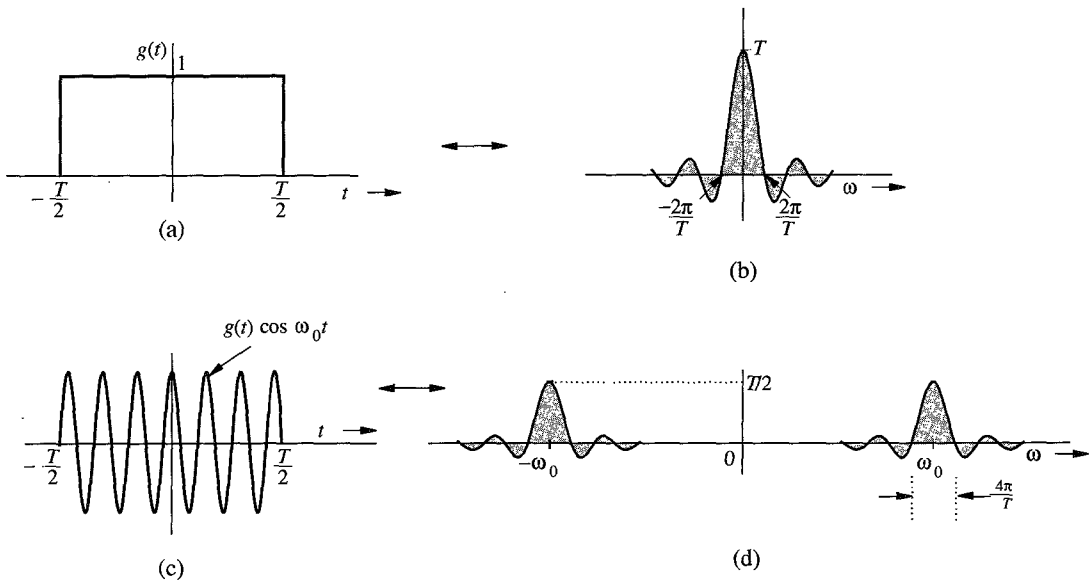


Figure 3.22 Example of spectral shifting by amplitude modulation.

The pulse  $g(t)$  is the same rectangular pulse shown in Fig. 3.10a (with  $\tau = T$ ). From pair 17 of Table 3.1, we find  $G(\omega)$ , the Fourier transform of  $g(t)$ , as

$$\text{rect}\left(\frac{t}{T}\right) \iff T \text{sinc}\left(\frac{\omega T}{2}\right)$$

This spectrum  $G(\omega)$  is shown in Fig. 3.22b. The signal  $g(t) \cos \omega_0 t$  is shown in Fig. 3.22c. From Eq. (3.35) it follows that

$$g(t) \cos \omega_0 t \iff \frac{1}{2}[G(\omega + \omega_0) + G(\omega - \omega_0)]$$

This spectrum of  $g(t) \cos \omega_0 t$  is obtained by shifting  $G(\omega)$  in Fig. 3.22b to the left by  $\omega_0$  and also to the right by  $\omega_0$  and then multiplying it by half, as shown in Fig. 3.22d.

### Application of Modulation

Modulation is used to shift signal spectra. Some of the situations where spectrum shifting is necessary are given next.

1. If several signals, each occupying the same frequency band, are transmitted simultaneously over the same transmission medium, they will all interfere; it will be impossible to separate or retrieve them at a receiver. For example, if all radio stations decide to broadcast audio signals simultaneously, the receiver will not be able to separate them. This problem is solved by using modulation, whereby each radio station is assigned a distinct carrier frequency. Each station transmits a modulated signal, thus shifting the signal spectrum to its allocated band, which is not occupied by any other station. A radio receiver can pick up any station by tuning to the band of the desired station. The receiver must now demodulate the received signal (undo the effect of modulation). Demodulation therefore consists of another spectral shift required to restore the signal to its original band. Note that both modulation and demodulation implement spectral shifting. Consequently, demodulation operation is similar to modulation (see Prob. 3.3-10). This method of transmitting several signals simultaneously over a channel by sharing its frequency band is known as **frequency-division multiplexing (FDM)**.
2. For effective radiation of power over a radio link, the antenna size must be on the order of the wavelength of the signal to be radiated. Audio signal frequencies are so low (wavelengths are so large) that impracticably large antennas will be required for radiation. Here, shifting the spectrum to a higher frequency (a smaller wavelength) by modulation solves the problem.

### Bandpass Signals

Figure 3.21d and f shows that if  $g_c(t)$  and  $g_s(t)$  are low-pass signals, each with a bandwidth  $B$  Hz or  $2\pi B$  rad/s, then the signals  $g_c(t) \cos \omega_0 t$  and  $g_s(t) \sin \omega_0 t$  are both bandpass signals occupying the same band, and each having a bandwidth of  $4\pi B$  rad/s. Hence, a linear combination of both these signals will also be a bandpass signal occupying the same band as that of the either signal, and with the same bandwidth ( $4\pi B$  rad/s). Hence, a general bandpass signal  $g_{bp}(t)$  can be expressed as\*

$$g_{bp}(t) = g_c(t) \cos \omega_0 t + g_s(t) \sin \omega_0 t \quad (3.38)$$

\* See Sec. 11.5 for a rigorous proof of this statement.

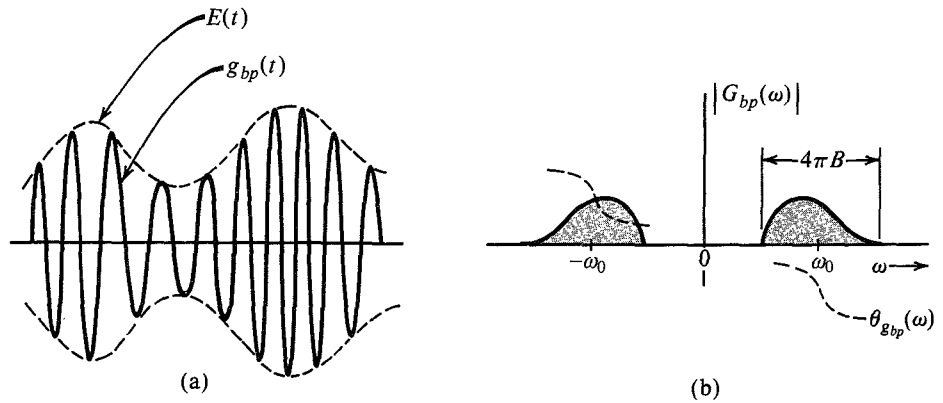


Figure 3.23 Bandpass signal and its spectrum.

The spectrum of  $g_{bp}(t)$  is centered at  $\pm\omega_0$  and has a bandwidth  $4\pi B$ , as shown in Fig. 3.23. Although the magnitude spectra of both  $g_c(t) \cos \omega_0 t$  and  $g_s(t) \sin \omega_0 t$  are symmetrical about  $\pm\omega_0$ , the magnitude spectrum of their sum,  $g_{bp}(t)$ , is not necessarily symmetrical about  $\pm\omega_0$ . This is due to the fact that the amplitudes of the two signals do not add directly because of their phases for the reason that

$$a_1 e^{j\varphi_1} + a_2 e^{j\varphi_2} \neq (a_1 + a_2) e^{j(\varphi_1 + \varphi_2)}$$

A typical bandpass signal  $g_{bp}(t)$  and its spectra are shown in Fig. 3.23. Using a well-known trigonometric identity, Eq. (3.38) can be expressed as

$$g_{bp}(t) = E(t) \cos [\omega_0 t + \psi(t)] \quad (3.39)$$

where

$$E(t) = +\sqrt{g_c^2(t) + g_s^2(t)} \quad (3.40a)$$

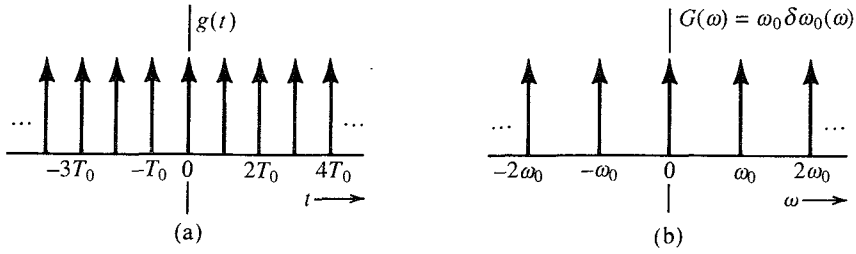
$$\psi(t) = -\tan^{-1} \left[ \frac{g_s(t)}{g_c(t)} \right] \quad (3.40b)$$

Because  $g_c(t)$  and  $g_s(t)$  are low-pass signals,  $E(t)$  and  $\psi(t)$  are also low-pass signals. Because  $E(t)$  is nonnegative [Eq. (3.40a)], it follows from Eq. (3.39) that  $E(t)$  is a slowly varying envelope and  $\psi(t)$  is a slowly varying phase of the bandpass signal  $g_{bp}(t)$ , as shown in Fig. 3.23. Thus, the bandpass signal  $g_{bp}(t)$  will appear as a sinusoid of slowly varying amplitude. Because of the time-varying phase  $\psi(t)$  the frequency of the sinusoid also varies slowly\* with time about the center frequency  $\omega_0$ .

**EXAMPLE 3.13** Find the Fourier transform of a general periodic signal  $g(t)$  of period  $T_0$ , and hence, determine the Fourier transform of the periodic impulse train  $\delta_{T_0}(t)$  shown in Fig. 3.24a.

\* It is necessary that  $2\pi B \ll \omega_0$  for a well-defined envelope. Otherwise the variations of  $E(t)$  are of the same order as the carrier, and it will be difficult to separate the envelope from the carrier.





**Figure 3.24** Impulse train and its spectrum.

A periodic signal  $g(t)$  can be expressed as an exponential Fourier series as

$$g(t) = \sum_{n=-\infty}^{\infty} D_n e^{jn\omega_0 t} \quad \omega_0 = \frac{2\pi}{T_0}$$

Therefore,

$$g(t) \iff \sum_{n=-\infty}^{\infty} \mathcal{F}[D_n e^{jn\omega_0 t}]$$

Now from Eq. (3.20a), it follows that

$$g(t) \iff 2\pi \sum_{n=-\infty}^{\infty} D_n \delta(\omega - n\omega_0) \quad (3.41)$$

Equation (2.89) shows that the impulse train  $\delta_{T_0}(t)$  can be expressed as an exponential Fourier series as

$$\delta_{T_0}(t) = \frac{1}{T_0} \sum_{n=-\infty}^{\infty} e^{jn\omega_0 t} \quad \omega_0 = \frac{2\pi}{T_0}$$

Here  $D_n = 1/T_0$ . Therefore, from Eq. (3.41),

$$\begin{aligned} \delta_{T_0}(t) &\iff \frac{2\pi}{T_0} \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \\ &= \omega_0 \delta_{\omega_0}(\omega) \quad \omega_0 = \frac{2\pi}{T_0} \end{aligned} \quad (3.42)$$

Thus, the spectrum of the impulse train also happens to be an impulse train (in the frequency domain), as shown in Fig. 3.24b.

### 3.3.6 Convolution

The convolution of two functions  $g(t)$  and  $w(t)$ , denoted by  $g(t) * w(t)$ , is defined by the integral

$$g(t) * w(t) = \int_{-\infty}^{\infty} g(\tau) w(t - \tau) d\tau$$

The time convolution property and its dual, the frequency convolution property, state that if

$$g_1(t) \iff G_1(\omega) \quad \text{and} \quad g_2(t) \iff G_2(\omega)$$

then (**time convolution**)

$$g_1(t) * g_2(t) \iff G_1(\omega)G_2(\omega) \quad (3.43)$$

and (**frequency convolution**)

$$g_1(t)g_2(t) \iff \frac{1}{2\pi}G_1(\omega) * G_2(\omega) \quad (3.44)$$

*Proof:* By definition,

$$\begin{aligned} \mathcal{F}[g_1(t) * g_2(t)] &= \int_{-\infty}^{\infty} e^{-j\omega t} \left[ \int_{-\infty}^{\infty} g_1(\tau)g_2(t - \tau) d\tau \right] dt \\ &= \int_{-\infty}^{\infty} g_1(\tau) \left[ \int_{-\infty}^{\infty} e^{-j\omega t} g_2(t - \tau) dt \right] d\tau \end{aligned}$$

The inner integral is the Fourier transform of  $g_2(t - \tau)$ , given by [time-shifting property in Eq. (3.30)]  $G_2(\omega)e^{-j\omega\tau}$ . Hence,

$$\begin{aligned} \mathcal{F}[g_1(t) * g_2(t)] &= \int_{-\infty}^{\infty} g_1(\tau)e^{-j\omega\tau} G_2(\omega) d\tau \\ &= G_2(\omega) \int_{-\infty}^{\infty} g_1(\tau)e^{-j\omega\tau} d\tau = G_1(\omega)G_2(\omega) \end{aligned}$$

The frequency convolution property (3.44) can be proved in exactly the same way by reversing the roles of  $g(t)$  and  $G(\omega)$ .

### Bandwidth of the Product of Two Signals

If  $g_1(t)$  and  $g_2(t)$  have bandwidths  $B_1$  and  $B_2$  Hz, respectively, the bandwidth of  $g_1(t)g_2(t)$  is  $B_1 + B_2$  Hz. This result follows from the application of the width property of convolution<sup>3</sup> to Eq. (3.44). This property states that the width of  $x * y$  is the sum of the widths of  $x$  and  $y$ . Consequently, if the bandwidth of  $g(t)$  is  $B$  Hz, then the bandwidth of  $g^2(t)$  is  $2B$  Hz, and the bandwidth of  $g^n(t)$  is  $nB$  Hz.\*

---

**EXAMPLE 3.14** Using the time convolution property, show that if

$$g(t) \iff G(\omega)$$

then

$$\int_{-\infty}^t g(\tau) d\tau \iff \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \quad (3.45)$$

---

\* The width property of convolution does not hold in some pathological cases. This happens when the convolution of two functions is zero over a range even when both functions are nonzero, e.g.,  $\sin \omega_0 t u(t) * u(t)$ . Technically the property holds even in this case if in calculating the width of the convolved function we take into account that range where the convolution is zero.

Because

$$u(t - \tau) = \begin{cases} 1 & \tau \leq t \\ 0 & \tau > t \end{cases}$$

it follows that

$$g(t) * u(t) = \int_{-\infty}^{\infty} g(\tau)u(t - \tau)d\tau = \int_{-\infty}^t g(\tau) d\tau$$

Now from the time convolution property [Eq. (3.43)], it follows that

$$\begin{aligned} g(t) * u(t) &\iff G(\omega)U(\omega) \\ &= G(\omega) \left[ \frac{1}{j\omega} + \pi\delta(\omega) \right] \\ &= \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \end{aligned}$$

In deriving the last result we used pair 11 of Table 3.1 and Eq. (2.18a).

### 3.3.7 Time Differentiation and Time Integration

If

$$g(t) \iff G(\omega)$$

then (**time differentiation**)\*

$$\frac{dg}{dt} \iff j\omega G(\omega) \quad (3.46)$$

and (**time integration**)

$$\int_{-\infty}^t g(\tau)d\tau \iff \frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega) \quad (3.47)$$

*Proof:* Differentiation of both sides of Eq. (3.8b) yields

$$\frac{dg}{dt} = \frac{1}{2\pi} \int_{-\infty}^{\infty} j\omega G(\omega)e^{j\omega t} d\omega$$

This shows that

$$\frac{dg}{dt} \iff j\omega G(\omega)$$

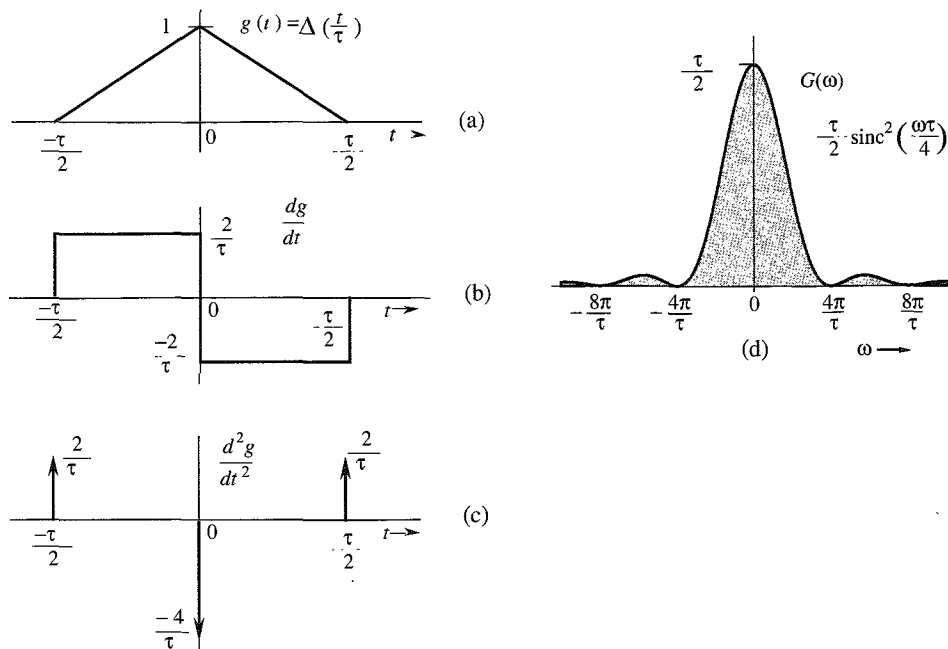
Repeated application of this property yields

$$\frac{d^n g}{dt^n} \iff (j\omega)^n G(\omega) \quad (3.48)$$

The time integration property [Eq. (3.47)] already has been proved in Example 3.14.

\* Valid only if the transform of  $dg/dt$  exists.

**EXAMPLE 3.15** Using the time differentiation property, find the Fourier transform of the triangle pulse  $\Delta(t/\tau)$  shown in Fig. 3.25a.



**Figure 3.25** Finding the Fourier transform of a piecewise-linear signal using the time differentiation property.

To find the Fourier transform of this pulse we differentiate it successively, as shown in Fig. 3.25b and c. The second derivative consists of a sequence of impulses, as shown in Fig. 3.25c. Recall that the derivative of a signal at a jump discontinuity is an impulse of strength equal to the amount of jump. The function  $dg/dt$  has a positive jump of  $2/\tau$  at  $t = \pm\tau/2$ , and a negative jump of  $4/\tau$  at  $t = 0$ . Therefore,

$$\frac{d^2g}{dt^2} = \frac{2}{\tau} \left[ \delta\left(t + \frac{\tau}{2}\right) - 2\delta(t) + \delta\left(t - \frac{\tau}{2}\right) \right] \quad (3.49)$$

From the time differentiation property (3.48),

$$\frac{d^2g}{dt^2} \iff (j\omega)^2 G(\omega) = -\omega^2 G(\omega) \quad (3.50a)$$

Also, from the time-shifting property (3.30),

$$\delta(t - t_0) \iff e^{-j\omega t_0} \quad (3.50b)$$

Taking the Fourier transform of Eq. (3.49) and using the results in Eqs. (3.50), we obtain

$$-\omega^2 G(\omega) = \frac{2}{\tau} \left( e^{j\frac{\omega\tau}{2}} - 2 + e^{-j\frac{\omega\tau}{2}} \right) = \frac{4}{\tau} \left( \cos \frac{\omega\tau}{2} - 1 \right) = -\frac{8}{\tau} \sin^2 \left( \frac{\omega\tau}{4} \right)$$

and

$$G(\omega) = \frac{8}{\omega^2 \tau} \sin^2 \left( \frac{\omega \tau}{4} \right) = \frac{\tau}{2} \left[ \frac{\sin(\omega \tau / 4)}{\omega \tau / 4} \right]^2 = \frac{\tau}{2} \operatorname{sinc}^2 \left( \frac{\omega \tau}{4} \right) \quad (3.51)$$

The spectrum  $G(\omega)$  is shown in Fig. 3.25d. This procedure of finding the Fourier transform can be applied to any function  $g(t)$  made up of straight-line segments with  $g(t) \rightarrow 0$  as  $|t| \rightarrow \infty$ . The second derivative of such a signal yields a sequence of impulses whose Fourier transform can be found by inspection. This example suggests a numerical method of finding the Fourier transform of an arbitrary signal  $g(t)$  by approximating the signal by straight-line segments.

### 3.4 SIGNAL TRANSMISSION THROUGH A LINEAR SYSTEM

For a linear, time-invariant, continuous-time system the input-output relationship is given by

$$y(t) = g(t) * h(t) \quad (3.52)$$

where  $g(t)$  is the input,  $y(t)$  is the output, and  $h(t)$  is the unit impulse response of the linear time-invariant system. If

$$g(t) \iff G(\omega), \quad y(t) \iff Y(\omega), \quad \text{and} \quad h(t) \iff H(\omega)$$

where  $H(\omega)$  is the system transfer function, then application of the time convolution property to Eq. (3.52) yields

$$Y(\omega) = G(\omega)H(\omega) \quad (3.53)$$

#### 3.4.1 Signal Distortion during Transmission

The transmission of an input signal  $g(t)$  through a system changes it into the output signal  $y(t)$ . Equation (3.53) shows the nature of this change or modification. Here  $G(\omega)$  and  $Y(\omega)$  are the

**Table 3.2**  
**Fourier Transform Operations**

Operation	$g(t)$	$G(\omega)$
Addition	$g_1(t) + g_2(t)$	$G_1(\omega) + G_2(\omega)$
Scalar multiplication	$kg(t)$	$kG(\omega)$
Symmetry	$G(t)$	$2\pi g(-\omega)$
Scaling	$g(at)$	$\frac{1}{ a } G\left(\frac{\omega}{a}\right)$
Time shift	$g(t - t_0)$	$G(\omega)e^{-j\omega t_0}$
Frequency shift	$g(t)e^{j\omega_0 t}$	$G(\omega - \omega_0)$
Time convolution	$g_1(t) * g_2(t)$	$G_1(\omega)G_2(\omega)$
Frequency convolution	$g_1(t)g_2(t)$	$\frac{1}{2\pi} G_1(\omega) * G_2(\omega)$
Time differentiation	$\frac{d^n g}{dt^n}$	$(j\omega)^n G(\omega)$
Time integration	$\int_{-\infty}^t g(x) dx$	$\frac{G(\omega)}{j\omega} + \pi G(0)\delta(\omega)$

spectra of the input and the output, respectively. Therefore,  $H(\omega)$  is the spectral response of the system. The output spectrum is given by the input spectrum multiplied by the spectral response of the system. Equation (3.53) clearly brings out the spectral shaping (or modification) of the signal by the system. Equation (3.53) can be expressed in polar form as

$$|Y(\omega)|e^{j\theta_y(\omega)} = |G(\omega)||H(\omega)|e^{j[\theta_g(\omega)+\theta_h(\omega)]}$$

Therefore,

$$|Y(\omega)| = |G(\omega)| |H(\omega)| \quad (3.54a)$$

$$\theta_y(\omega) = \theta_g(\omega) + \theta_h(\omega) \quad (3.54b)$$

During the transmission, the input signal amplitude spectrum  $|G(\omega)|$  is changed to  $|G(\omega)||H(\omega)|$ . Similarly, the input signal phase spectrum  $\theta_g(\omega)$  is changed to  $\theta_g(\omega) + \theta_h(\omega)$ . An input signal spectral component of frequency  $\omega$  is modified in amplitude by a factor  $|H(\omega)|$  and is shifted in phase by an angle  $\theta_h(\omega)$ . Clearly,  $|H(\omega)|$  is the amplitude response, and  $\theta_h(\omega)$  is the phase response of the system. The plots of  $|H(\omega)|$  and  $\theta_h(\omega)$  as functions of  $\omega$  show at a glance how the system modifies the amplitudes and phases of various sinusoidal inputs. This is why  $H(\omega)$  is called the **frequency response** of the system. During transmission through the system, some frequency components may be boosted in amplitude, while others may be attenuated. The relative phases of the various components also change. In general, the output waveform will be different from the input waveform.

### Distortionless Transmission

In several applications, such as signal amplification or message signal transmission over a communication channel, we require the output waveform to be a replica of the input waveform. In such cases, we need to minimize the distortion caused by the amplifier or the communication channel. It is therefore of practical interest to determine the characteristics of a system that allows a signal to pass without distortion (**distortionless transmission**).

Transmission is said to be distortionless if the input and the output have identical wave shapes within a multiplicative constant. A delayed output that retains the input waveform is also considered distortionless. Thus, in distortionless transmission, the input  $g(t)$  and the output  $y(t)$  satisfy the condition

$$y(t) = kg(t - t_d) \quad (3.55)$$

The Fourier transform of this equation yields

$$Y(\omega) = kG(\omega)e^{-j\omega t_d}$$

But

$$Y(\omega) = G(\omega)H(\omega)$$

Therefore,

$$H(\omega) = k e^{-j\omega t_d}$$

This is the transfer function required for distortionless transmission. From this equation it follows that

$$|H(\omega)| = k \quad (3.56a)$$

$$\theta_h(\omega) = -\omega t_d \quad (3.56b)$$

This shows that for distortionless transmission, the amplitude response  $|H(\omega)|$  must be a constant, and the phase response  $\theta_h(\omega)$  must be a linear function of  $\omega$ , as shown in Fig. 3.26. The slope of  $\theta_h(\omega)$  with respect to  $\omega$  is  $-t_d$ , where  $t_d$  is the delay of the output with respect to the input.\*

### Intuitive Explanation of the Distortionless Transmission Conditions

It is instructive to derive the conditions for distortionless transmission heuristically. Once again, imagine  $g(t)$  to be composed of various sinusoids (its spectral components), which are being passed through a distortionless system. For the distortionless case, the output signal is the input signal multiplied by  $k$  and delayed by  $t_d$ . To synthesize such a signal, we need exactly the same components as those of  $g(t)$ , with each component multiplied by  $k$  and delayed by  $t_d$ . This means that the system transfer function  $H(\omega)$  should be such that each sinusoidal component suffers the same attenuation  $k$  and each component undergoes the same time delay of  $t_d$  seconds. The first condition requires that

$$|H(\omega)| = k$$

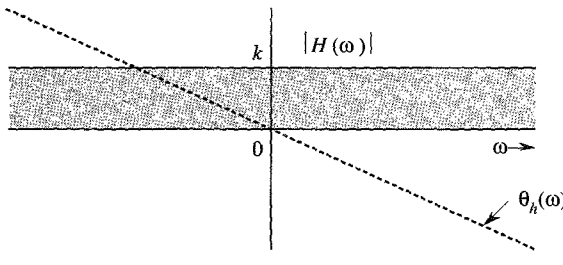
We have seen earlier (sec. 3.3) that to achieve the same time delay  $t_d$  for every frequency component requires a linear phase delay  $\omega t_d$  (Fig. 3.19). Therefore,

$$\theta_h(\omega) = -\omega t_d$$

The time delay resulting from the signal transmission through a system is the negative of the slope of the system phase response  $\theta_h$ ; that is,

$$t_d(\omega) = -\frac{d\theta_h}{d\omega} \quad (3.57)$$

If the slope of  $\theta_h$  is constant (that is, if  $\theta_h$  is linear with  $\omega$ ), all the components are delayed by the same time interval  $t_d$ . But if the slope is not constant,  $t_d$ , the time delay, varies with frequency. This means that different frequency components undergo different amounts of time delay, and consequently the output waveform will not be a replica of the input waveform. A good way of judging phase linearity is to plot  $t_d$  as a function of frequency. For a distortionless system,  $t_d$  should be constant over the band of interest.†



**Figure 3.26** Linear time-invariant system frequency response for distortionless transmission.

\* In addition, we require that  $\theta_h(0)$  either be 0 (as shown in Fig. 3.26) or have a constant value  $n\pi$  ( $n$  an integer), that is,  $\theta_h(\omega) = n\pi - \omega t_d$ . The addition of the excess phase of  $n\pi$  may at most change the sign of the signal.

† Figure 3.26 shows that for distortionless transmission, the phase response is not only linear, but must also pass through the origin. This latter requirement can be somewhat relaxed for bandpass signals. The phase at the origin may be any constant [ $\theta_h(\omega) = \theta_0 - \omega t_d$  or  $\theta_h(0) = \theta_0$ ]. The reason for this can be found in Eq. (3.36), which shows that the addition of a constant phase  $\theta_0$  to a spectrum of a bandpass signal amounts to a phase shift of the carrier by  $\theta_0$ . The modulating signal (the envelope) is not affected. The output envelope is the same as the input envelope delayed

It is often thought (erroneously) that flatness of amplitude response  $|H(\omega)|$  alone can guarantee signal quality. A system may have a flat amplitude response and yet distort a signal beyond recognition if the phase response is not linear ( $t_d$  not constant).

### The Nature of Distortion in Audio and Video Signals

Generally speaking, a human ear can readily perceive amplitude distortion, although it is relatively insensitive to phase distortion. For the phase distortion to become noticeable, the variation in delay (variation in the slope of  $\theta_h$ ) should be comparable to the signal duration (or the physically perceptible duration, in case the signal itself is long). In the case of audio signals, each spoken syllable can be considered as an individual signal. The average duration of a spoken syllable is of a magnitude on the order of 0.01 to 0.1 second. The audio systems may have nonlinear phases, yet no noticeable signal distortion results because in practical audio systems, maximum variation in the slope of  $\theta_h$  is only a small fraction of a millisecond. This is the real reason behind the statement that “the human ear is relatively insensitive to phase distortion.”<sup>4</sup> As a result, the manufacturers of audio equipment make available only  $|H(\omega)|$ , the amplitude response characteristic of their systems.

For video signals, on the other hand, the situation is exactly the opposite. The human eye is sensitive to phase distortion but is relatively insensitive to amplitude distortion. The amplitude distortion in television signals manifests itself as a partial destruction of the relative half-tone values of the resulting picture, which is not readily apparent to the human eye. The phase distortion (nonlinear phase), on the other hand, causes different time delays in different picture elements. This results in a smeared picture, which is readily apparent to the human eye. Phase distortion is also very important in digital communication systems because the nonlinear phase characteristic of a channel causes pulse dispersion (spreading out), which in turn causes pulses to interfere with neighboring pulses. This interference can cause an error in the pulse amplitude at the receiver: a binary **1** may read as **0**, and vice versa.

**EXAMPLE 3.16** If  $g(t)$  and  $y(t)$  are the input and the output, respectively, of a simple  $RC$  low-pass filter (Fig. 3.27a), determine the transfer function  $H(\omega)$  and sketch  $|H(\omega)|$ ,  $\theta_h(\omega)$ , and  $t_d(\omega)$ . For distortionless transmission through this filter, what is the requirement on the bandwidth of  $g(t)$  if amplitude response variation within 2% and time delay variation within 5% are tolerable? What is the transmission delay? Find the output  $y(t)$ .

Application of the voltage division rule to this circuit yields

$$H(\omega) = \frac{1/j\omega C}{R + (1/j\omega C)} = \frac{1}{1 + j\omega RC} = \frac{a}{a + j\omega}$$

where

$$a = \frac{1}{RC} = 10^6$$

Hence,

$$|H(\omega)| = \frac{a}{\sqrt{a^2 + \omega^2}} \simeq 1 \quad \omega \ll a$$

$$\theta_h(\omega) = -\tan^{-1} \frac{\omega}{a} \simeq -\frac{\omega}{a} \quad \omega \ll a$$

by  $t_g = -d\theta_h/d\omega$ , called the **group** or **envelope delay**, and the output carrier is the same as the input carrier delayed by  $t_p = -\theta_h(\omega_0)/\omega_0$ , called the **phase delay**, where  $\omega_0$  is the center frequency of the passband.



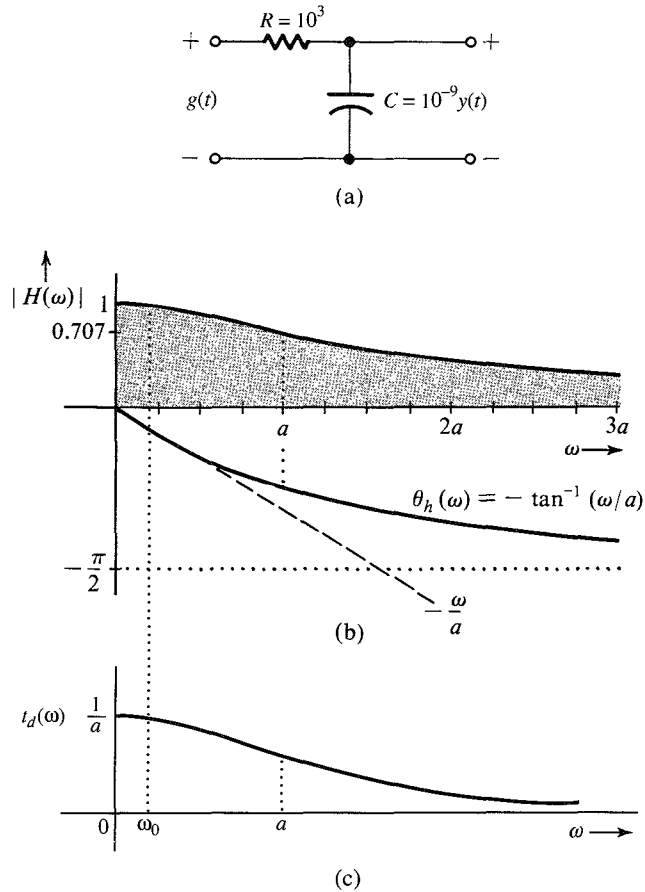


Figure 3.27 Simple RC filter, its frequency response and time delay.

Finally, the time delay is given by [Eq. (3.57)]

$$t_d(\omega) = -\frac{d\theta_h}{d\omega} = \frac{a}{\omega^2 + a^2} \simeq \frac{1}{a} = 10^{-6} \quad \omega \ll a$$

The amplitude and phase response characteristics are given in Fig. 3.27b. The time delay  $t_d$  as a function of  $\omega$  is shown in Fig. 3.27c. For  $\omega \ll a$  ( $a = 10^6$ ), the amplitude response is practically constant and the phase shift is nearly linear. The phase linearity results in a constant time delay characteristic. The filter therefore can transmit low-frequency signals with negligible distortion.

In our case, amplitude response variation within 2% and time delay variation within 5% are tolerable. Let  $\omega_0$  be the highest bandwidth of a signal that can be transmitted within these specifications. To compute  $\omega_0$  observe that the filter is a low-pass filter with gain and time delay both at maximum when  $\omega = 0$  and

$$|H(0)| = 1 \quad \text{and} \quad t_d(0) = \frac{1}{a}$$

Therefore,  $|H(\omega_0)| \geq 0.98$  and  $t_d(\omega_0) \geq 0.95/a$ , so that

$$|H(\omega_0)| = \frac{a}{\sqrt{\omega_0^2 + a^2}} \geq 0.98 \implies \omega_0 \leq 0.203a = 203,000$$

$$t_d(\omega_0) = \frac{a}{\omega_0^2 + a^2} \geq \frac{0.95}{a} \implies \omega_0 \leq 0.2294a = 229,400$$

The smaller of the two values,  $\omega_0 = 203,000$  rad/s or 32.31 kHz, is the highest bandwidth that satisfies both constraints on  $|H(\omega)|$  and  $t_d$ .

The time delay  $t_d \approx 1/a = 10^{-6}$  over this band (see Fig. 3.27c). Also the amplitude response is almost unity (Fig. 3.27b). Therefore, the output  $y(t) \approx g(t - 10^{-6})$ .

### 3.5 IDEAL AND PRACTICAL FILTERS

Ideal filters allow distortionless transmission of a certain band of frequencies and suppress all the remaining frequencies. The ideal low-pass filter (Fig. 3.28), for example, allows all components below  $\omega = W$  rad/s to pass without distortion and suppresses all components above  $\omega = W$ . Figure 3.29 shows ideal high-pass and bandpass filter characteristics.

The ideal low-pass filter in Fig. 3.28a has a linear phase of slope  $-t_d$ , which results in a time delay of  $t_d$  seconds for all its input components of frequencies below  $W$  rad/s. Therefore, if the input is a signal  $g(t)$  band-limited to  $W$  rad/s, the output  $y(t)$  is  $g(t)$  delayed by  $t_d$ , that is,

$$y(t) = g(t - t_d)$$

The signal  $g(t)$  is transmitted by this system without distortion, but with time delay  $t_d$ . For this filter  $|H(\omega)| = \text{rect}(\omega/2W)$ , and  $\theta_h(\omega) = -\omega t_d$ , so that

$$H(\omega) = \text{rect}\left(\frac{\omega}{2W}\right) e^{-j\omega t_d} \quad (3.58a)$$

The unit impulse response  $h(t)$  of this filter is found from pair 18 in Table 3.1 and the time-shifting property:

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\left[\text{rect}\left(\frac{\omega}{2W}\right) e^{-j\omega t_d}\right] \\ &= \frac{W}{\pi} \text{sinc}[W(t - t_d)] \end{aligned} \quad (3.58b)$$

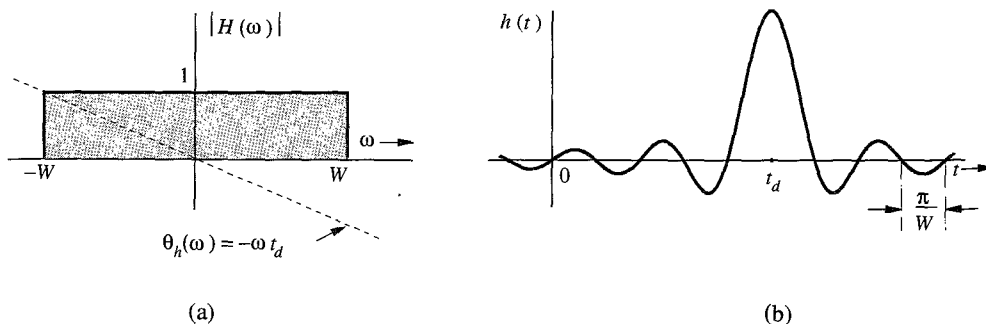
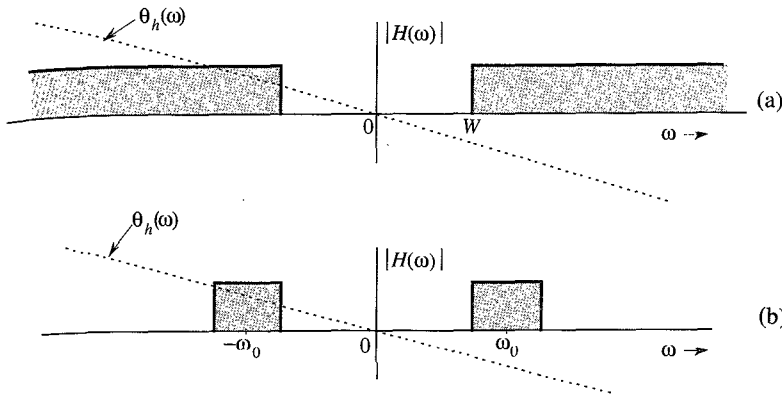


Figure 3.28 Ideal low-pass filter frequency response and its impulse response.



**Figure 3.29** Ideal high-pass and band-pass filter frequency responses.

Recall that  $h(t)$  is the system response to impulse input  $\delta(t)$ , which is applied at  $t = 0$ . Figure 3.28b shows a curious fact: the response  $h(t)$  begins even before the input is applied (at  $t = 0$ ). Clearly, the filter is noncausal and therefore physically unrealizable. Similarly, one can show that other ideal filters (such as the ideal high-pass or the ideal bandpass filters shown in Fig. 3.29) are also physically unrealizable.

For a physically realizable system,  $h(t)$  must be causal; that is,

$$h(t) = 0 \quad \text{for } t < 0$$

In the frequency domain, this condition is equivalent to the well-known **Paley-Wiener criterion**, which states that the necessary and sufficient condition for the amplitude response  $|H(\omega)|$  to be realizable is\*

$$\int_{-\infty}^{\infty} \frac{|\ln |H(\omega)||}{1 + \omega^2} d\omega < \infty \quad (3.59)$$

If  $H(\omega)$  does not satisfy this condition, it is unrealizable. Note that if  $|H(\omega)| = 0$  over any finite band,  $|\ln |H(\omega)|| = \infty$  over that band, and the condition (3.59) is violated. If, however,  $H(\omega) = 0$  at a single frequency (or a set of discrete frequencies), the integral in Eq. (3.59) may still be finite even though the integrand is infinite. Therefore, for a physically realizable system,  $H(\omega)$  may be zero at some discrete frequencies, but it cannot be zero over any finite band. According to this criterion, ideal filter characteristics (Figs. 3.28 and 3.29) are clearly unrealizable.

The impulse response  $h(t)$  in Fig. 3.28 is not realizable. One practical approach to filter design is to cut off the tail of  $h(t)$  for  $t < 0$ . The resulting causal impulse response  $\hat{h}(t)$ , where

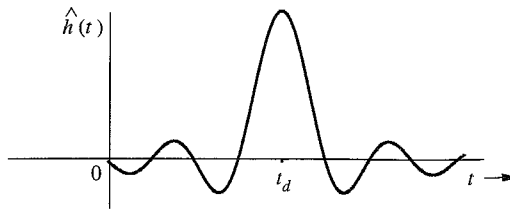
$$\hat{h}(t) = h(t)u(t)$$

is physically realizable because it is causal (Fig. 3.30). If  $t_d$  is sufficiently large,  $\hat{h}(t)$  will be a close approximation of  $h(t)$ , and the resulting filter  $\hat{H}(\omega)$  will be a good approximation of an ideal filter. This close realization of the ideal filter is achieved because of the increased value

\*  $|H(\omega)|$  is assumed to be square integrable, that is,

$$\int_{-\infty}^{\infty} |H(\omega)|^2 d\omega$$

is finite. Note that the Paley-Wiener criterion is a criterion for the realizability of the amplitude response  $|H(\omega)|$ .



**Figure 3.30** Approximate realization of an ideal low-pass filter by truncating its impulse response.

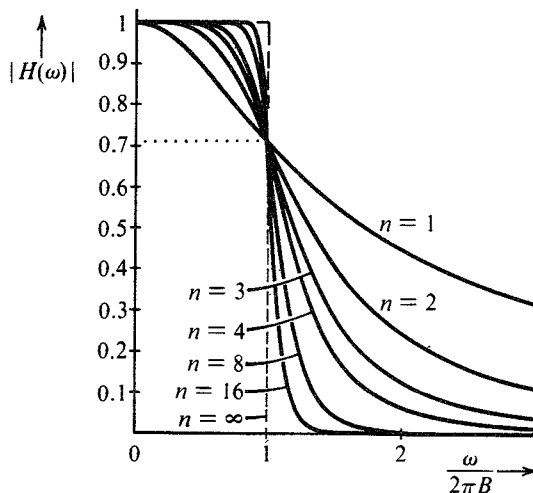
of time delay  $t_d$ . This means that the price of close realization is higher delay in the output; this is often true of noncausal systems. Of course, theoretically a delay  $t_d = \infty$  is needed to realize the ideal characteristics. But a glance at Fig. 3.28b shows that a delay  $t_d$  of three or four times  $\pi/W$  will make  $\hat{h}(t)$  a reasonably close version of  $h(t - t_d)$ . For instance, an audio filter is required to handle frequencies of up to 20 kHz. In this case a  $t_d$  of about  $10^{-4}$  (0.1 ms) would be a reasonable choice. The truncation operation [cutting the tail of  $h(t)$  to make it causal], however, creates some unsuspected problems of spectral spread and leakage, and can be partly corrected by truncating  $h(t)$  gradually (rather than abruptly) using a tapered window function.<sup>5</sup>

In practice, we can realize a variety of filter characteristics to approach ideal characteristics. Practical (realizable) filter characteristics are gradual, without jump discontinuities in the amplitude response  $|H(\omega)|$ . The well-known Butterworth filters, for example, have amplitude response

$$|H(\omega)| = \frac{1}{\sqrt{1 + (\omega/2\pi B)^{2n}}}$$

These characteristics are shown in Fig. 3.31 for several values of  $n$  (the order of the filter). Note that the amplitude response approaches an ideal low-pass behavior as  $n \rightarrow \infty$ .

The half-power bandwidth of a filter is defined as the bandwidth over which the amplitude response  $|H(\omega)|$  remains constant within variations of 3 dB (or a ratio of  $1/\sqrt{2}$ , that is, 0.707). Figure 3.31 shows that for all  $n$ , the Butterworth filter (half-power) bandwidth is  $B$  Hz. The



**Figure 3.31** Butterworth filter characteristic.

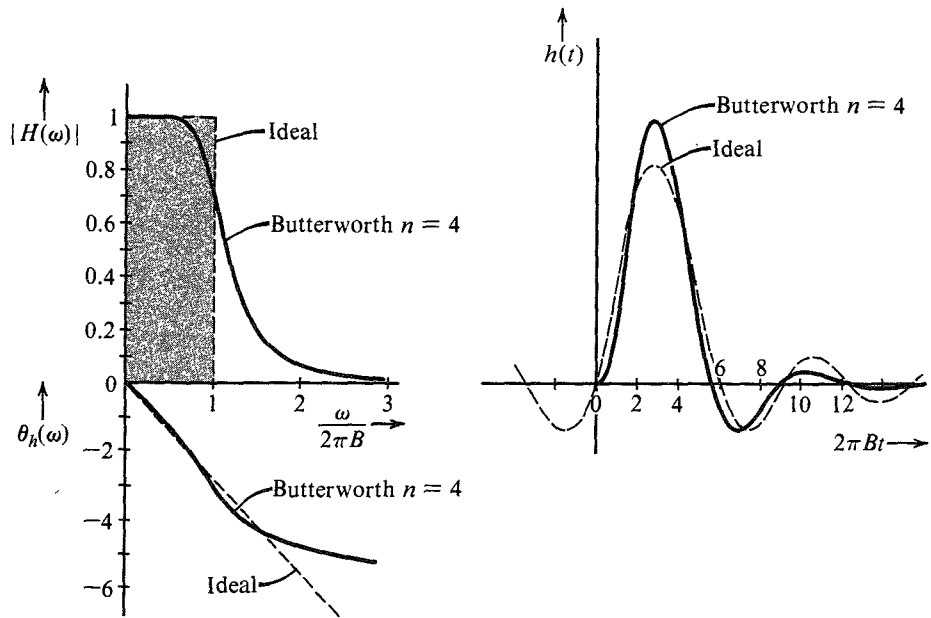


Figure 3.32 Comparison of Butterworth filter ( $n = 4$ ) with an ideal filter.

half-power bandwidth of a low-pass filter is also called the **cutoff frequency**. Figure 3.32 shows  $|H(\omega)|$ ,  $\theta_h(\omega)$ , and  $h(t)$  for the case of  $n = 4$ .

It should be remembered that the magnitude  $|H(\omega)|$  and the phase  $\theta_h(\omega)$  of a system are interdependent; that is, we cannot choose  $|H(\omega)|$  and  $\theta_h(\omega)$  independently as we please. A certain trade-off exists between ideal magnitude and ideal phase characteristics. If we try to perfect  $|H(\omega)|$  more,  $\theta_h(\omega)$  deviates more from the ideal, and vice versa. As  $n \rightarrow \infty$ , the amplitude response approaches ideal, but the corresponding phase response is badly distorted in the vicinity of the cutoff frequency  $B$  Hz.

### Digital Filters

Analog signals can also be processed by digital means (A/D conversion). This involves sampling, quantizing, and coding. The resulting digital signal can be processed by a small, special-purpose digital computer designed to convert the input sequence into a desired output sequence. The output sequence is converted back into the desired analog signal. A special algorithm of the processing digital computer can be used to achieve a given signal operation (e.g., low-pass, bandpass, or high-pass filtering).

Digital processing of analog signals has several advantages. A small, special-purpose computer can be time-shared for several uses, and the cost of digital implementation is often considerably lower than that of its analog counterpart. The accuracy of a digital filter is dependent only on the computer word length, the quantizing interval, and the sampling rate (aliasing error). Digital filters employ simple elements, such as adders, multipliers, shifters, and delay elements, rather than *RLC* components and operational amplifiers. As a result, they are generally unaffected by such factors as component accuracy, temperature stability, long-term drift, and so on, that afflict analog filter circuits. Also, many of the circuit restrictions

imposed by physical limitations of analog devices can be removed, or at least circumvented, in a digital processor. Moreover, filters of a high order can be realized easily. Finally, digital filters can be modified simply by changing the algorithm of the computer, in contrast to an analog system, which may have to be physically rebuilt.

The subject of digital filtering is somewhat beyond our scope in this course. Several excellent books are available on the subject.<sup>3</sup>

### 3.6 SIGNAL DISTORTION OVER A COMMUNICATION CHANNEL

A signal transmitted over a channel is distorted because of various channel imperfections. The nature of signal distortion will now be studied.

#### Linear Distortion

We shall first consider linear time-invariant channels. Signal distortion can be caused over such a channel by nonideal characteristics of either the magnitude, the phase, or both. We can identify the effects these nonidealities will have on a pulse  $g(t)$  transmitted through such a channel. Let the pulse exist over the interval  $(a, b)$  and be zero outside this interval. We recall the discussion in Sec. 3.1.1 about the marvelous balance of the Fourier spectrum. The components of the Fourier spectrum of the pulse have such a perfect and delicate balance of magnitudes and phases that they add up precisely to the pulse  $g(t)$  over the interval  $(a, b)$  and to zero outside this interval. The transmission of  $g(t)$  through an ideal channel that satisfies the conditions of distortionless transmission also leaves this balance undisturbed, because a distortionless channel multiplies each component by the same factor and delays each component by the same amount of time. Now, if the amplitude response of the channel is not ideal [that is,  $|H(\omega)|$  is not equal to a constant], this delicate balance will be disturbed, and the sum of all the components cannot be zero outside the interval  $(a, b)$ . In short, the pulse will spread out (see the following example). The same thing happens if the channel phase characteristic is not ideal, that is,  $\theta_h(\omega) \neq -\omega t_d$ . Thus, spreading, or **dispersion**, of the pulse will occur if either the amplitude response or the phase response, or both, are nonideal.

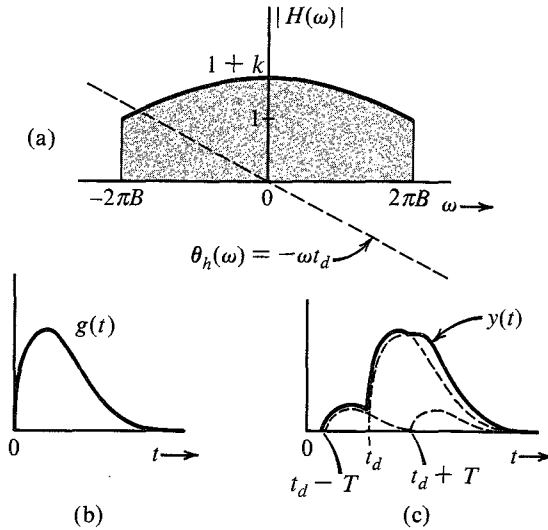
This type of distortion is undesirable in a TDM system, because pulse spreading causes interference with a neighboring pulse and consequently with a neighboring channel (crosstalk). For an FDM system, this type of distortion causes distortion (dispersion) in each multiplexed signal, but no interference occurs with a neighboring channel. This is because in FDM, each of the multiplexed signals occupies a band not occupied by any other signal. The amplitude and phase nonidealities of a channel will distort the spectrum of each signal, but because they are all nonoverlapping, no interference occurs among them.

---

**EXAMPLE 3.17** A low-pass filter (Fig. 3.33a) transfer function  $H(\omega)$  is given by

$$H(\omega) = \begin{cases} (1 + k \cos T\omega)e^{-j\omega t_d} & |\omega| < 2\pi B \\ 0 & |\omega| > 2\pi B \end{cases} \quad (3.60)$$

A pulse  $g(t)$  band-limited to  $B$  Hz (Fig. 3.33b) is applied at the input of this filter. Find the output  $y(t)$ .



**Figure 3.33** Pulse is dispersed when it passes through a system that is not distortionless.

This filter has ideal phase and nonideal magnitude characteristics. Because  $g(t) \leftrightarrow G(\omega)$ ,  $y(t) \leftrightarrow Y(\omega)$  and

$$\begin{aligned} Y(\omega) &= G(\omega)H(\omega) \\ &= G(\omega)(1 + k \cos T\omega)e^{-j\omega t_d} \\ &= G(\omega)e^{-j\omega t_d} + k[G(\omega) \cos T\omega]e^{-j\omega t_d} \end{aligned} \quad (3.61)$$

Using the time-shifting property and Eqs. (3.30a) and (3.32), we have

$$y(t) = g(t - t_d) + \frac{k}{2}[g(t - t_d - T) + g(t - t_d + T)] \quad (3.62)$$

The output is actually  $g(t) + (k/2)[g(t - T) + g(t + T)]$  delayed by  $t_d$ . It consists of  $g(t)$  and its echoes shifted by  $\pm t_d$ . The dispersion of the pulse caused by its echoes is evident from Fig. 3.33c. Ideal amplitude but nonideal phase response of  $H(\omega)$  has a similar effect (see Prob. 3.6-1).

### Distortion Caused by Channel Nonlinearities

Until now we considered the channel to be linear. This approximation is valid only for small signals. For large amplitudes, nonlinearities cannot be ignored. A general discussion of nonlinear systems is beyond our scope. Here we shall consider a simple case of a memoryless nonlinear channel where the input  $g$  and the output  $y$  are related by some nonlinear equation,

$$y = f(g)$$

The right-hand side of this equation can be expanded in a McLaurin's series as

$$y(t) = a_0 + a_1 g(t) + a_2 g^2(t) + a_3 g^3(t) + \cdots + a_k g^k(t) + \cdots$$

Recall the result in Sec. 3.3.6 (convolution) that if the bandwidth of  $g(t)$  is  $B$  Hz, then the bandwidth of  $g^k(t)$  is  $kB$  Hz. Hence, the bandwidth of  $y(t)$  is  $kB$  Hz. Consequently, the output spectrum spreads well beyond the input spectrum, and the output signal contains new frequency components not contained in the input signal. In broadcast communication, we need to amplify signals at very high power levels, where high-efficiency amplifiers (class C) are desirable. Unfortunately, these amplifiers are nonlinear, and their use to amplify signals causes distortion. This is one of the serious problems in AM signals. However, FM signals are not affected by nonlinear distortion, as shown in Chapter 5. If a signal is transmitted over a nonlinear channel, the nonlinearity not only distorts the signal, but also causes interference with other signals on the channel because of its spectral dispersion (spreading). The spectral dispersion will cause a serious interference problem in FDM systems (but not in TDM systems).

**EXAMPLE 3.18** The input  $x(t)$  and the output  $y(t)$  of a certain nonlinear channel are related as

$$y(t) = x(t) + 0.001x^2(t)$$

Find the output signal  $y(t)$  and its spectrum  $Y(\omega)$  if the input signal is  $x(t) = (1000/\pi) \operatorname{sinc}(1000t)$ . Verify that the bandwidth of the output signal is twice that of the input signal. This is the result of signal squaring. Can the signal  $x(t)$  be recovered (without distortion) from the output  $y(t)$ ?

Since

$$x(t) = \frac{1000}{\pi} \operatorname{sinc}(1000t)$$

$$X(\omega) = \operatorname{rect}\left(\frac{\omega}{2000}\right)$$

We have

$$y(t) = x(t) + 0.001x^2(t) = \frac{1000}{\pi} \operatorname{sinc}(1000t) + \frac{1000}{\pi^2} \operatorname{sinc}^2(1000t)$$

$$Y(\omega) = \operatorname{rect}\left(\frac{\omega}{2000}\right) + 0.316 \Delta\left(\frac{\omega}{4000}\right)$$

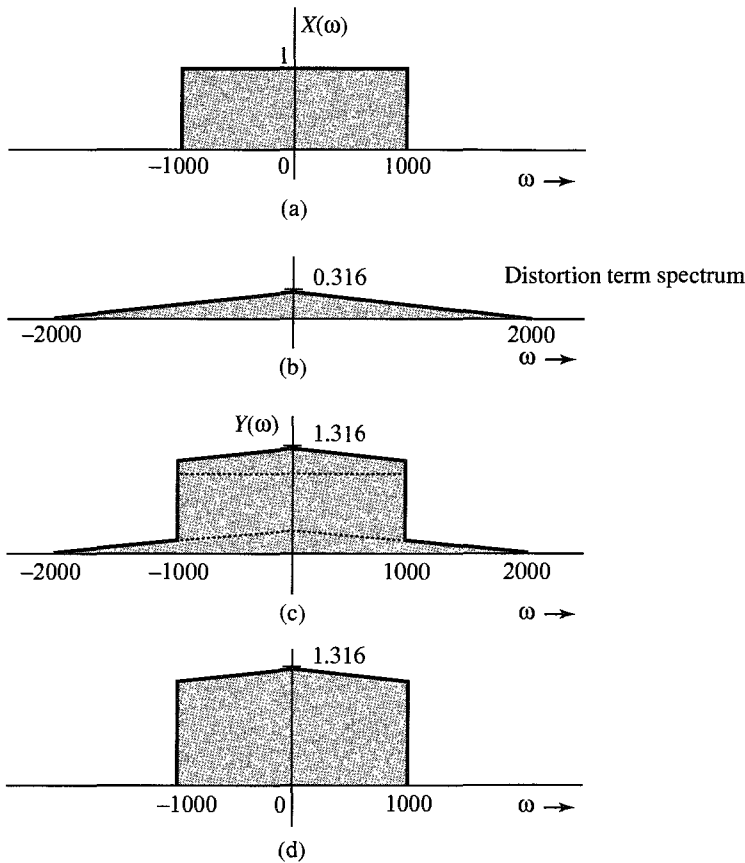
Observe that  $0.316 \operatorname{sinc}^2(1000t)$  is the unwanted (distortion) term in the received signal. Figure 3.34a shows the input (desired) signal spectrum  $X(\omega)$ ; Fig. 3.34b shows the spectrum of the undesired (distortion) term; and Fig. 3.34c shows the received signal spectrum  $Y(\omega)$ . We make the following observations:

1. The bandwidth of the received signal  $y(t)$  is twice that of the input signal  $x(t)$  (because of signal squaring).
2. The received signal contains the input signal  $x(t)$  plus an unwanted signal  $(1000/\pi) \operatorname{sinc}^2(1000t)$ . The spectra of these two signals are shown in Fig. 3.34a and b. Figure 3.34c shows  $Y(\omega)$ , the spectrum of the received signal. Note that the desired signal and the distortion signal spectra overlap, and it is impossible to recover the signal  $x(t)$  from the received signal  $y(t)$  without some distortion.
3. We can reduce the distortion by passing the received signal through a low-pass filter of bandwidth 1000 rad/s. The spectrum of the output of this filter is shown in Fig. 3.34d.



Observe that the output of this filter is the desired input signal  $x(t)$  with some residual distortion.

4. We have an additional problem of interference with other signals if the input signal  $x(t)$  is frequency-division multiplexed along with several other signals on this channel. This means that several signals occupying nonoverlapping frequency bands are transmitted simultaneously on the same channel. Spreading of the spectrum  $X(\omega)$  outside its original band of 1000 rad/s will interfere with the signal in the band of 1000 to 2000 rad/s. Thus, in addition to the distortion of  $x(t)$ , we also have an interference with the neighboring band.
5. If  $x(t)$  were a digital signal consisting of a pulse train, each pulse would be distorted, but there would be no interference with the neighboring pulses. Moreover even with distorted pulses, data can be received without loss because digital communication can withstand considerable pulse distortion without loss of information. Thus, if this channel were used to transmit a TDM signal consisting of two interleaved pulse trains, the data in the two trains would be recovered at the receiver.



**Figure 3.34** Signal distortion caused by nonlinear operation. (a) Desired (input) signal spectrum. (b) Spectrum of the unwanted signal (distortion) in the received signal. (c) Spectrum of the received signal. (d) Spectrum of the received signal after low-pass filtering.

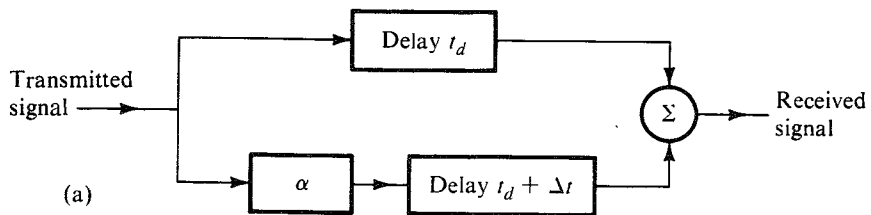
### Distortion Caused by Multipath Effects

A multipath transmission takes place when a transmitted signal arrives at the receiver by two or more paths of different delays. For example, if a signal is transmitted over a cable that has impedance irregularities (mismatching) along the path, the signal will arrive at the receiver in the form of a direct wave plus various reflections with various delays. In radio links, the signal can be received by direct path between the transmitting and the receiving antennas and also by reflections from other objects, such as hills, buildings, and so on. In long-distance radio links using the ionosphere, similar effects occur because of one-hop and multihop paths. In each of these cases, the transmission channel can be represented as several channels in parallel, each with a different relative attenuation and a different time delay. Let us consider the case of only two paths: one with a unity gain and a delay  $t_d$ , and the other with a gain  $\alpha$  and a delay  $t_d + \Delta t$ , as shown in Fig. 3.35a. The transfer functions of the two paths are given by  $e^{-j\omega t_d}$  and  $\alpha e^{-j\omega(t_d+\Delta t)}$ , respectively. The overall transfer function of such a channel is  $H(\omega)$ , given by

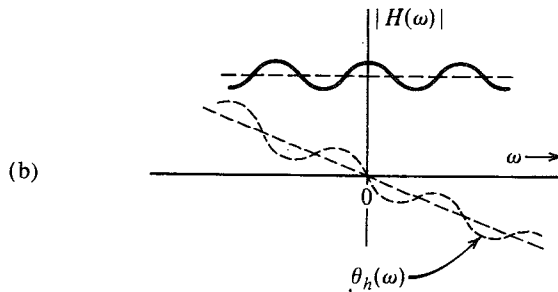
$$\begin{aligned} H(\omega) &= e^{-j\omega t_d} + \alpha e^{-j\omega(t_d+\Delta t)} \\ &= e^{-j\omega t_d} (1 + \alpha e^{-j\omega \Delta t}) \end{aligned} \tag{3.63a}$$

$$\begin{aligned} &= e^{-j\omega t_d} (1 + \alpha \cos \omega \Delta t - j\alpha \sin \omega \Delta t) \\ &= \underbrace{\sqrt{1 + \alpha^2 + 2\alpha \cos \omega \Delta t}}_{|H(\omega)|} e^{-j \underbrace{[\omega t_d + \tan^{-1} \frac{\alpha \sin \omega \Delta t}{1 + \alpha \cos \omega \Delta t}]}_{\theta_h(\omega)}} \end{aligned} \tag{3.63b}$$

Both the magnitude and the phase characteristics of  $H(\omega)$  are periodic in  $\omega$  with a period of  $2\pi/\Delta t$  (Fig. 3.35b). The multipath transmission, therefore, causes nonidealities in



(a)



(b)

Figure 3.35 Multipath transmission.

the magnitude and the phase characteristics of the channel and will cause linear distortion (pulse dispersion), as discussed earlier. If, for instance, the gains of the two paths are very close, that is,  $\alpha \approx 1$ , the signals received by the two paths can very nearly cancel each other at certain frequencies, where their phases are  $\pi$  rad apart (signal annihilation by destructive interference). Equation (3.63b) shows that at frequencies where  $\omega = n\pi/\Delta t$  ( $n$  odd),  $\cos \omega\Delta t = -1$ , and  $|H(\omega)| \approx 0$  when  $\alpha \approx 1$ . These frequencies are the multipath null frequencies. At frequencies  $\omega = n\pi/\Delta t$  ( $n$  even), the two signals interfere constructively to enhance the gain. Such channels cause **frequency-selective fading** of transmitted signals. Such distortion can be partly corrected by using the tapped delay-line equalizer, as shown in Prob. 3.6-2. These equalizers are useful in several applications in communications, discussed in Chapters 6 and 7.

### Fading Channels

Thus far, the channel characteristics were assumed to be constant with time. In practice, we encounter channels whose transmission characteristics vary with time. These include troposcatter channels and channels using the ionosphere for radio reflection to achieve long-distance communication. The time variations of the channel properties arise because of semiperiodic and random changes in the propagation characteristics of the medium. The reflection properties of the ionosphere, for example, are related to meteorological conditions that change seasonally, daily, and even from hour to hour, much the same way as does the weather. Periods of sudden storms also occur. Hence, the effective channel transfer function varies semiperiodically and randomly, causing random attenuation of the signal. This phenomenon is known as **fading**. One way to reduce the effects of fading is to use **automatic gain control (AGC)**.\*

Fading may be strongly frequency dependent where different frequency components are affected unequally. Such fading is known as frequency-selective fading and can cause serious problems in communication. Multipath propagation can cause frequency-selective fading.

## 3.7 SIGNAL ENERGY AND ENERGY SPECTRAL DENSITY

The energy  $E_g$  of a signal  $g(t)$  is defined as the area under  $|g(t)|^2$ . We can also determine the signal energy from its Fourier transform  $G(\omega)$  through Parseval's theorem.

### Parseval's Theorem

Signal energy can be related to the signal spectrum  $G(\omega)$  by substituting Eq. (3.8b) in Eq. (2.2):

$$E_g = \int_{-\infty}^{\infty} g(t)g^*(t) dt = \int_{-\infty}^{\infty} g(t) \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} G^*(\omega)e^{-j\omega t} d\omega \right] dt$$

Here, we used the fact that  $g^*(t)$ , being the conjugate of  $g(t)$ , can be expressed as the conjugate of the right-hand side of Eq. (3.8b). Now, interchanging the order of integration yields

$$\begin{aligned} E_g &= \frac{1}{2\pi} \int_{-\infty}^{\infty} G^*(\omega) \left[ \int_{-\infty}^{\infty} g(t)e^{-j\omega t} dt \right] d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega)G^*(\omega) d\omega \end{aligned}$$

---

\* AGC will also suppress slow variations of the original signal.

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\omega)|^2 d\omega \quad (3.64)$$

This is the statement of the well-known Parseval's theorem. A similar result was obtained for a periodic signal and its Fourier series in Eq. (2.59). This result allows us to determine the signal energy from either the time-domain specification  $g(t)$  or the frequency-domain specification  $G(\omega)$  of the same signal.

**EXAMPLE 3.19** Verify Parseval's theorem for the signal  $g(t) = e^{-at}u(t)$  ( $a > 0$ ).

We have

$$E_g = \int_{-\infty}^{\infty} g^2(t) dt = \int_0^{\infty} e^{-2at} dt = \frac{1}{2a} \quad (3.65)$$

We now determine  $E_g$  from the signal spectrum  $G(\omega)$  given by

$$G(\omega) = \frac{1}{j\omega + a}$$

and from Eq. (3.64),

$$E_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\omega)|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{\omega^2 + a^2} d\omega = \frac{1}{2\pi a} \tan^{-1} \frac{\omega}{a} \Big|_{-\infty}^{\infty} = \frac{1}{2a}$$

which verifies Parseval's theorem.

### Energy Spectral Density (ESD)

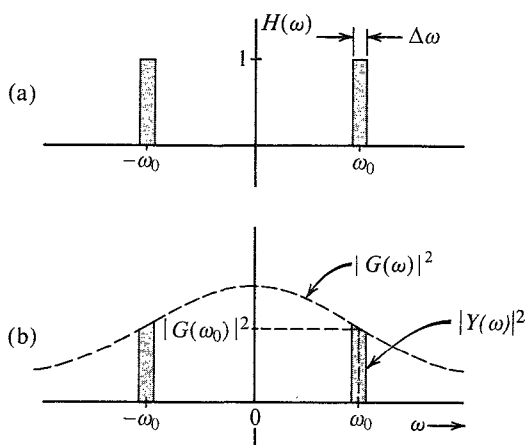
Equation (3.64) can be interpreted to mean that the energy of a signal  $g(t)$  is the result of energies contributed by all the spectral components of the signal  $g(t)$ . The contribution of a spectral component of frequency  $\omega$  is proportional to  $|G(\omega)|^2$ . To elaborate this further, consider a signal  $g(t)$  applied at the input of an ideal bandpass filter, whose transfer function  $H(\omega)$  is shown in Fig. 3.36a. This filter suppresses all frequencies except a narrow band  $\Delta\omega$  ( $\Delta\omega \rightarrow 0$ ) centered at a frequency  $\omega_0$  (Fig. 3.36b). If the filter output is  $y(t)$ , then its Fourier transform  $Y(\omega) = G(\omega)H(\omega)$ , and  $E_y$ , the energy of the output  $y(t)$ , is

$$E_y = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\omega)H(\omega)|^2 d\omega \quad (3.66)$$

Because  $H(\omega)=1$  over the passband  $\Delta\omega$ , and zero everywhere else, the intergral on the right-hand side is the sum of the two shaded areas in Fig. 3.36b, and we have (for  $\Delta\omega \rightarrow 0$ )

$$E_y = 2 \frac{1}{2\pi} |G(\omega_0)|^2 d\omega = 2|G(\omega_0)|^2 df$$

Thus,  $2|G(\omega)|^2 df$  is the energy contributed by the spectral components within the two narrow bands, each of width  $\Delta f$  Hz, centered at  $\pm\omega_0$ . Therefore, we can interpret  $|G(\omega)|^2$  as the energy per unit bandwidth (in hertz) of the spectral components of  $g(t)$  centered at frequency  $\omega$ . In other words,  $|G(\omega)|^2$  is the energy spectral density (per unit bandwidth in hertz) of  $g(t)$ . Actually the energy contributed per unit bandwidth is  $2|G(\omega)|^2$  because both the positive and the negative frequency components combine to form the components in the band  $\Delta f$ . However, for the sake of convenience we consider the positive and negative frequency components being



**Figure 3.36** Interpretation of the energy spectral density of a signal.

independent. Some authors *do* define  $2|G(\omega)|^2$  as the energy spectral density. The **energy spectral density (ESD)**  $\Psi_g(t)$  is thus defined as

$$\Psi_g(\omega) = |G(\omega)|^2 \quad (3.67)$$

and Eq. (3.64) can be expressed as

$$E_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_g(\omega) d\omega = \int_{-\infty}^{\infty} \Psi_g(\omega) df \quad (3.68)$$

From the results in Example 3.19, the ESD of the signal  $g(t) = e^{-at}u(t)$  is

$$\Psi_g(\omega) = |G(\omega)|^2 = \frac{1}{\omega^2 + a^2}$$

### Essential Bandwidth of a Signal

The spectra of most signals extend to infinity. However, because the energy of a practical signal is finite, the signal spectrum must approach 0 as  $\omega \rightarrow \infty$ . Most of the signal energy is contained within a certain band of  $B$  Hz, and the energy content of the components of frequencies greater than  $B$  Hz is negligible. We can therefore suppress the signal spectrum beyond  $B$  Hz with little effect on the signal shape and energy. The bandwidth  $B$  is called the **essential bandwidth** of the signal. The criterion for selecting  $B$  depends on the error tolerance in a particular application. We may, for instance, select  $B$  to be that band which contains 95% of the signal energy\*. This figure may be higher or lower than 95%, depending on the precision needed. Using such a criterion, we can determine the essential bandwidth of a signal. Suppression of all the spectral components of  $g(t)$  beyond the essential bandwidth results in a signal  $\hat{g}(t)$ , which is a close approximation of  $g(t)$ .<sup>†</sup> If we use the 95% criterion for the essential bandwidth, the energy of the error (the difference)  $g(t) - \hat{g}(t)$  is 5% of  $E_g$ . The following example demonstrates the bandwidth estimation procedure.

\* Essential bandwidth for a low-pass signal may also be defined as a frequency at which the value of the amplitude spectrum is a small fraction (about 5 to 10%) of its peak value. In Example 3.19, the peak of  $|G(\omega)|$  is  $1/a$ , and it occurs at  $\omega = 0$ .

† In practice the truncation is performed gradually using tapered windows in order to avoid excessive spectral leakage resulting from the abrupt truncation.<sup>5</sup>

**EXAMPLE 3.20** Estimate the essential bandwidth  $W$  rad/s of the signal  $e^{-at}u(t)$  if the essential band is required to contain 95% of the signal energy.

In this case,

$$G(\omega) = \frac{1}{j\omega + a}$$

and the ESD is

$$|G(\omega)|^2 = \frac{1}{\omega^2 + a^2}$$

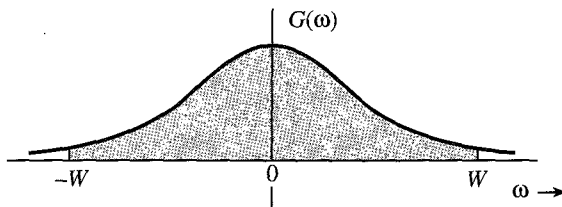
This ESD is shown in Fig. 3.37. Moreover, the signal energy  $E_g$  is  $1/2\pi$  times the area under this ESD, which has already been found to be  $1/2a$ . Let  $W$  rad/s be the essential bandwidth, which contains 95% of the total signal energy  $E_g$ . This means  $1/2\pi$  times the shaded area in Fig. 3.37 is  $0.95/2a$ , that is,

$$\begin{aligned} \frac{0.95}{2a} &= \frac{1}{2\pi} \int_{-W}^W \frac{d\omega}{\omega^2 + a^2} \\ &= \frac{1}{2\pi a} \tan^{-1} \frac{\omega}{a} \Big|_{-W}^W = \frac{1}{\pi a} \tan^{-1} \frac{W}{a} \end{aligned}$$

or

$$\frac{0.95\pi}{2} = \tan^{-1} \frac{W}{a} \implies W = 12.706a \text{ rad/s}$$

This means that the spectral components of  $g(t)$  in the band from 0 (dc) to 12.706 rad/s (2.02 Hz) contribute 95% of the total signal energy; all the remaining spectral components (in the band from 12.706 rad/s to  $\infty$ ) contribute only 5% of the signal energy.\*



**Figure 3.37** Estimating the essential bandwidth of a signal.

**EXAMPLE 3.21** Estimate the essential bandwidth of a rectangular pulse  $g(t) = \text{rect}(t/T)$  (Fig. 3.38a), where the essential bandwidth is to contain at least 90% of the pulse energy.

For this pulse, the energy  $E_g$  is

$$E_g = \int_{-\infty}^{\infty} g^2(t) dt = \int_{-T/2}^{T/2} dt = T$$

\* Note that although the ESD exists over the band  $-\infty$  to  $\infty$ , the trigonometric spectrum exists only over the band 0 to  $\infty$ . The spectrum range  $-\infty$  to  $\infty$  applies to the exponential spectrum. In practice, whenever we talk about a bandwidth, we mean it in the trigonometric sense. Hence, the essential band is from 0 to  $W$ , not  $-W$  to  $W$ .

Also because

$$\text{rect}\left(\frac{t}{T}\right) \iff T \text{ sinc}\left(\frac{\omega T}{2}\right)$$

the ESD for this pulse is

$$\Psi_g(\omega) = |G(\omega)|^2 = T^2 \text{ sinc}^2\left(\frac{\omega T}{2}\right)$$

This ESD is shown in Fig. 3.38b as a function of  $\omega T$  as well as  $fT$ , where  $f$  is the frequency in hertz. The energy  $E_W$  within the band from 0 to  $W$  rad/s is given by

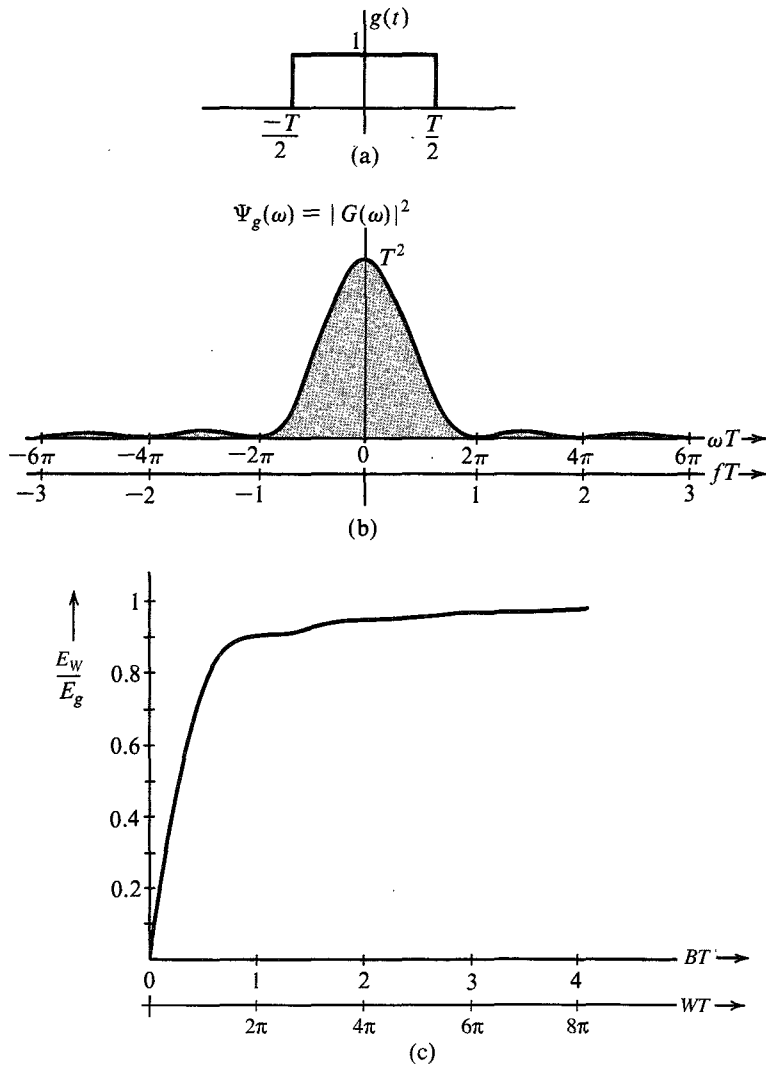


Figure 3.38 Gate function and its energy spectral density.

$$E_W = \frac{1}{2\pi} \int_{-W}^W T^2 \operatorname{sinc}^2\left(\frac{\omega T}{2}\right) d\omega$$

Setting  $\omega T = x$  in this integral so that  $d\omega = (1/T) dx$ , we obtain

$$E_W = \frac{T}{\pi} \int_0^{WT} \operatorname{sinc}^2\left(\frac{x}{2}\right) dx$$

Also because  $E_g = T$ , we have

$$\frac{E_W}{E_g} = \frac{1}{\pi} \int_0^{WT} \operatorname{sinc}^2\left(\frac{x}{2}\right) dx$$

The integral on the right-hand side is numerically computed, and the plot of  $E_W/E_g$  vs.  $WT$  is shown in Fig. 3.38c. Note that 90.28% of the total energy of the pulse  $g(t)$  is contained within the band  $W = 2\pi/T$  rad/s or  $B = 1/T$  Hz. Therefore, using the 90% criterion, the bandwidth of a rectangular pulse of width  $T$  seconds is  $1/T$  Hz. A similar result was obtained from Example 3.2.

### Energy of Modulated Signals

We have seen that modulation shifts the signal spectrum  $G(\omega)$  to the left and right by  $\omega_0$ . We now show that a similar thing happens to the ESD of the modulated signal.

Let  $g(t)$  be a baseband signal band-limited to  $B$  Hz. The amplitude-modulated signal  $\varphi(t)$  is

$$\varphi(t) = g(t) \cos \omega_0 t$$

and the spectrum (Fourier transform) of  $\varphi(t)$  is

$$\Phi(\omega) = \frac{1}{2} [G(\omega + \omega_0) + G(\omega - \omega_0)]$$

The ESD of the modulated signal  $\varphi(t)$  is  $|\Phi(\omega)|^2$ , that is,

$$\Psi_\varphi(\omega) = \frac{1}{4} |G(\omega + \omega_0) + G(\omega - \omega_0)|^2$$

If  $\omega_0 \geq 2\pi B$ , then  $G(\omega + \omega_0)$  and  $G(\omega - \omega_0)$  are nonoverlapping (see Fig. 3.39), and

$$\Psi_\varphi(\omega) = \frac{1}{4} \left[ |G(\omega + \omega_0)|^2 + |G(\omega - \omega_0)|^2 \right] \quad (3.69a)$$

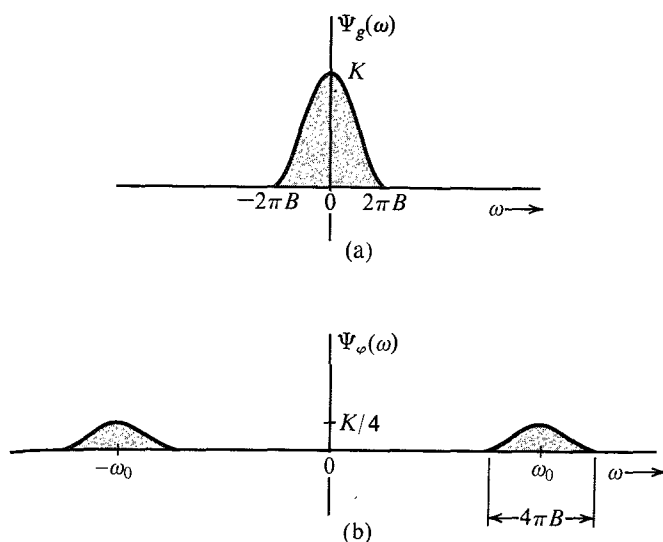
$$= \frac{1}{4} \left[ \Psi_g(\omega + \omega_0) + \Psi_g(\omega - \omega_0) \right] \quad (3.69b)$$

The ESDs of both  $g(t)$  and the modulated signal  $\varphi(t)$  are shown in Fig. 3.39. It is clear that modulation shifts the ESD of  $g(t)$  by  $\pm\omega_0$ . Observe that the area under  $\Psi_\varphi(\omega)$  is half the area under  $\Psi_g(\omega)$ . Because the energy of a signal is proportional to the area under its ESD, it follows that the energy of  $\varphi(t)$  is half the energy of  $g(t)$ , that is,

$$E_\varphi = \frac{1}{2} E_g \quad \omega_0 \geq 2\pi B \quad (3.70)$$

It may seem surprising that a signal  $\varphi(t)$ , which appears so energetic compared to  $g(t)$ , should have only half the energy of  $g(t)$ . Appearances are deceiving, as usual. The energy of a signal





**Figure 3.39** Energy spectral densities of modulating and modulated signals.

is proportional to the square of its amplitude, and higher amplitudes contribute more energy. Signal  $g(t)$  remains at higher amplitude levels most of the time. On the other hand,  $\varphi(t)$ , because of the factor  $\cos \omega_0 t$ , dips to zero amplitude levels many times, which reduces its energy.

### Time Autocorrelation Function and the Energy Spectral Density

In Chapter 2, we showed that a good measure of comparing two signals  $g(t)$  and  $z(t)$  is the correlation function  $\psi_{gz}(\tau)$  defined in Eq. (2.49). We also defined the correlation of a signal  $g(t)$  with itself [the autocorrelation function  $\psi_g(\tau)$ ] in Eq. (2.50). For a real signal  $g(t)$ , the autocorrelation function  $\psi_g(\tau)$  is given by\*

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t + \tau) dt \quad (3.71a)$$

Setting  $x = t + \tau$  in Eq. (3.71a) yields

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(x)g(x - \tau) dx$$

In this equation,  $x$  is a dummy variable and could be replaced by  $t$ . Thus,

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t \pm \tau) dt \quad (3.71b)$$

This shows that for a real  $g(t)$ , the autocorrelation function is an even function of  $\tau$ , that is,

$$\psi_g(\tau) = \psi_g(-\tau) \quad (3.72)$$

\* For a complex signal  $g(t)$ , we define

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g^*(t)g(t + \tau) dt \quad (3.71c)$$

We now show that the ESD  $\Psi_g(\omega) = |G(\omega)|^2$  is the Fourier transform of the autocorrelation function  $\psi_g(\tau)$ . Although the result is proved here for real signals, it is valid for complex signals also. Note that the autocorrelation function is a function of  $\tau$ , not  $t$ . Hence, its Fourier transform is  $\int \psi(\tau)e^{-j\omega\tau} d\tau$ . Thus,

$$\begin{aligned}\mathcal{F}[\psi_g(\tau)] &= \int_{-\infty}^{\infty} e^{-j\omega\tau} \left[ \int_{-\infty}^{\infty} g(t)g(t+\tau) dt \right] d\tau \\ &= \int_{-\infty}^{\infty} g(t) \left[ \int_{-\infty}^{\infty} g(\tau+t)e^{-j\omega\tau} d\tau \right] dt\end{aligned}$$

The inner integral is the Fourier transform of  $g(\tau+t)$ , which is  $g(\tau)$  left-shifted by  $t$ . Hence, it is given by [time-shifting property in Eq. (3.30)]  $G(\omega)e^{j\omega t}$ . Therefore,

$$\mathcal{F}[\psi_g(\tau)] = G(\omega) \int_{-\infty}^{\infty} g(t)e^{j\omega t} dt = G(\omega)G(-\omega) = |G(\omega)|^2$$

This shows that

$$\psi_g(\tau) \iff \Psi_g(\omega) = |G(\omega)|^2 \quad (3.73)$$

A careful observation of the operation of correlation shows close connection to convolution. Indeed, the autocorrelation function  $\psi_g(\tau)$  is the convolution of  $g(\tau)$  with  $g(-\tau)$  because

$$g(\tau) * g(-\tau) = \int_{-\infty}^{\infty} g(x)g[-(\tau-x)] dx = \int_{-\infty}^{\infty} g(x)g(x-\tau) dx = \psi_g(\tau)$$

Application of the time convolution property [Eq. (3.43)] to this equation yields Eq. (3.73).

**EXAMPLE 3.22** Find the time autocorrelation function of the signal  $g(t) = e^{-at}u(t)$ , and from it determine the ESD of  $g(t)$ .

In this case,

$$g(t) = e^{-at}u(t) \quad \text{and} \quad g(t-\tau) = e^{-a(t-\tau)}u(t-\tau)$$

Recall that  $g(t-\tau)$  is  $g(t)$  right-shifted by  $\tau$ , as shown in Fig. 3.40a (for positive  $\tau$ ). The autocorrelation function  $\psi_g(\tau)$  is given by the area under the product  $g(t)g(t-\tau)$  [see Eq. (3.71b)]. Therefore,

$$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t-\tau) dt = e^{a\tau} \int_{\tau}^{\infty} e^{-2at} dt = \frac{1}{2a}e^{-a\tau}$$

This is valid for positive  $\tau$ . We can perform a similar procedure for negative  $\tau$ . However, we know that for a real  $g(t)$ ,  $\psi_g(\tau)$  is an even function of  $\tau$ . Therefore,

$$\psi_g(\tau) = \frac{1}{2a}e^{-a|\tau|}$$

Figure 3.40b shows the autocorrelation function  $\psi_g(\tau)$ . The ESD  $\Psi_g(\omega)$  is the Fourier transform of  $\psi_g(\tau)$ . From Table 3.1 (pair 3), it follows that

$$\Psi_g(\omega) = \frac{1}{\omega^2 + a^2}$$

which confirms the earlier result.

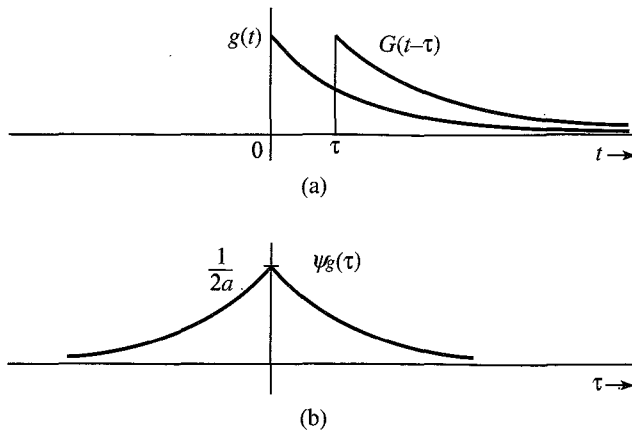


Figure 3.40 Computation of the time autocorrelation function.

### ESD of the Input and the Output

If  $g(t)$  and  $y(t)$  are the input and the corresponding output of a linear time-invariant (LTI) system, then

$$Y(\omega) = H(\omega)G(\omega)$$

Therefore,

$$|Y(\omega)|^2 = |H(\omega)|^2 |G(\omega)|^2$$

This shows that

$$\Psi_y(\omega) = |H(\omega)|^2 \Psi_g(\omega) \quad (3.74)$$

Thus, the output signal ESD is  $|H(\omega)|^2$  times the input signal ESD.

## 3.8 SIGNAL POWER AND POWER SPECTRAL DENSITY

For a power signal, a meaningful measure of its size is its power [defined in Eq. (2.4)] as the time average of the signal energy averaged over the infinite time interval. The power  $P_g$  of a real signal  $g(t)$  is given by

$$P_g = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t) dt \quad (3.75)$$

The signal power and the related concepts can be readily understood by defining a truncated signal  $g_T(t)$  as

$$g_T(t) = \begin{cases} g(t) & |t| \leq T/2 \\ 0 & |t| > T/2 \end{cases}$$

The truncated signal is shown in Fig. 3.41. The integral on the right-hand side of Eq. (3.75) is the energy of the truncated signal  $g_T(t)$ . Thus,

$$P_g = \lim_{T \rightarrow \infty} \frac{E_{g_T}}{T} \quad (3.76)$$

This equation serves as a link between power and energy. Understanding this relationship will be very helpful in understanding and relating all the power concepts to the energy concepts. Because the signal power is just the time average of energy, all the concepts and results of signal energy apply to signal power also if we modify the concepts properly by taking their time averages.

### Power Spectral Density (PSD)

If the signal  $g(t)$  is a power signal, then its power is finite, and the truncated signal  $g_T(t)$  is an energy signal as long as  $T$  is finite. If  $g_T(t) \iff G_T(\omega)$ , then from Parseval's theorem,

$$E_{g_T} = \int_{-\infty}^{\infty} g_T^2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G_T(\omega)|^2 d\omega$$

Hence,  $P_g$ , the power of  $g(t)$ , is given by

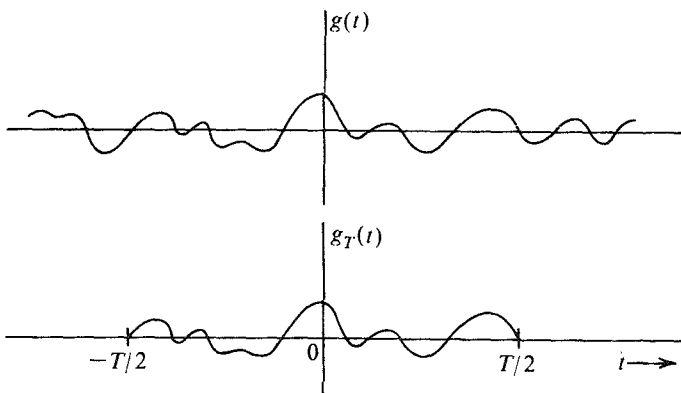
$$P_g = \lim_{T \rightarrow \infty} \frac{E_{g_T}}{T} = \lim_{T \rightarrow \infty} \frac{1}{T} \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} |G_T(\omega)|^2 d\omega \right] \quad (3.77)$$

As  $T$  increases, the duration of  $g_T(t)$  increases, and its energy  $E_{g_T}$  also increases proportionately. This means  $|G_T(\omega)|^2$  also increases with  $T$ , and as  $T \rightarrow \infty$ ,  $|G_T(\omega)|^2$  also approaches  $\infty$ . However,  $|G_T(\omega)|^2$  must approach  $\infty$  at the same rate as  $T$  because for a power signal, the right-hand side of Eq. (3.77) must converge. This convergence permits us to interchange the order of the limiting process and integration in Eq. (3.77), and we have

$$P_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} \lim_{T \rightarrow \infty} \frac{|G_T(\omega)|^2}{T} d\omega \quad (3.78)$$

We define the **power spectral density (PSD)**  $S_g(\omega)$  as

$$S_g(\omega) = \lim_{T \rightarrow \infty} \frac{|G_T(\omega)|^2}{T} \quad (3.79)$$



**Figure 3.41** Limiting process in derivation of PSD.

Consequently,\*

$$P_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_g(\omega) d\omega \quad (3.80a)$$

$$= \frac{1}{\pi} \int_0^{\infty} S_g(\omega) d\omega \quad (3.80b)$$

This result is parallel to the result [Eq. (3.68)] for energy signals. The power is  $1/2\pi$  times the area under the PSD. Observe that the PSD is the time average of the ESD of  $g_T(t)$  [Eq. (3.79)].

As is the case with ESD, the PSD is also a positive, real, and even function of  $\omega$ . If  $g(t)$  is a voltage signal, the units of PSD are volts squared per hertz. Equations (3.80) can be expressed in a more compact form using the variable  $f$  (in hertz) as

$$P_g = \int_{-\infty}^{\infty} S_g(\omega) df = 2 \int_0^{\infty} S_g(\omega) df \quad (3.81)$$

### Time-Autocorrelation Function of Power Signals

The (time) autocorrelation function  $\mathcal{R}_g(\tau)$  of a real power signal  $g(t)$  is defined as†

$$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t + \tau) dt \quad (3.82a)$$

Using the same argument as that used for energy signals [Eqs. (3.71b) and (3.72)], we can show that  $\mathcal{R}_g(\tau)$  is an even function of  $\tau$ . This means for a real  $g(t)$

$$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t - \tau) dt \quad (3.82b)$$

and

$$\mathcal{R}_g(\tau) = \mathcal{R}_g(-\tau) \quad (3.83)$$

For energy signals, the ESD  $\Psi_g(\omega)$  is the Fourier transform of the autocorrelation function  $\psi_g(\tau)$ . A similar result applies to power signals. We now show that for a power signal, the PSD  $S_g(\omega)$  is the Fourier transform of the autocorrelation function  $\mathcal{R}_g(\tau)$ . From Eq. (3.82a) and Fig. 3.41,

$$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-\infty}^{\infty} g_T(t)g_T(t + \tau) dt = \lim_{T \rightarrow \infty} \frac{\psi_{g_T}(\tau)}{T}$$

Recall that  $\psi_{g_T}(\tau) \iff |G_T(\omega)|^2$ . Hence, the Fourier transform of the preceding equation yields

$$\mathcal{R}_g(\tau) \iff \lim_{T \rightarrow \infty} \frac{|G_T(\omega)|^2}{T} = S_g(\omega) \quad (3.84)$$

\* One should use caution in using a unilateral expression such as  $P_g = 2(1/2\pi) \int_0^{\infty} S_g(\omega) d\omega$  when  $S_g(\omega)$  contains an impulse at the origin (a dc component). The impulse part should not be multiplied by the factor 2.

† For a complex  $g(t)$ , we define

$$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g^*(t)g(t + \tau) dt \quad (3.82n)$$

Although we have proved these results for a real  $g(t)$ , Eqs. (3.79), (3.80), (3.81), and (3.84) are equally valid for a complex  $g(t)$ .

The concept and relationships for signal power are parallel to those for signal energy. This is brought out in Table 3.3.

### Signal Power Is Its Mean Square Value

A glance at Eq. (3.75) shows that the signal power is the time average or mean of its squared value. In other words  $P_g$  is the mean square value of  $g(t)$ . We must remember, however, that this is a time mean, not a statistical mean (to be discussed in later chapters). Statistical means are denoted by overbars. Thus, the (statistical) mean square of a variable  $x$  is denoted by  $\overline{x^2}$ . To distinguish from this kind of mean, we shall use a wavy overline to denote a time average. Thus, the time mean square value of  $g(t)$  will be denoted by  $\overline{\overline{g^2(t)}}$ . The time averages are conventionally denoted by pointed brackets, such as  $\langle g^2(t) \rangle$ . We shall, however, use the wavy overline notation because it is much easier to associate means with a bar on top rather than the brackets. Using this notation, we see that

$$P_g = \overline{\overline{g^2(t)}} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t) dt \quad (3.85a)$$

Note that the rms value of a signal is the square root of its mean square value. Therefore,

$$[g(t)]_{\text{rms}} = \sqrt{P_g} \quad (3.85b)$$

From Eqs. (3.82), it is clear that for a real signal  $g(t)$ , the time autocorrelation function  $\mathcal{R}_g(\tau)$  is the time mean of  $g(t)g(t + \tau)$ . Thus,

$$\mathcal{R}_g(\tau) = \overline{\overline{g(t)g(t \pm \tau)}} \quad (3.86)$$

This discussion also explains why we have been using the term time autocorrelation rather than just autocorrelation. This is to distinguish clearly the present autocorrelation function (a time average) from the statistical autocorrelation function (a statistical average) to be introduced in a future chapter.

### Interpretation of Power Spectral Density

Because the PSD is a time average of the ESD of  $g(t)$ , we can argue along the lines used in the interpretation of ESD. We can readily show that the PSD  $S_g(\omega)$  represents the power per unit

Table 3.3

$E_g = \int_{-\infty}^{\infty} g^2(t) dt$	$P_g = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g^2(t) dt = \lim_{T \rightarrow \infty} \frac{E_{gT}}{T}$
$\psi_g(\tau) = \int_{-\infty}^{\infty} g(t)g(t + \tau) dt$	$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t + \tau) dt = \lim_{T \rightarrow \infty} \frac{\Psi_{gT}(\tau)}{T}$
$\Psi_g(\omega) =  G(\omega) ^2$	$S_g(\omega) = \lim_{T \rightarrow \infty} \frac{ G_T(\omega) ^2}{T} = \lim_{T \rightarrow \infty} \frac{\Psi_{gT}(\omega)}{T}$
$\psi_g(\tau) \iff \Psi_g(\omega)$	$\mathcal{R}_g(\tau) \iff S_g(\omega)$
$E_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} \Psi_g(\omega) d\omega$	$P_g = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_g(\omega) d\omega$

bandwidth (in hertz) of the spectral components at the frequency  $\omega$ . The power contributed by the spectral components within the band  $\omega_1$  to  $\omega_2$  is given by

$$\Delta P_g = \frac{1}{\pi} \int_{\omega_1}^{\omega_2} S_g(\omega) d\omega \quad (3.87)$$

### Autocorrelation Method: A Powerful Tool

For a signal  $g(t)$ , the ESD, which is equal to  $|G(\omega)|^2$ , can also be found by taking the Fourier transform of its autocorrelation function. If the Fourier transform of a signal is enough to determine its ESD, then why do we needlessly complicate our lives by talking about autocorrelation functions? The reason for following this alternate route is to lay a foundation for dealing with power signals and random signals. The Fourier transform of a power signal generally does not exist. Moreover, the luxury of finding the Fourier transform is available only for deterministic signals, which can be described as functions of time. The random message signals that occur in communication problems (e.g., random binary pulse train) cannot be described as functions of time, and it is impossible to find their Fourier transforms. However, the autocorrelation function for such signals can be determined from their statistical information. This allows us to determine the PSD (the spectral information) of such a signal. Indeed, we may consider the autocorrelation approach as the generalization of Fourier techniques to power signals and random signals. The following example of a random binary pulse train dramatically illustrates the power of this technique.

**EXAMPLE 3.23** Figure 3.42a shows a random binary pulse train  $g(t)$ . The pulse width is  $T_b/2$ , and one binary digit is transmitted every  $T_b$  seconds. A binary **1** is transmitted by the positive pulse, and a binary **0** is transmitted by the negative pulse. The two symbols are equally likely and occur randomly. We shall determine the autocorrelation function, the PSD, and the essential bandwidth of this signal.

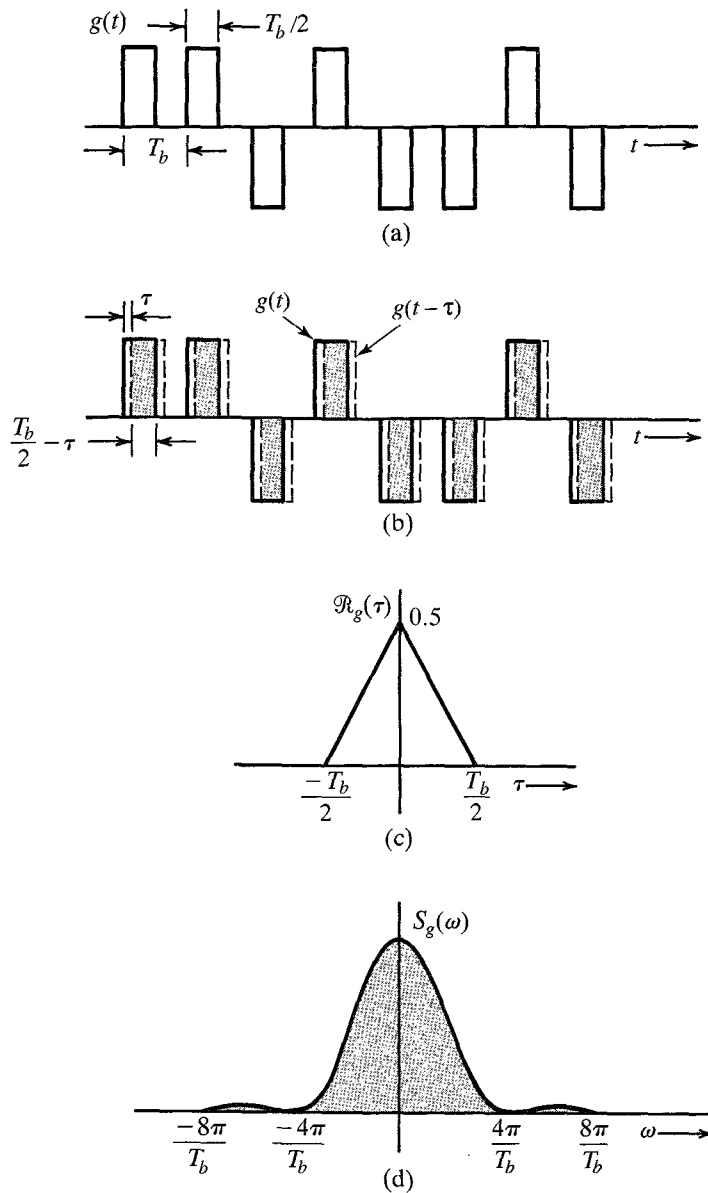
We cannot describe this signal as a function of time because the precise waveform is not known due to its random nature. We do, however, know its behavior in terms of the averages (the statistical information). The autocorrelation function, being an average parameter (time average) of the signal, is determinable from the given statistical (average) information. We have [Eq. (3.82b)]

$$\mathcal{R}_g(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} g(t)g(t - \tau) dt$$

Figure 3.42b shows  $g(t)$  by solid lines and  $g(t - \tau)$ , which is  $g(t)$  delayed by  $\tau$ , by dashed lines. To determine the integrand on the right-hand side of the above equation, we multiply  $g(t)$  with  $g(t - \tau)$ , find the area under the product  $g(t)g(t - \tau)$ , and divide it by the averaging interval  $T$ . Let there be  $N$  bits (pulses) during this interval  $T$  so that  $T = NT_b$ , and as  $T \rightarrow \infty$ ,  $N \rightarrow \infty$ . Thus,

$$\mathcal{R}_g(\tau) = \lim_{N \rightarrow \infty} \frac{1}{NT_b} \int_{-NT_b/2}^{NT_b/2} g(t)g(t - \tau) dt$$

Let us first consider the case of  $\tau < T_b/2$ . In this case there is an overlap (shown by the shaded region) between each pulse of  $g(t)$  and that of  $g(t - \tau)$ . The area under



**Figure 3.42** Autocorrelation function and power spectral density function of a random binary pulse train.

the product  $g(t)g(t - \tau)$  is  $T_b/2 - \tau$  for each pulse. Since there are  $N$  pulses during the averaging interval, then the total area under  $g(t)g(t - \tau)$  is  $N(T_b/2 - \tau)$ , and

$$\mathcal{R}_g(\tau) = \lim_{N \rightarrow \infty} \frac{1}{NT_b} \left[ N \left( \frac{T_b}{2} - \tau \right) \right]$$



$$= \frac{1}{2} \left( 1 - \frac{2\tau}{T_b} \right) \quad \tau < \frac{T_b}{2}$$

Because  $\mathcal{R}_g(\tau)$  is an even function of  $\tau$ ,

$$\mathcal{R}_g(\tau) = \frac{1}{2} \left( 1 - \frac{2|\tau|}{T_b} \right) \quad |\tau| < \frac{T_b}{2} \quad (3.88a)$$

as shown in Fig. 3.42c.

As we increase  $\tau$  beyond  $T_b/2$ , there will be overlap between each pulse and its immediate neighbor. The two overlapping pulses are equally likely to be of the same polarity or of opposite polarity. Their product is equally likely to be 1 or  $-1$  over the overlapping interval. On the average, half the pulse products will be 1 (positive-positive or negative-negative pulse combinations), and the remaining half pulse products will be  $-1$  (positive-negative or negative-positive combinations). Consequently, the area under  $g(t)g(t - \tau)$  will be zero when averaged over an infinitely large time ( $T \rightarrow \infty$ ), and

$$\mathcal{R}_g(\tau) = 0 \quad |\tau| > \frac{T_b}{2} \quad (3.88b)$$

The autocorrelation function in this case is the triangle function  $\frac{1}{2}\Delta(t/T_b)$  shown in Fig. 3.42c. The PSD is the Fourier transform of  $\frac{1}{2}\Delta(t/T_b)$ , which is found in Example 3.15 (or Table 3.1, pair 19) as

$$S_g(\omega) = \frac{T_b}{4} \text{sinc}^2 \left( \frac{\omega T_b}{4} \right) \quad (3.89)$$

The PSD is the square of the sinc function shown in Fig. 3.42d. From the result in Example 3.21, we conclude that the 90.28% of the area of this spectrum is contained within the band from 0 to  $4\pi/T_b$  rad/s, or from 0 to  $2/T_b$  Hz. Thus, the essential bandwidth may be taken as  $2/T_b$  Hz (assuming a 90% power criterion). This example illustrates dramatically how the autocorrelation function can be used to obtain the spectral information of a (random) signal where conventional means of obtaining the Fourier spectrum are not usable.

### Input and Output Power Spectral Densities

Because the PSD is a time average of ESDs, the relationship between the input and output signal PSDs of a linear time-invariant (LTI) system is similar to that of ESDs. Following the argument used for ESD [Eq. (3.74)], we can readily show that if  $g(t)$  and  $y(t)$  are the input and output signals of an LTI system with transfer function  $H(\omega)$ , then

$$S_y(\omega) = |H(\omega)|^2 S_g(\omega) \quad (3.90)$$

**EXAMPLE 3.24** A noise signal  $n_i(t)$  with PSD  $S_{n_i}(\omega) = K$  is applied at the input of an ideal differentiator (Fig. 3.43a). Determine the PSD and the power of the output noise signal  $n_o(t)$ .

The transfer function of an ideal differentiator is  $H(\omega) = j\omega$ . If the noise at the demodulator output is  $n_o(t)$ , then from Eq. (3.90),

$$S_{n_o}(\omega) = |H(\omega)|^2 S_{n_i}(\omega) = |j\omega|^2 K$$

The output PSD  $S_{n_o}(\omega)$  is parabolic, as shown in Fig. 3.43c. The output noise power  $N_o$  is  $1/2\pi$  times the area under the output PSD. Therefore,

$$N_o = \frac{1}{2\pi} \int_{-2\pi B}^{2\pi B} K \omega^2 d\omega = K \int_{-2\pi B}^{2\pi B} \omega^2 d\omega = \frac{8\pi^2 B^3 K}{3}$$

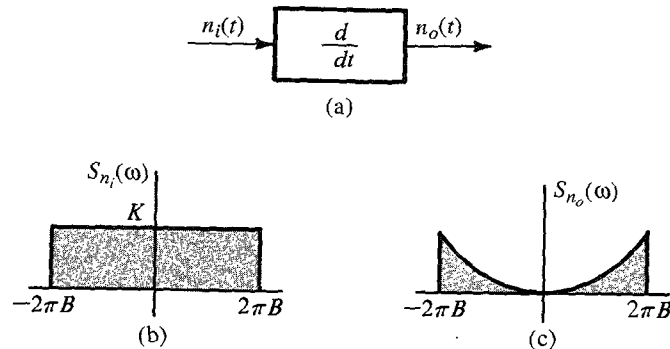


Figure 3.43 Power spectral densities at the input and the output of an ideal differentiator.

### PSD of Modulated Signals

Following the argument in deriving Eqs. (3.69) and (3.70) for energy signals, we can derive similar results for power signals by taking the time averages. We can show that for a power signal  $g(t)$ , if

$$\varphi(t) = g(t) \cos \omega_0 t$$

then the PSD  $S_\varphi(\omega)$  of the modulated signal  $\varphi(t)$  is given by

$$S_\varphi(\omega) = \frac{1}{4} [S_g(\omega + \omega_0) + S_g(\omega - \omega_0)] \quad (3.91)$$

Thus, modulation shifts the PSD of  $g(t)$  by  $\pm\omega_0$ . The power of  $\varphi(t)$  is half the power of  $g(t)$ , that is,

$$P_\varphi = \frac{1}{2} P_g \quad \omega_0 \geq 2\pi B \quad (3.92)$$

## 3.9 NUMERICAL COMPUTATION OF FOURIER TRANSFORM: THE DFT

To compute  $G(\omega)$ , the Fourier transform of  $g(t)$ , numerically, we have to use the samples of  $g(t)$ . Moreover, we can determine  $G(\omega)$  only at some finite number of frequencies. Thus, we can only compute the samples of  $G(\omega)$ . For this reason, we shall now find the relationships between the samples of  $g(t)$  and the samples of  $G(\omega)$ .

In numerical computations, the data must be finite. This means that the number of samples of  $g(t)$  and  $G(\omega)$  must be finite. In other words, we must deal with time-limited signals. If the signal is not time-limited, then we need to truncate it to make its duration finite. The same

is true of  $G(\omega)$ . To begin with, let us consider a signal  $g(t)$  of duration  $\tau$  seconds, starting at  $t = 0$ , as shown in Fig. 3.44a. However, for reasons that will become clear as we go along, we shall consider the duration of  $g(t)$  to be  $T_0$ , where  $T_0 \geq \tau$ , which makes  $g(t) = 0$  in the interval  $\tau < t \leq T_0$ , as shown in Fig. 3.44a. Clearly, this makes no difference in the computation of  $G(\omega)$ . Let us take samples of  $g(t)$  at uniform intervals of  $T_s$  seconds. There are a total of  $N_0$  samples, where

$$N_0 = \frac{T_0}{T_s} \quad (3.93)$$

Now,\*

$$\begin{aligned} G(\omega) &= \int_0^{T_0} g(t) e^{-j\omega t} dt \\ &= \lim_{T_s \rightarrow 0} \sum_{k=0}^{N_0-1} g(kT_s) e^{-j\omega k T_s} T_s \end{aligned} \quad (3.94)$$

Let us consider the samples of  $G(\omega)$  at uniform intervals of  $\omega_0$ . If  $G_r$  is the  $r$ th sample, that is,  $G_r = G(r\omega_0)$ , then from Eq. (3.94), we obtain

$$\begin{aligned} G_r &= \sum_{k=0}^{N_0-1} T_s g(kT_s) e^{-jr\omega_0 T_s k} \\ &= \sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \end{aligned} \quad (3.95)$$

where

$$g_k = T_s g(kT_s), \quad G_r = G(r\omega_0), \quad \Omega_0 = \omega_0 T_s \quad (3.96)$$

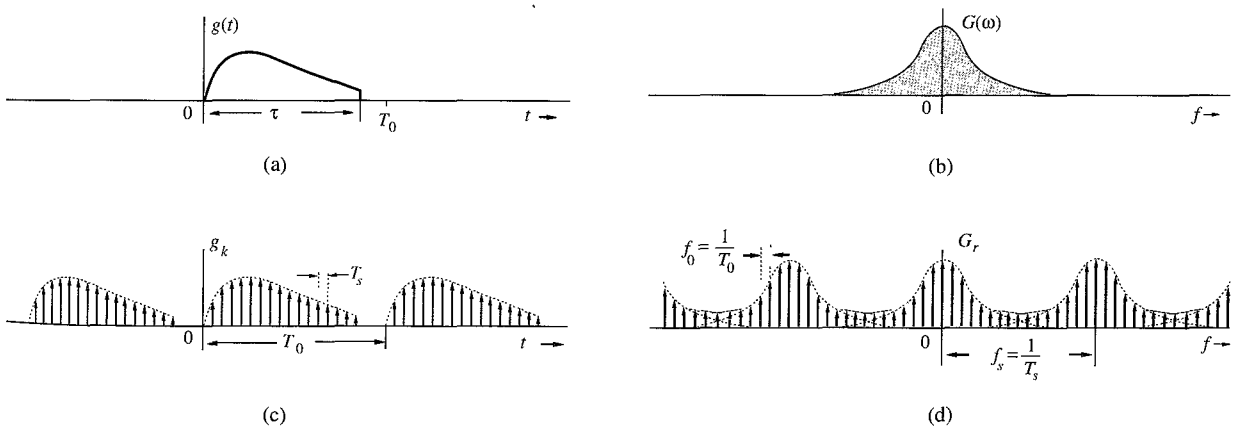


Figure 3.44 Relationship between the samples of  $g(t)$  and of  $G(\omega)$ .

\* The upper limit on the summation in Eq. (3.94) is  $N_0 - 1$  (not  $N_0$ ) because the last term in the sum starts at  $(N_0 - 1)T_s$  and covers the area under the summand up to  $N_0 T_s = T_0$ .

Thus, Eq. (3.95) relates the samples of  $g(t)$  to the samples of  $G(\omega)$ . In this derivation, we have assumed that  $T_s \rightarrow 0$ . In practice, it is not possible to make  $T_s \rightarrow 0$  because it will increase the data enormously. We strive to make  $T_s$  as small as is practicable. This will result in some computational error.

We make an interesting observation from Eq. (3.95). The samples  $G_r$  are periodic with a period of  $2\pi/\Omega_0$  samples. This follows from Eq. (3.95), which shows that  $G_{(r+2\pi/\Omega_0)} = G_r$ . Thus, only  $2\pi/\Omega_0$  number of samples  $G_r$  can be independent. Equation (3.95) shows that  $G_r$  is determined by  $N_0$  independent values  $g_k$ . Hence, for unique inverses of these equations, there can be only  $N_0$  independent sample values  $G_r$ . This means

$$N_0 = \frac{2\pi}{\Omega_0} = \frac{2\pi}{\omega_0 T_s} = \frac{2\pi N_0}{\omega_0 T_0} \quad (3.97)$$

Hence,

$$\omega_0 = \frac{2\pi}{T_0} \quad \text{and} \quad f_0 = \frac{1}{T_0} \quad (3.98)$$

Thus, the spectral sampling interval  $\omega_0$  rad/sec. (or  $f_0$  Hz) can be adjusted by a proper choice of  $T_0$ : the larger the  $T_0$ , the smaller the  $\omega_0$ . The wisdom of selecting  $T_0 \geq \tau$  is now clear. When  $T_0$  is greater than  $\tau$ , we shall have several zero-valued samples  $g_k$  in the interval from  $\tau$  to  $T_0$ . Thus, by increasing the number of zero-valued samples of  $g_k$ , we reduce  $\omega_0$  [more closely spaced samples of  $G(\omega)$ ], yielding more details of  $G(\omega)$ . This process of reducing  $\omega_0$  by the inclusion of zero-valued samples  $g_k$  is known as **zero padding**. Also, for a given sampling interval  $T_s$ , larger  $T_0$  implies larger  $N_0$ . Thus, by selecting suitably large value of  $N_0$ , we can obtain samples of  $G(\omega)$  as close as possible.

To find the inverse relationship, we multiply both sides of Eq. (3.95) by  $e^{jm\Omega_0 r}$  and sum over  $r$  as

$$\sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} = \sum_{r=0}^{N_0-1} \left[ \sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \right] e^{jm\Omega_0 r}$$

Interchanging the order of summation on the right-hand side,

$$\sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} = \sum_{k=0}^{N_0-1} g_k \left[ \sum_{r=0}^{N_0-1} e^{j(m-k)\Omega_0 r} \right] \quad (3.99)$$

In order to find the inner sum on the right-hand side, we shall now show that

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \begin{cases} N_0 & n = 0, \pm N_0, \pm 2N_0, \dots \\ 0 & \text{otherwise} \end{cases} \quad (3.100)$$

To show this, recall that  $\Omega_0 N_0 = 2\pi$  and  $e^{jn\Omega_0 k} = 1$  for  $n = 0, \pm N_0, \pm 2N_0, \dots$ , so that

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \sum_{k=0}^{N_0-1} 1 = N_0 \quad n = 0, \pm N_0, \pm 2N_0, \dots$$

To compute the sum for other values of  $n$ , we note that the sum on the left-hand side of Eq. (3.100) is a geometric progression with common ratio  $\alpha = e^{jn\Omega_0}$ . Therefore (see Appendix E),

$$\sum_{k=0}^{N_0-1} e^{jn\Omega_0 k} = \frac{e^{jn\Omega_0 N_0} - 1}{e^{jn\Omega_0} - 1} = 0 \quad e^{jn\Omega_0 N_0} = e^{j2\pi n} = 1$$

This proves Eq. (3.100). It now follows that the inner sum on the right-hand side of Eq. (3.99) is zero for  $k \neq m$ , and the sum is  $N_0$  when  $k = m$ . Therefore, the outer sum will have only one nonzero term when  $k = m$ , and it is  $N_0 g_k = N_0 g_m$ . Therefore,

$$g_m = \frac{1}{N_0} \sum_{r=0}^{N_0-1} G_r e^{jm\Omega_0 r} \quad \Omega_0 = \frac{2\pi}{N_0} \quad (3.101)$$

Equation (3.101) reveals the interesting fact that  $g_{(m+N_0)} = g_m$ . This means that the sequence  $g_k$  is also periodic with a period of  $N_0$  samples (representing the time duration  $N_0 T_s = T_0$  seconds). Moreover,  $G_r$  is also periodic with a period of  $N_0$  samples, representing a frequency interval  $N_0 \omega_0 = (T_0/T_s)(2\pi/T_0) = 2\pi/T_s = \omega_s$  rad/s. This is equal to  $1/T_s$  Hz. But  $1/T_s$  is the number of samples of  $g(t)$  per second. Thus,  $1/T_s = f_s$  is the sampling frequency (in hertz) of  $g(t)$ . This means  $G_r$  is  $N_0$ -periodic, repeating every  $f_s$  Hz. Let us summarize the results derived so far. We have proved the discrete Fourier transform (DFT) pair

$$G_r = \sum_{k=0}^{N_0-1} g_k e^{-jr\Omega_0 k} \quad (3.102a)$$

$$g_k = \frac{1}{N_0} \sum_{r=0}^{N_0-1} G_r e^{jk\Omega_0 r} \quad (3.102b)$$

where

$$\begin{aligned} g_k &= T_s g(kT_s) & G_r &= G(r\omega_0) \\ \omega_0 &= \frac{2\pi}{T_0} = 2\pi f_0 & \omega_s &= \frac{2\pi}{T_s} = 2\pi f_s \\ N_0 &= \frac{T_0}{T_s} = \frac{\omega_s}{\omega_0} = \frac{f_s}{f_0} & \Omega_0 &= \omega_0 T_s = \frac{2\pi}{N_0} \end{aligned} \quad (3.103)$$

Both the sequences  $g_k$  and  $G_r$  are periodic with a period of  $N_0$  samples. This results in  $g_k$  repeating with period  $T_0$  seconds and  $G_r$  repeating with period  $\omega_s = 2\pi/T_s$  rad/s, or  $f_s = 1/T_s$  Hz (the sampling frequency). The sampling interval of  $g_k$  is  $T_s$  seconds and the sampling interval of  $G_r$  is  $\omega_0 = 2\pi/T_0$  rad/s, or  $f_0 = 1/T_0$  Hz. This is shown in Fig. 3.44c and d. For convenience, we have used the frequency variable  $f$  (in hertz) rather than  $\omega$  (in radians per second).

We have assumed  $g(t)$  to be time-limited to  $\tau$  seconds. This makes  $G(\omega)$  non-band-limited.\* Hence, the periodic repetition of the spectra  $G_r$ , as shown in Fig. 3.44d, will cause overlapping of spectral components, resulting in error. The nature of this error, known as **aliasing error**, is explained in more detail in Chapter 6. The spectrum  $G_r$  repeats every  $f_s$  Hz. The aliasing error is reduced by increasing  $f_s$ , the repetition frequency (see Fig. 3.44d). To summarize, the computation of  $G_r$  using DFT has aliasing error when  $g(t)$  is time-limited. This error can be made as small as desired by increasing the sampling frequency  $f_s = 1/T_s$  (or reducing the sampling interval  $T_s$ ). The aliasing error is the direct result of the nonfulfillment of the requirement  $T_s \rightarrow 0$  in Eq. (3.94).

\* We can show that a signal cannot be simultaneously time-limited and band-limited. If it is one, it cannot be the other, and vice versa.<sup>3</sup>

When  $g(t)$  is not time-limited, we need to truncate it to make it time-limited. This will cause further error in  $G_r$ . This error can be reduced as much as desired by appropriately increasing the truncating interval  $T_0$ .\*

In computation of the inverse Fourier transform [by using the inverse DFT in Eq. (3.102b)] we have similar problems. If  $G(\omega)$  is band-limited,  $g(t)$  is not time-limited, and the periodic repetition of samples  $g_k$  will overlap (aliasing in the time domain). We can reduce the aliasing error by increasing  $T_0$ , the period of  $g_k$  (in seconds). This is equivalent to reducing the frequency sampling interval  $f_0 = 1/T_0$  of  $G(\omega)$ . Moreover, if  $G(\omega)$  is not band-limited, we need to truncate it. This will cause an additional error in the computation of  $g_k$ . By increasing the truncation bandwidth, we can reduce this error. In practice, (tapered) window functions are often used for truncation<sup>5</sup> in order to reduce the severity of some problems caused by straight truncation (also known as rectangular windowing).

Because  $G_r$  is  $N_0$ -periodic, we need to determine the values of  $G_r$  over any one period. It is customary to determine  $G_r$  over the range  $(0, N_0 - 1)$  rather than over the range  $(-N_0/2, N_0/2 - 1)$ . The identical remark applies to  $g_k$ .

### Choice of $T_s$ , $T_0$ , and $N_0$

In DFT computation, we first need to select suitable values for  $N_0$ ,  $T_s$ , and  $T_0$ . For this purpose we should first decide on  $B$ , the essential bandwidth of  $g(t)$ . From Fig. 3.44d, it is clear that the spectral overlapping (aliasing) occurs at the frequency  $f_s/2$  Hz. This spectral overlapping may also be viewed as the spectrum beyond  $f_s/2$  folding back at  $f_s/2$ . Hence, this frequency is also called the **folding frequency**. If the folding frequency is chosen such that the spectrum  $G(\omega)$  is negligible beyond the folding frequency, aliasing (the spectral overlapping) is not significant. Hence, the folding frequency should at least be equal to the highest significant frequency, that is, the frequency beyond which  $G(\omega)$  is negligible. We shall call this frequency the **essential bandwidth**  $B$  (in hertz). If  $g(t)$  is band-limited, then clearly, its bandwidth is identical to the essential bandwidth. Thus,

$$\frac{f_s}{2} \geq B \quad (3.104a)$$

Moreover, the sampling interval  $T_s = 1/f_s$  [Eq. (3.103)]. Hence,

$$T_s \leq \frac{1}{2B} \quad (3.104b)$$

Once we pick  $B$ , we can choose  $T_s$  according to Eq. (3.104b). Also,

$$f_0 = \frac{1}{T_0} \quad (3.105)$$

where  $f_0$  is the **frequency resolution** [separation between samples of  $G(\omega)$ ]. Hence, if  $f_0$  is given, we can pick  $T_0$  according to Eq. (3.105). Knowing  $T_0$  and  $T_s$ , we determine  $N_0$  from

$$N_0 = \frac{T_0}{T_s} \quad (3.106)$$

In general, if the signal is time-limited,  $G(\omega)$  is not band-limited, and there is aliasing in the computation of  $G_r$ . To reduce the aliasing effect, we need to increase the folding frequency,

\* The DFT relationships represent a transform in their own right, and they are exact. If, however, we identify  $g_k$  and  $G_r$  as the samples of a signal  $g(t)$  and its Fourier transform  $G(\omega)$ , respectively, then the DFT relationships are approximations because of the aliasing and truncating effects.

that is, reduce  $T_s$  (the sampling interval) as much as is practicable. If the signal is band-limited,  $g(t)$  is not time-limited, and there is aliasing (overlapping) in the computation of  $g_k$ . To reduce this aliasing, we need to increase  $T_0$ , the period of  $g_k$ . This results in reducing the frequency sampling interval  $f_0$  (in hertz). In either case (reducing  $T_s$  in the time-limited case or increasing  $T_0$  in the band-limited case), for higher accuracy, we need to increase the number of samples  $N_0$  because  $N_0 = T_0/T_s$ . There are also signals that are neither time-limited nor band-limited. In such cases, we need to reduce  $T_s$  and increase  $T_0$ .

### Points of Discontinuity

If  $g(t)$  has a jump discontinuity at a sampling point, the sample value should be taken as the average of the values on the two sides of the discontinuity because the Fourier representation at a point of discontinuity converges to the average value.

### DFT Computations Using the FFT Algorithm

The number of computations required in performing the DFT was dramatically reduced by an algorithm developed by Tukey and Cooley in 1965.<sup>6</sup> This algorithm, known as the **fast Fourier transform (FFT)**, reduces the number of computations from something on the order of  $N_0^2$  to  $N_0 \log N_0$ . To compute one sample  $G_r$  from Eq. (3.102a), we require  $N_0$  complex multiplications and  $N_0 - 1$  complex additions. To compute  $N_0$  values of  $G_r$  ( $r = 0, 1, \dots, N_0 - 1$ ), we require a total of  $N_0^2$  complex multiplications and  $N_0(N_0 - 1)$  complex additions. For large  $N_0$ , this can be prohibitively time-consuming, even for a very high-speed computer. The FFT is, thus, a life saver in signal processing applications. The FFT algorithm is simplified if we choose  $N_0$  to be a power of 2, although this is not necessary, in general. Details of the FFT can be found in any book on signal processing.<sup>3</sup>

Let us consider two examples illustrating the use of DFT in finding the Fourier transform. We shall use MATLAB to find DFT by the FFT algorithm. In the first example, the signal  $g(t) = e^{-2t}u(t)$  starts at  $t = 0$ . In the second example, we use  $g(t) = \text{rect}(t)$ , which starts at  $t = -\frac{1}{2}$ .

---

#### Computer Example C3.1

Use DFT (implemented by the FFT algorithm) to compute the Fourier transform of  $e^{-2t}u(t)$ . Plot the resulting Fourier spectra.

We first determine  $T_s$  and  $T_0$ . The Fourier transform of  $e^{-2t}u(t)$  is  $1/(j\omega + 2)$ . This low-pass signal is not band-limited. Let us take its essential bandwidth to be that frequency where  $|G(\omega)|$  becomes 1% of its peak value, which occurs at  $\omega = 0$ . Observe that

$$|G(\omega)| = \frac{1}{\sqrt{\omega^2 + 4}} \approx \frac{1}{\omega} \quad \omega \gg 2$$

Also, the peak of  $|G(\omega)|$  is at  $\omega = 0$ , where  $|G(0)| = 0.5$ . Hence, the essential bandwidth  $B$  is at  $\omega = 2\pi B$ , where

$$|G(\omega)| \approx \frac{1}{2\pi B} = 0.5 \times 0.01 \quad \Rightarrow B = \frac{100}{\pi} \text{ Hz}$$

and from Eq. (3.104b),

$$T_s \leq \frac{1}{2B} = 0.005\pi = 0.0157$$

Let us round this value down to  $T_s = 0.015625$  second so that we have 64 samples per second. The second issue is to determine  $T_0$ . The signal is not time-limited. We need to truncate it at  $T_0$  such that  $g(T_0) \ll 1$ . We shall pick  $T_0 = 4$  (eight time constants of the signal), which yields  $N_0 = T_0/T_s = 256$ . This is a power of 2. Note that there is a great deal of flexibility in determining  $T_s$  and  $T_0$ , depending on the accuracy desired and the computational capacity available. We could just as well have picked  $T_0 = 8$  and  $T_s = 1/32$ , yielding  $N_0 = 256$ , although this would have given a slightly higher aliasing error.

Because the signal has a jump discontinuity at  $t = 0$ , the first sample (at  $t = 0$ ) is 0.5, the averages of the values on the two sides of the discontinuity. The MATLAB program, which implements the DFT using the FFT algorithm is as follows:

```
Ts=1/64; T0=4; N0=T0/Ts;
t=0:Ts:Ts*(N0-1);t=t';
g=Ts*exp(-2*t);
g(1)=Ts*0.5;
G=fft(g);
[Gp,Gm]=cart2pol(real(G),imag(G));
k=0:N0-1; k=k';
w=2*pi*k/T0;
subplot(211),stem(w(1:32),Gm(1:32));
subplot(212),stem(w(1:32),Gp(1:32));
```

Because  $G_r$  is  $N_0$ -periodic,  $G_r = G_{(r+256)}$  so that  $G_{256} = G_0$ . Hence, we need to plot  $G_r$  over the range  $r = 0$  to 255 (not 256). Moreover, because of this periodicity,  $G_{-r} = G_{(-r+256)}$ ,

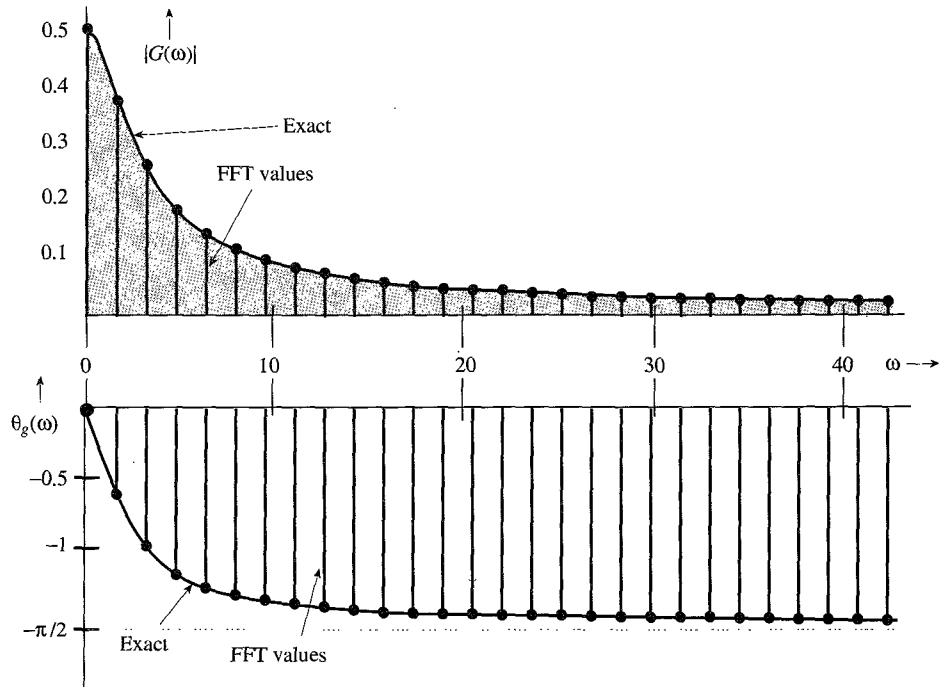


Figure 3.45 Discrete Fourier transform of an exponential signal  $e^{-2t}u(t)$ .



and the  $G_r$  over the range of  $r = -127$  to  $-1$  are identical to the  $G_r$  over the range of  $r = 129$  to  $255$ . Thus,  $G_{-127} = G_{129}$ ,  $G_{-126} = G_{130}$ ,  $\dots$ ,  $G_{-1} = G_{255}$ . In addition, because of the property of conjugate symmetry of the Fourier transform,  $G_{-r} = G_r^*$ , it follows that  $G_{129} = G_{127}^*$ ,  $G_{130} = G_{126}^*$ ,  $\dots$ ,  $G_{255} = G_1^*$ . Thus, the plots beyond  $r = N_0/2$  (128 in this case) are not necessary for real signals (because they are conjugates of  $G_r$  for  $r = 0$  to 128).

The plot of the Fourier spectra in Fig. 3.45 shows the samples of magnitude and phase of  $G(\omega)$  at the intervals of  $1/T_0 = 1/4$  Hz or  $\omega_0 = 1.5708$  rad/s. In Fig. 3.45, we have shown only the first 28 points (rather than all 128 points) to avoid too much crowding of the data.

In this example, we knew  $G(\omega)$  beforehand and hence could make intelligent choices for  $B$  (or the sampling frequency  $f_s$ ). In practice, we generally do not know  $G(\omega)$  beforehand. In fact, that is the very thing we are trying to determine. In such a case, we must make an intelligent guess for  $B$  or  $f_s$  from circumstantial evidence. We should then continue reducing the value of  $T_s$  and recomputing the transform until the result stabilizes within the desired number of significant digits.

Next, we compute the Fourier transform of  $g(t) = 8 \text{ rect}(t)$ .

### Computer Example C3.2

Use DFT (implemented by the FFT algorithm) to compute the Fourier transform of  $8 \text{ rect}(t)$ . Plot the resulting Fourier spectra.

This gate function and its Fourier transform are shown in Fig. 3.46a and b. To determine the value of the sampling interval  $T_s$ , we must first decide on the essential bandwidth  $B$ . From Fig. 3.46b, we see that  $G(\omega)$  decays rather slowly with  $\omega$ . Hence, the essential bandwidth  $B$  is rather large. For instance, at  $B = 15.5$  Hz (97.39 rad/s),  $G(\omega) = -0.1643$ , which is about 2% of the peak at  $G(0)$ . Hence, the essential bandwidth may be taken as 16 Hz. However, we shall deliberately take  $B = 4$  for two reasons: (1) to show the effect of aliasing and (2) the use of  $B > 4$  will give an enormous number of samples, which cannot be conveniently displayed on the book-sized sheet without losing sight of the essentials. Thus, we shall intentionally accept approximation in order to clarify the concepts of DFT graphically.

The choice of  $B = 4$  results in the sampling interval  $T_s = 1/2B = \frac{1}{8}$ . Looking again at the spectrum in Fig. 3.46b, we see that the choice of the frequency resolution  $f_0 = \frac{1}{4}$  Hz is reasonable. This will give four samples in each lobe of  $G(\omega)$ . In this case  $T_0 = 1/f_0 = 4$  seconds and  $N_0 = T_0/T = 32$ . The duration of  $g(t)$  is only 1 second. We must repeat it every 4 seconds ( $T_0 = 4$ ), as shown in Fig. 3.46c, and take samples every  $\frac{1}{8}$  second. This gives us 32 samples ( $N_0 = 32$ ). Also,

$$\begin{aligned} g_k &= T_s g(kT) \\ &= \frac{1}{8} g(kT) \end{aligned}$$

Since  $g(t) = 8 \text{ rect}(t)$ , the values of  $g_k$  are 1, 0, or 0.5 (at the points of discontinuity), as shown in Fig. 3.46c. In this figure,  $g_k$  is shown as a function of  $t$  as well as  $k$ , for convenience.

In the derivation of the DFT, we assumed that  $g(t)$  begins at  $t = 0$  (Fig. 3.44a), and then took  $N_0$  samples over the interval  $(0, T_0)$ . In the present case, however,  $g(t)$  begins at  $-\frac{1}{2}$ . This difficulty is easily resolved when we realize that the DFT found by this procedure is actually the DFT of  $g_k$  repeating periodically every  $T_0$  seconds. From Fig. 3.46c, it is clear that repeating the segment of  $g_k$  over the interval from  $-2$  to  $2$  seconds periodically is identical to repeating the segment of  $g_k$  over the interval from  $0$  to  $4$  seconds. Hence, the DFT of the samples taken from  $-2$  to  $2$  seconds is the same as that of the samples taken from  $0$  to  $4$  seconds. Therefore, regardless of where  $g(t)$

starts, we can always take the samples of  $g(t)$  and its periodic extension over the interval from 0 to  $T_0$ . In the present example, the 32 sample values are

$$g_k = \begin{cases} 1 & 0 \leq k \leq 3 \text{ and } 29 \leq k \leq 31 \\ 0 & 5 \leq k \leq 27 \\ 0.5 & k = 4, 28 \end{cases}$$

Observe that the last sample is at  $t = 31/8$ , not at 4, because the signal repetition starts at  $t = 4$ , and the sample at  $t = 4$  is the same as the sample at  $t = 0$ . Now,  $N_0 = 32$  and  $\Omega_0 = 2\pi/32 = \pi/16$ . Therefore [see Eq. (3.102a)],

$$G_r = \sum_{k=0}^{31} g_k e^{-jr \frac{\pi}{16} k}$$

The MATLAB program, which implements this DFT equation using the FFT algorithm, is given next. First we write a MATLAB program to generate 32 samples of  $g_k$ , and then we compute the DFT.

```
% (c32.m)
B=4; f0=1/4;
Ts=1/(2*B); T0=1/f0;
N0=T0/Ts;
k=0:N0; k=k';
for m=1:length(k)
    if k(m)>=0 & k(m)<=3, gk(m)=1; end
    if k(m)==4 & k(m)==28 gk(m)=0.5; end
    if k(m)>=5 & k(m)<=27, gk(m)=0; end
    if k(m)>=29 & k(m)<=31, gk(m)=1; end
end
gk=gk';
Gr=fft(gk);
subplot(211), stem(k, gk)
subplot(212), stem(k, Gr)
```

Figure 3.46d shows the plot of  $G_r$ .

The samples  $G_r$  are separated by  $f_0 = 1/T_0$  Hz. In this case  $T_0 = 4$ , so the frequency resolution  $f_0$  is  $\frac{1}{4}$  Hz, as desired. The folding frequency  $f_s/2 = B = 4$  Hz corresponds to  $r = N_0/2 = 16$ . Because  $G_r$  is  $N_0$ -periodic ( $N_0 = 32$ ), the values of  $G_r$  for  $r = -16$  to  $n = -1$  are the same as those for  $r = 16$  to  $n = 31$ . The DFT gives us the samples of the spectrum  $G(\omega)$ .

For the sake of comparison, Fig. 3.46d also shows the shaded curve  $8 \operatorname{sinc}(\omega/2)$ , which is the Fourier transform of  $8 \operatorname{rect}(t)$ . The values of  $G_r$  computed from the DFT equation show aliasing error, which is clearly seen by comparing the two superimposed plots. The error in  $G_2$  is just about 1.3%. However, the aliasing error increases rapidly with  $r$ . For instance, the error in  $G_6$  is about 12%, and the error in  $G_{10}$  is 33%. The error in  $G_{14}$  is a whopping 72%. The percent error increases rapidly near the folding frequency ( $r = 16$ ) because  $g(t)$  has a jump discontinuity, which makes  $G(\omega)$  decay slowly as  $1/\omega$ . Hence, near the folding frequency, the inverted tail (due to aliasing) is very nearly equal to  $G(\omega)$  itself. Moreover, the final values are the difference between the exact and the folded values (which are very close to the exact values). Hence, the percent error near the folding frequency ( $r = 16$  in this case) is very high, although the absolute error is very small. Clearly, for signals with jump discontinuities, the aliasing error near the folding frequency

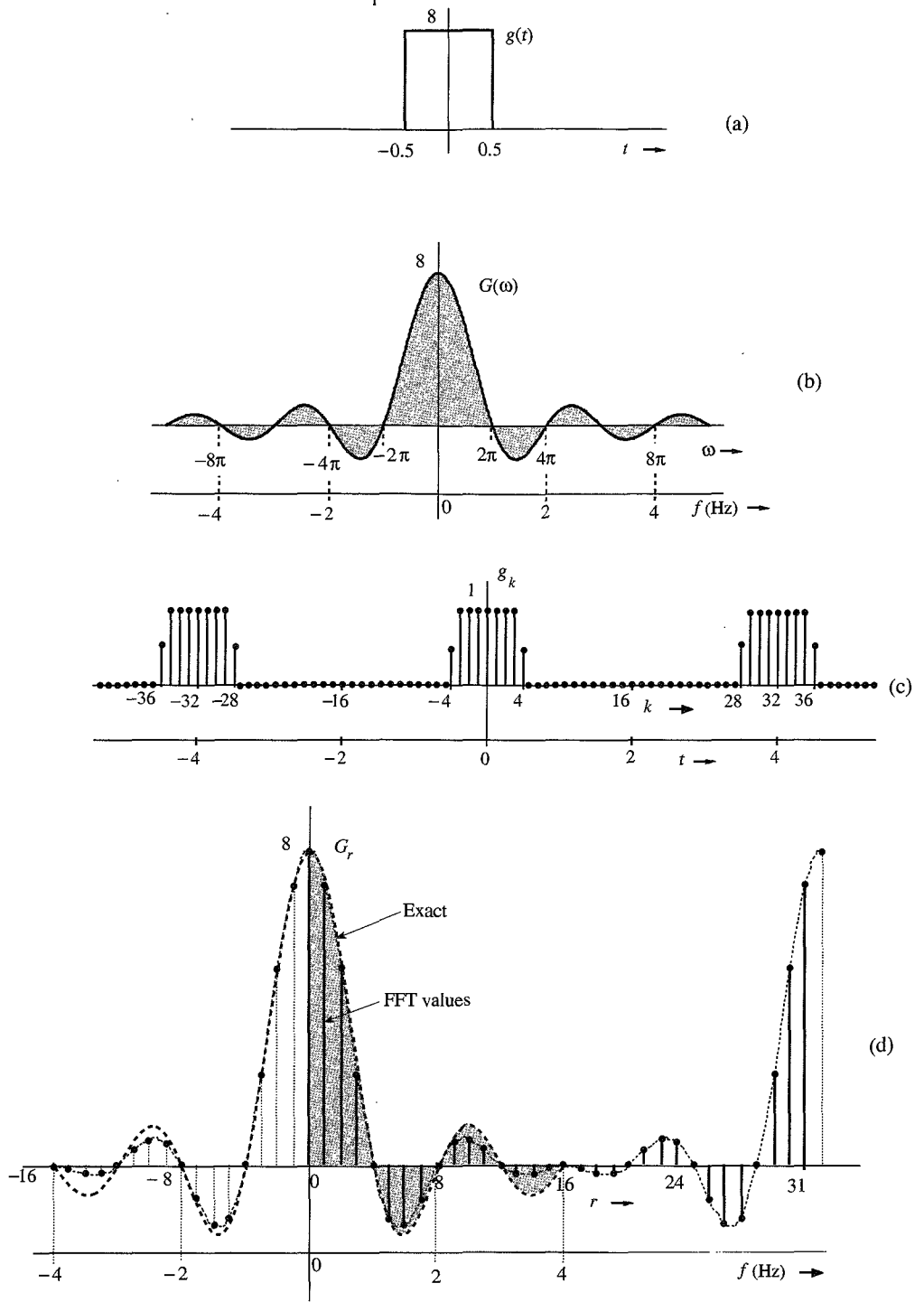


Figure 3.46 Discrete Fourier transform of a gate pulse.

will always be high (in percentage terms), regardless of the choice of  $N_0$ . To ensure a negligible aliasing error at any value  $r$ , we must make sure that  $N_0 \gg r$ . This observation is valid for all signals with jump discontinuities.

### Filtering

We generally think of filtering in terms of some hardware-oriented solution (namely, building a circuit with  $RLC$  components and operational amplifiers). However, filtering also has a software-oriented solution [a computer algorithm that yields the filtered output  $y(t)$  for a given input  $g(t)$ ]. This can be conveniently accomplished by using the DFT. If  $g(t)$  is the signal to be filtered, then  $G_r$ , the DFT of  $g_k$ , is found. The spectrum  $G_r$  is then shaped (filtered) as desired by multiplying  $G_r$  by  $H_r$ , where  $H_r$  are the samples of the filter transfer function  $H(\omega)$  [ $H_r = H(r\omega_0)$ ]. Finally, we take the IDFT of  $G_r H_r$  to obtain the filtered output  $y_k$  [ $y_k = T_s y(kT)$ ]. This procedure is demonstrated in the following example.

#### Computer Example C3.3

The signal  $g(t)$  in Fig. 3.47a is passed through an ideal low-pass filter of transfer function  $H(\omega)$ , shown in Fig. 3.47b. Using DFT, find the filter output.

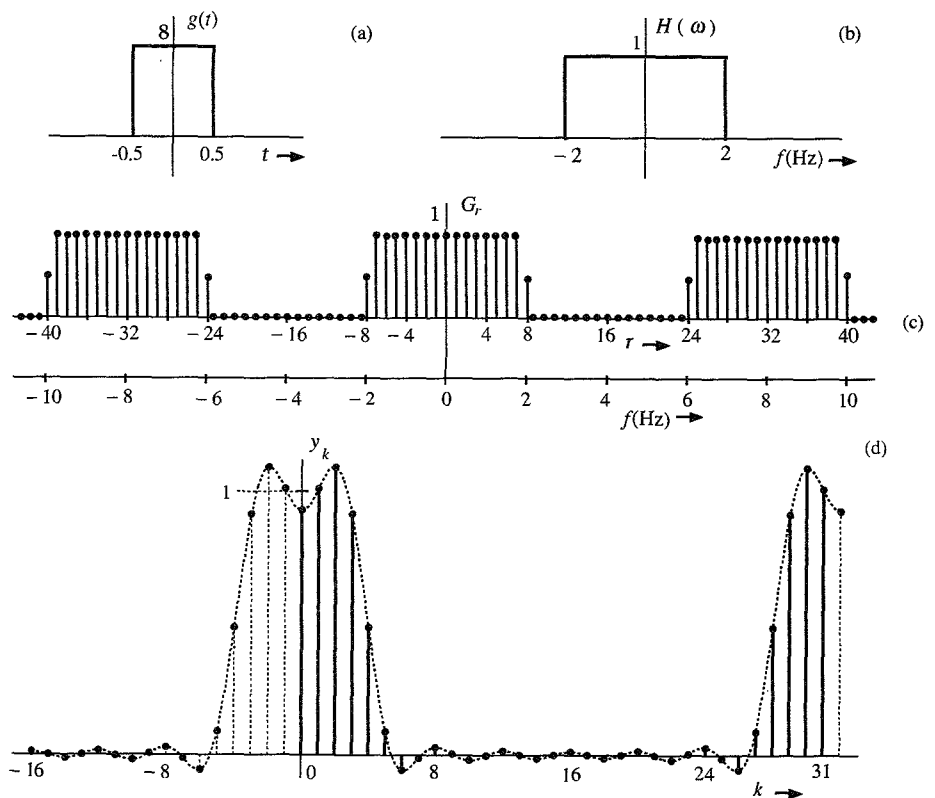


Figure 3.47 Filtering  $g(t)$  through  $H(\omega)$ .

We have already found the 32-point DFT of  $g(t)$  (see Fig. 3.46d). Next we multiply  $G_r$  by  $H_r$ . To compute  $H_r$ , we remember that in computing the 32-point DFT of  $g(t)$ , we have used  $f_0 = \frac{1}{4}$ . Because  $G_r$  is 32-periodic,  $H_r$  must also be 32-periodic with samples separated by  $\frac{1}{4}$  Hz. This means that  $H_r$  must be repeated every 8 Hz or  $16\pi$  rad/s (see Fig. 3.47c). This gives the 32 samples of  $H_r$  over  $0 \leq \omega \leq 16\pi$  as follows:

$$H_r = \begin{cases} 1 & 0 \leq r \leq 7 \text{ and } 25 \leq r \leq 31 \\ 0 & 9 \leq r \leq 23 \\ 0.5 & r = 8, 24 \end{cases}$$

We multiply  $G_r$  by  $H_r$  and take the inverse DFT. The resulting output signal is shown in Fig. 3.47d. Table 3.4 gives a printout of  $g_k$ ,  $G_r$ ,  $H_r$ ,  $Y_r$ , and  $y_k$ .

We have already found the 32-point DFT ( $G_r$ ) of  $g(t)$  in Example C3.2. The MATLAB program of Example C3.2 should be saved as an m-file, e.g., "c32.m." We can import  $G_r$  in the MATLAB environment by the command "c32." Next, we generate 32-point samples of  $H_r$ , multiply  $G_r$  by  $H_r$ , and take the inverse DFT to compute  $y_k$ . We can also find  $y_k$  by convolving  $g_k$  with  $h_k$ .

```
c32;
r=0:32; r=r';
for m=1:length(r)
    if r(m)>=0 & r(m)<=7, Hr(m)=1; end
    if r(m)>=25 & r(m)<=31, Hr(m)=1; end
    if r(m)>=9 & r(m)<=23, Hr(m)=0; end
    if r(m)==8 & r(m)==24, Hr(m)=0.5; end
```

Table 3.4

No.	$g_k$	$G_r$	$H_r$	$G_r H_r$	$y_k$
0	1	8.000	1	8.000	.9285
1	1	7.179	1	7.179	1.009
2	1	5.027	1	5.027	1.090
3	1	2.331	1	2.331	.9123
4	0.5	0.000	1	0.000	.4847
5	0	-1.323	1	-1.323	.08884
6	0	-1.497	1	-1.497	-.05698
7	0	-.8616	1	-.8616	-.01383
8	0	0.000	0.5	0.000	.02933
9	0	.5803	0	0.000	.004837
10	0	.6682	0	0.000	-.01966
11	0	.3778	0	0.000	-.002156
12	0	0.000	0	0.000	.01534
13	0	-.2145	0	0.000	.0009828
14	0	-.1989	0	0.000	-.01338
15	0	-.06964	0	0.000	.0002876
16	0	0.000	0	0.000	.01280
17	0	-.06964	0	0.000	.0002876
18	0	-.1989	0	0.000	-.01338
19	0	-.2145	0	0.000	.0009828
20	0	0.000	0	0.000	.01534
21	0	.3778	0	0.000	-.002156
22	0	.6682	0	0.000	-.01966
23	0	.5803	0	0.000	.004837
24	0	0.000	0.5	0.000	.03933
25	0	-.8616	1	-.8616	-.01383
26	0	-1.497	1	-1.497	-.05698
27	0	-.1.323	1	-1.323	.08884
28	0.5	0.000	1	0.000	.4847
29	1	2.331	1	2.331	.9123
30	1	5.027	1	5.027	1.090
31	1	7.179	1	7.179	1.009

```

end
Hr=Hr';
Yr=Gr.*Hr;
yk=ifft(Yr);
clg,stem(k,yk)

```

## REFERENCES

1. R. V. Churchill, and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed., McGraw-Hill, New York, 1978.
2. R. N. Bracewell, *Fourier Transform and Its Applications*, rev. 2nd ed., McGraw-Hill, New York, 1986.
3. B. P. Lathi, *Signal Processing and Linear Systems*, Berkeley-Cambridge Press, Carmichael, CA, 1998.
4. E. A. Guillemin, *Theory of Linear Physical Systems*, Wiley, New York, 1963.
5. F. J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, vol. 66, pp. 51–83, January 1978.
6. Tukey and Cooley, "An Algorithm for the Machine Calculation of Complex Fourier Series," *Mathematics of Computation*, Vol. 19, pp. 297–301, April, 1965.

**3.1-1** Show that the Fourier transform of  $g(t)$  may be expressed as

$$G(\omega) = \int_{-\infty}^{\infty} g(t) \cos \omega t \, dt - j \int_{-\infty}^{\infty} g(t) \sin \omega t \, dt$$

Hence, show that if  $g(t)$  is an even function of  $t$ , then

$$G(\omega) = 2 \int_0^{\infty} g(t) \cos \omega t \, dt$$

and if  $g(t)$  is an odd function of  $t$ , then

$$G(\omega) = -2j \int_0^{\infty} g(t) \sin \omega t \, dt$$

Hence, prove that:

If  $g(t)$  is:

- a real and even function of  $t$
- a real and odd function of  $t$
- an imaginary and even function of  $t$
- a complex and even function of  $t$
- a complex and odd function of  $t$

Then  $G(\omega)$  is:

- a real and even function of  $\omega$
- an imaginary and odd function of  $\omega$
- an imaginary and even function of  $\omega$
- a complex and even function of  $\omega$
- a complex and odd function of  $\omega$

**3.1-2 (a)** Show that for a real  $g(t)$ , the inverse transform, Eq. (3.8b), can be expressed as

$$g(t) = \frac{1}{\pi} \int_0^{\infty} |G(\omega)| \cos[\omega t + \theta_g(\omega)] \, d\omega$$

This is the trigonometric form of the (inverse) Fourier transform. Compare this with the compact trigonometric Fourier series.

**(b)** Express the Fourier integral (inverse Fourier transform) for  $g(t) = e^{-at} u(t)$  in the trigonometric form given in part (a).

3.1-3 If  $g(t) \iff G(\omega)$ , then show that  $g^*(t) \iff G^*(-\omega)$ .

3.1-4 From definition (3.8a), find the Fourier transforms of the signals  $g(t)$  shown in Fig. P3.1-4.

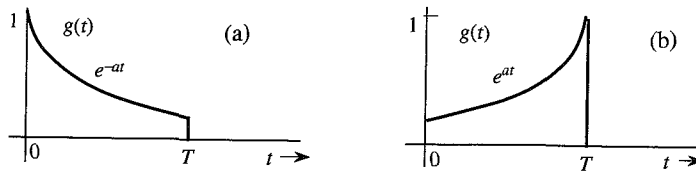


Figure P3.1-4

3.1-5 From definition (3.8a), find the Fourier transforms of the signals shown in Fig. P3.1-5.

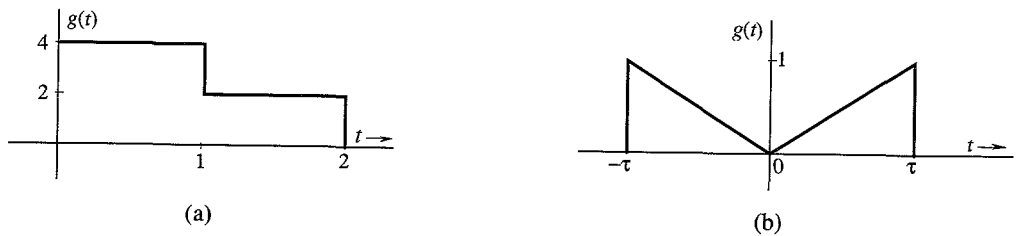


Figure P3.1-5

3.1-6 From definition (3.8b), find the inverse Fourier transforms of the spectra shown in Fig. P3.1-6.

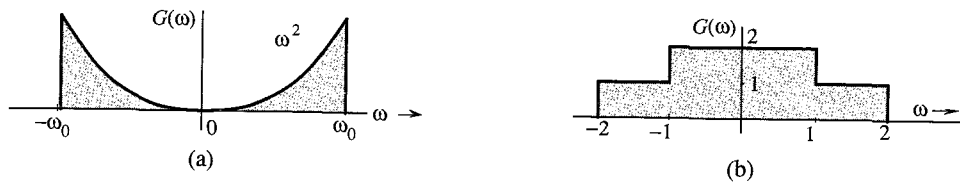


Figure P3.1-6

3.1-7 From definition (3.8b), find the inverse Fourier transforms of the spectra shown in Fig. P3.1-7.

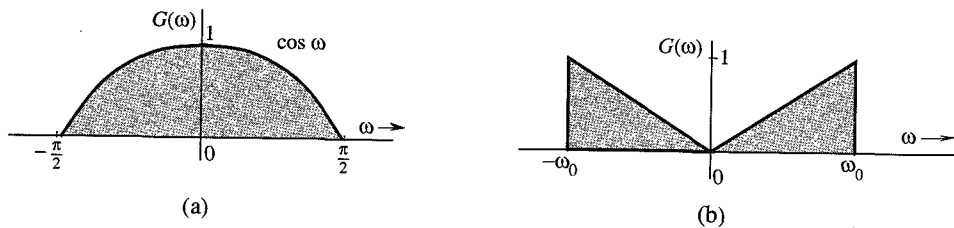


Figure P3.1-7

3.1-8 Find the inverse Fourier transform of  $G(\omega)$  for the spectra shown in Fig. P3.1-8.

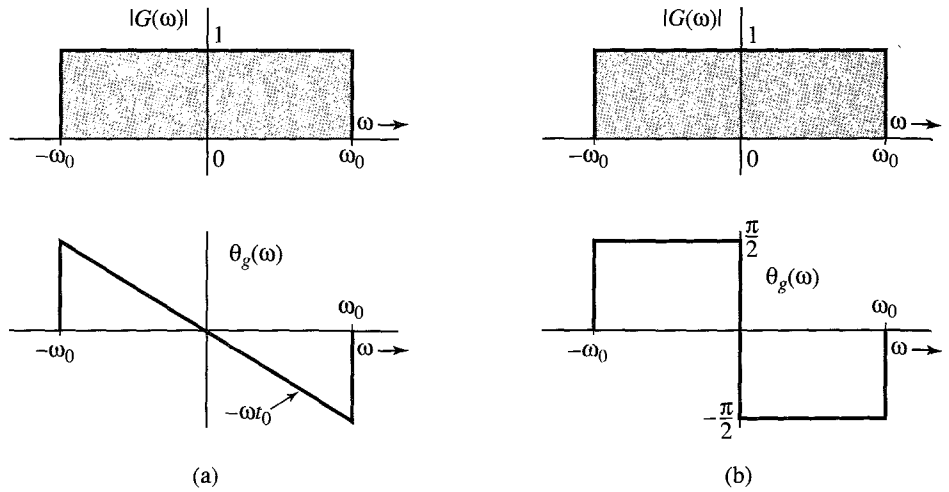


Figure P3.1-8

Hint:  $G(\omega) = |G(\omega)|e^{j\theta_g(\omega)}$ . For part (a),  $G(\omega) = 1e^{-j\omega t_0}$ ,  $|\omega| \leq \omega_0$ . For part (b),

$$G(\omega) = \begin{cases} 1e^{-j\pi/2} = -j & 0 < \omega \leq \omega_0 \\ 1e^{j\pi/2} = j & 0 > \omega \geq -\omega_0 \end{cases}$$

This problem illustrates how different phase spectra (both with the same amplitude spectrum) represent entirely different signals.

3.2-1 Sketch the following functions:

- (a)  $\text{rect}(t/2)$ ; (b)  $\Delta(3\omega/100)$ ; (c)  $\text{rect}(t - 10/8)$ ; (d)  $\text{sinc}(\pi\omega/5)$ ; (e)  $\text{sinc}[(\omega - 10\pi)/5]$ ; (f)  $\text{sinc}(t/5) \text{rect}(t/10\pi)$ . Hint:  $g(\frac{x-a}{b})$  is  $g(\frac{x}{b})$  right-shifted by  $a$ .

3.2-2 From definition (3.8a), show that the Fourier transform of  $\text{rect}(t - 5)$  is  $\text{sinc}(\omega/2)e^{-j5\omega}$ .

3.2-3 From definition (3.8b), show that the inverse Fourier transform of  $\text{rect}[(\omega - 10)/2\pi]$  is  $\text{sinc}(\pi t) e^{j10t}$ .

3.2-4 Using pairs 7 and 12 (Table 3.1) show that  $u(t) \iff \pi\delta(\omega) + 1/j\omega$ .

3.2-5 Show that  $\cos(\omega_0 t + \theta) \iff \pi[\delta(\omega + \omega_0)e^{-j\theta} + \delta(\omega - \omega_0)e^{j\theta}]$ . Hint: Express  $\cos(\omega_0 t + \theta)$  in terms of exponentials using Euler's formula.

3.3-1 Apply the symmetry property to the appropriate pair in Table 3.1 to show that:

- (a)  $0.5[\delta(t) + (j/\pi t)]$   
 (b)  $\delta(t + T) + \delta(t - T) \iff 2 \cos T\omega$  ;  
 (c)  $\delta(t + T) - \delta(t - T) \iff 2j \sin T\omega$ . Hint:  $g(-t) \iff G(-\omega)$  and  $\delta(t) = \delta(-t)$ .

3.3-2 The Fourier transform of the triangular pulse  $g(t)$  in Fig. P3.3-2a is given as

$$G(\omega) = \frac{1}{\omega^2}(e^{j\omega} - j\omega e^{j\omega} - 1)$$

Using this information, and the time-shifting and time-scaling properties, find the Fourier transforms of the signals shown in Fig. P3.3-2b, c, d, e, and f. Hint: Time inversion in  $g(t)$



results in the pulse  $g_1(t)$  in Fig. P3.3-2b; consequently  $g_1(t) = g(-t)$ . The pulse in Fig. P3.3-2c can be expressed as  $g(t - T) + g_1(t - T)$  [the sum of  $g(t)$  and  $g_1(t)$  both delayed by  $T$ ]. The pulses in Fig. P3.3-2d and e both can be expressed as  $g(t - T) + g_1(t + T)$  [the sum of  $g(t)$  delayed by  $T$  and  $g_1(t)$  advanced by  $T$ ] for some suitable choice of  $T$ . The pulse in Fig. P3.3-2f can be obtained by time-expanding  $g(t)$  by a factor of 2 and then delaying the resulting pulse by 2 seconds [or by first delaying  $g(t)$  by 1 second and then time-expanding by a factor of 2].

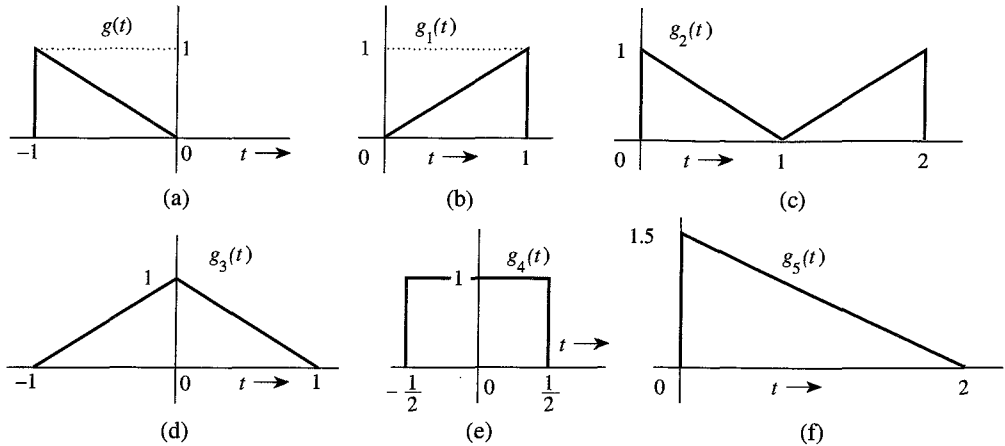


Figure P3.3-2

**3.3-3** Using only the time-shifting property and Table 3.1, find the Fourier transforms of the signals shown in Fig. P3.3-3. *Hint:* The signal in Fig. P3.3-3a is a sum of two shifted gate pulses. The signal in Fig. P3.3-3b is  $\sin t [u(t) - u(t - \pi)] = \sin t u(t) - \sin t u(t - \pi) = \sin t u(t) + \sin(t - \pi) u(t - \pi)$ . The reader should verify that the addition of these two sinusoids indeed results in the pulse in Fig. P3.3-3b. In the same way we can express the signal in Figs. P3.3-3c as  $\cos t u(t) + \sin(t - \pi/2)u(t - \pi/2)$  (verify this by sketching these signals). The signal in Fig. P3.3-3d is  $e^{-at}[u(t) - u(t - T)] = e^{-at}u(t) - e^{-aT}e^{-a(t-T)}u(t - T)$ .

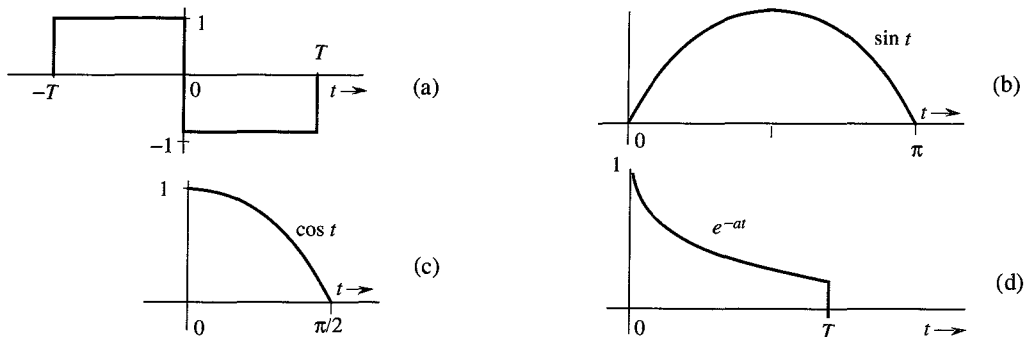


Figure P3.3-3

**3.3-4** Using the time-shifting property, show that if  $g(t) \iff G(\omega)$ , then

$$g(t + T) + g(t - T) \iff 2G(\omega) \cos T\omega$$

This is the dual of Eq. (3.35). Using this result and pairs 17 and 19 in Table 3.1, find the Fourier transforms of the signals shown in Fig. P3.3-4.

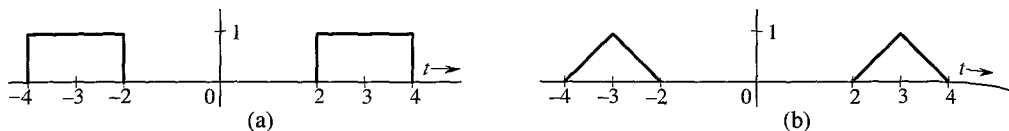


Figure P3.3-4

3.3-5 Prove the following results:

$$g(t) \sin \omega_0 t \iff \frac{1}{2j} [G(\omega - \omega_0) - G(\omega + \omega_0)]$$

$$\frac{1}{2j} [g(t + T) - g(t - T)] \iff G(\omega) \sin T\omega$$

Using the latter result and Table 3.1, find the Fourier transform of the signal in Fig. P3.3-5.

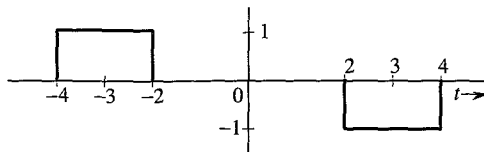


Figure P3.3-5

3.3-6 The signals in Fig. P3.3-6 are modulated signals with carrier  $\cos 10t$ . Find the Fourier transforms of these signals using the appropriate properties of the Fourier transform and Table 3.1. Sketch the amplitude and phase spectra for parts (a) and (b). *Hint:* These functions can be expressed in the form  $g(t) \cos \omega_0 t$ .

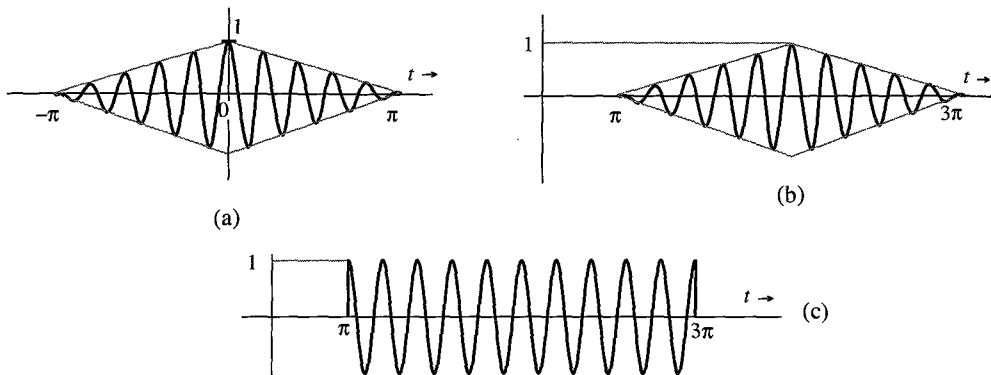


Figure P3.3-6

3.3-7 Using the frequency-shifting property and Table 3.1, find the inverse Fourier transform of the spectra shown in Fig. P3.3-7.

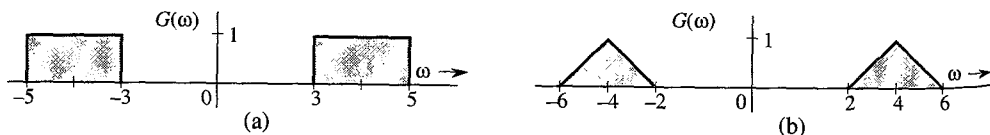


Figure P3.3-7

- 3.3-8** A signal  $g(t)$  is band-limited to  $B$  Hz. Show that the signal  $g^n(t)$  is band-limited to  $nB$  Hz. Hint:  $g^2(t) \iff [G(\omega) * G(\omega)]/2\pi$ , and so on. Use the width property of convolution.
- 3.3-9** Find the Fourier transform of the signal in Fig. P3.3-3a by three different methods:
- By direct integration using the definition (3.8a).
  - Using only pair 17 Table 3.1 and the time-shifting property.
  - Using the time-differentiation and time-shifting properties, along with the fact that  $\delta(t) \iff 1$ . Hint:  $1 - \cos 2x = 2 \sin^2 x$ .
- 3.3-10** The process of recovering a signal  $g(t)$  from the modulated signal  $g(t) \cos \omega_0 t$  is called **demodulation**. Show that the signal  $g(t) \cos \omega_0 t$  can be demodulated by multiplying it with  $2 \cos \omega_0 t$  and passing the product through a low-pass filter of bandwidth  $W$  rad/s [the bandwidth of  $g(t)$ ]. Assume  $W < \omega_0$ . Hint:  $2 \cos^2 \omega_0 t = 1 + \cos 2\omega_0 t$ . Recognize that the spectrum of  $g(t) \cos 2\omega_0 t$  is centered at  $2\omega_0$  and will be suppressed by a low-pass filter of bandwidth  $W$  rad/s.
- 3.4-1** Signals  $g_1(t) = 10^4 \text{rect}(10^4 t)$  and  $g_2(t) = \delta(t)$  are applied at the inputs of the ideal low-pass filters  $H_1(\omega) = \text{rect}(\omega/40,000\pi)$  and  $H_2(\omega) = \text{rect}(\omega/20,000\pi)$  (Fig. P3.4-1). The outputs  $y_1(t)$  and  $y_2(t)$  of these filters are multiplied to obtain the signal  $y(t) = y_1(t)y_2(t)$ .
- Sketch  $G_1(\omega)$  and  $G_2(\omega)$ .
  - Sketch  $H_1(\omega)$  and  $H_2(\omega)$ .
  - Sketch  $Y_1(\omega)$  and  $Y_2(\omega)$ .
  - Find the bandwidths of  $y_1(t)$ ,  $y_2(t)$ , and  $y(t)$ .

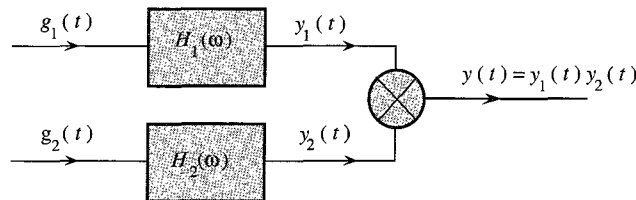


Figure P3.4-1

- 3.5-1** Consider a filter with the transfer function

$$H(\omega) = e^{-(k\omega^2 + j\omega t_0)}$$

Show that this filter is physically unrealizable by using the time-domain criterion [noncausal  $h(t)$ ] and the frequency-domain (Paley-Wiener) criterion. Can this filter be made approximately realizable by choosing a sufficiently large  $t_0$ ? Use your own (reasonable) criterion of approximate realizability to determine  $t_0$ . Hint: Use pair 22 in Table 3.1.

- 3.5-2** Show that a filter with transfer function

$$H(\omega) = \frac{2(10^5)}{\omega^2 + 10^{10}} e^{-j\omega t_0}$$

is unrealizable. Can this filter be made approximately realizable by choosing a sufficiently large  $t_0$ ? Use your own (reasonable) criterion of approximate realizability to determine  $t_0$ . Hint: Show that the impulse response is noncausal.

- 3.5-3** Determine the maximum bandwidth of a signal that can be transmitted through the low-pass  $RC$  filter in Fig. 3.27a with  $R = 1000$  and  $C = 10^{-9}$  if, over this bandwidth, the amplitude response (gain) variation is to be within 5% and the time delay variation is to be within 2%.

**3.5-4** A bandpass signal  $g(t)$  of bandwidth  $\Delta\omega = 2000$  centered at  $\omega = 10^5$  is passed through the RC filter in Example 3.16 (Fig. 3.27a) with  $RC = 10^{-3}$ . If over the passband, the variation of less than 2% in amplitude response and less than 1% in time delay is considered distortionless transmission, would  $g(t)$  be transmitted without distortion? Find the approximate expression for the output signal.

**3.6-1** A certain channel has ideal amplitude, but nonideal phase response (Fig. P3.6-1), given by

$$|H(\omega)| = 1$$

$$\theta_h(\omega) = -\omega t_0 - k \sin \omega T \quad k \ll 1$$

(a) Show that  $y(t)$ , the channel response to an input pulse  $g(t)$  band-limited to  $B$  Hz, is

$$y(t) = g(t - t_0) + \frac{k}{2}[g(t - t_0 - T) - g(t - t_0 + T)]$$

*Hint:* Use  $e^{-jk \sin \omega T} \approx 1 - jk \sin \omega T$ .

(b) Discuss how this channel will affect TDM and FDM systems from the viewpoint of interference among the multiplexed signals.

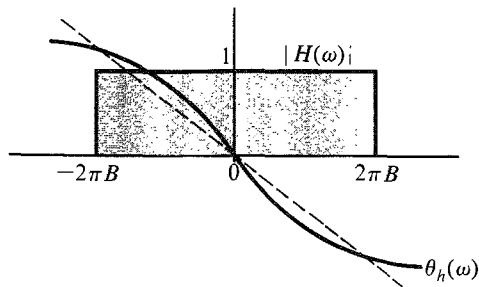


Figure P3.6-1

**3.6-2** The distortion caused by multipath transmission can be partly corrected by a tapped delay equalizer. Show that if  $\alpha \ll 1$ , the distortion in the multipath system in Fig. 3.35a can be approximately corrected if the received signal in Fig. 3.35a is passed through the tapped delay equalizer shown in Fig. P3.6-2. *Hint:* From Eq. (3.63a), it is clear that the equalizer filter transfer function should be  $H_{eq}(\omega) = 1/(1 + \alpha e^{-j\omega\Delta t})$ . Use the fact that  $1/(1-x) = 1 + x + x^2 + x^3 + \dots$  if  $x \ll 1$ .

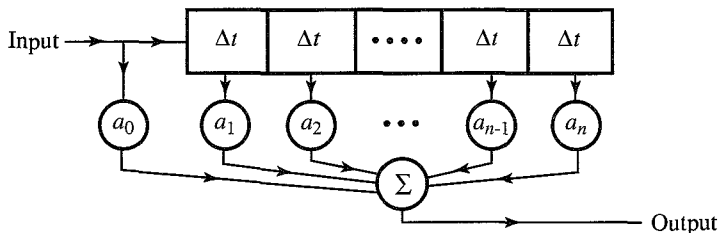


Figure P3.6-2

**3.7-1** Show that the energy of the gaussian pulse

$$g(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}}$$

is  $1/2\sigma\sqrt{\pi}$ . Verify this result by deriving the energy  $E_g$  from  $G(\omega)$  using Parseval's theorem. *Hint:* See pair 22 in Table 3.1. Use the fact that

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

3.7-2 Show that

$$\int_{-\infty}^{\infty} \text{sinc}^2(kx) dx = \frac{\pi}{k}$$

*Hint:* Recognize that the integral is the energy of  $g(t) = \text{sinc}(kt)$ . Find this energy by using Parseval's theorem.

3.7-3 Generalize Parseval's theorem to show that for real, Fourier transformable signals  $g_1(t)$  and  $g_2(t)$

$$\int_{-\infty}^{\infty} g_1(t)g_2(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_1(-\omega)G_2(\omega) d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_1(\omega)G_2(-\omega) d\omega$$

3.7-4 Show that

$$\int_{-\infty}^{\infty} \text{sinc}(2\pi Bt - m\pi) \text{sinc}(2\pi Bt - n\pi) dt = \begin{cases} 0 & m \neq n \\ \frac{1}{2B} & m = n \end{cases}$$

*Hint:* Recognize that

$$\text{sinc}(2\pi Bt - k\pi) = \text{sinc}\left[2\pi B\left(t - \frac{k}{2B}\right)\right] \iff \frac{1}{2B} \text{rect}\left(\frac{\omega}{4\pi B}\right) e^{-j\omega k/2B}$$

Use this fact and the result in Prob. 3.7-3 to show that

$$\int_{-\infty}^{\infty} \text{sinc}(2\pi Bt - m\pi) \text{sinc}(2\pi Bt - n\pi) dt = \frac{1}{8\pi B^2} \int_{-2\pi B}^{2\pi B} e^{j(n-m)/2B\omega} d\omega$$

The desired result follows from this integral.

3.7-5 For the signal

$$g(t) = \frac{2a}{t^2 + a^2}$$

determine the essential bandwidth  $B$  Hz of  $g(t)$  such that the energy contained in the spectral components of  $g(t)$  of frequencies below  $B$  Hz is 99% of the signal energy  $E_g$ . *Hint:* Determine  $G(\omega)$  by applying the symmetry property [Eq. (3.24)] to pair 3 of Table 3.1.

3.7-6 A low-pass signal  $g(t)$  is applied to a squaring device. The squarer output  $g^2(t)$  is applied to a unity gain ideal low-pass filter of bandwidth  $\Delta f$  Hz (Fig. P3.7-6). Show that if  $\Delta f$  is very small ( $\Delta f \rightarrow 0$ ), the filter output is a dc signal of amplitude  $2E_g\Delta f$ , where  $E_g$  is the energy of  $g(t)$ . *Hint:* The output  $y(t)$  is a dc signal because its spectrum  $Y(\omega)$  is concentrated at  $\omega = 0$  from  $-\Delta\omega$  to  $\Delta\omega$  with  $\Delta\omega \rightarrow 0$  (impulse at the origin). If  $g^2(t) \iff A(\omega)$ , and  $y(t) \iff Y(\omega)$ , then  $Y(\omega) \approx [4\pi A(0)\Delta f]\delta(\omega)$ .

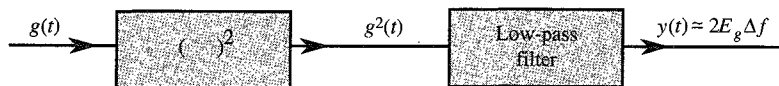


Figure P3.7-6

- 3.8-1** Show that the autocorrelation function of  $g(t) = C \cos(\omega_0 t + \theta_0)$  is given by  $\mathcal{R}_g(\tau) = (C^2/2) \cos \omega_0 \tau$ , and the corresponding PSD is  $S_g(\omega) = (C^2\pi/2)[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]$ . Hence, show that for a signal  $y(t)$  given by

$$y(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos(n\omega_0 t + \theta_n)$$

the autocorrelation function and the PSD are given by

$$\mathcal{R}_y(\tau) = C_0^2 + \frac{1}{2} \sum_{n=1}^{\infty} C_n^2 \cos n\omega_0 \tau$$

$$S_y(\omega) = 2\pi C_0^2 \delta(\omega) + \frac{\pi}{2} \sum_{n=1}^{\infty} C_n^2 [\delta(\omega - n\omega_0) + \delta(\omega + n\omega_0)]$$

*Hint:* Show that if  $g(t) = g_1(t) + g_2(t)$ , then  $\mathcal{R}_g(\tau) = \mathcal{R}_{g_1}(\tau) + \mathcal{R}_{g_2}(\tau) + \mathcal{R}_{g_1 g_2}(\tau) + \mathcal{R}_{g_2 g_1}(\tau)$ , where  $\mathcal{R}_{g_1 g_2}(\tau) = \lim_{T \rightarrow \infty} (1/T) \int_{-T/2}^{T/2} g_1(t) g_2(t + \tau) dt$ . If  $g_1(t)$  and  $g_2(t)$  represent any two of the infinite terms in  $y(t)$ , then show that  $\mathcal{R}_{g_1 g_2}(\tau) = \mathcal{R}_{g_2 g_1}(\tau) = 0$ . To show this use the fact that the area under any sinusoid over a very large time interval is at most equal to the area of the half-cycle of the sinusoid.

- 3.8-2** The random binary signal  $x(t)$  shown in Fig. P3.8-2 transmits one digit every  $T_b$  seconds. A binary **1** is transmitted by a pulse  $p(t)$  of width  $T_b/2$  and amplitude  $A$ ; a binary **0** is transmitted by no pulse. The digits **1** and **0** are equally likely and occur randomly. Determine the autocorrelation function  $\mathcal{R}_x(\tau)$  and the PSD  $S_x(\omega)$ .

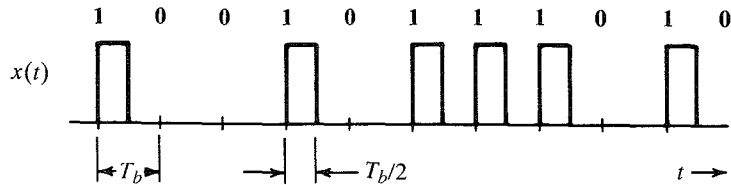


Figure P3.8-2

- 3.8-3** Find the mean square value (or power) of the output voltage  $y(t)$  of the RC network shown in Fig. 3.27a with  $RC = 1$  if the input voltage PSD  $S_x(\omega)$  is given by: (a)  $K$ ; (b)  $\text{rect}(\omega/2)$ ; (c)  $[\delta(\omega+1) + \delta(\omega-1)]$ . In each case calculate the power (mean square value) of the input signal  $x(t)$ .
- 3.8-4** Find the mean square value (or power) of the output voltage  $y(t)$  of the system shown in Fig. P3.8-4 if the input voltage PSD  $S_x(\omega) = \text{rect}(\omega/2)$ . Calculate the power (mean square value) of the input signal  $x(t)$ .

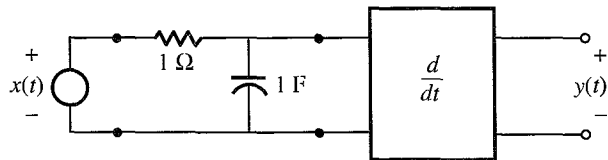
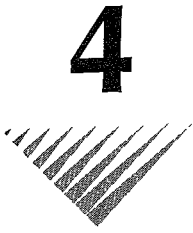


Figure P3.8-4



# 4 AMPLITUDE (LINEAR) MODULATION

**M**odulation is a process that causes a shift in the range of frequencies in a signal. It is used to gain certain advantages mentioned in Chapter 1. Before discussing modulation, it is important to distinguish between communication that does not use modulation (**baseband communication**) and communication that uses modulation (**carrier communication**).

## 4.1 BASEBAND AND CARRIER COMMUNICATION

The term **baseband** is used to designate the band of frequencies of the signal delivered by the source or the input transducer (see Fig. 1.2). In telephony, the baseband is the audio band (band of voice signals) of 0 to 3.5 kHz. In television, the baseband is the video band occupying 0 to 4.3 MHz. For digital data or PCM using bipolar signaling at a rate of  $R_b$  pulses per second, the baseband is 0 to  $R_b$  Hz.

In baseband communication, baseband signals are transmitted without modulation, that is, without any shift in the range of frequencies of the signal. Because the baseband signals have sizable power at low frequencies, they cannot be transmitted over a radio link but are suitable for transmission over a pair of wires, coaxial cables, or optical fibers. Local telephone communication, short-haul pulse-code modulation (PCM) (between two exchanges), and long-distance PCM over optical fibers use baseband communication. Modulation can be helpful in utilizing the vast spectrum of frequencies available because of technological advances. By modulating several baseband signals and shifting their spectra to nonoverlapping bands, one can use all the available bandwidth through frequency division multiplexing (FDM). Long-haul communication over a radio link also requires modulation to shift the signal spectrum to higher frequencies in order to enable efficient power radiation using antennas of reasonable dimensions. Yet another use of modulation is to exchange transmission bandwidth for the SNR.

Communication that uses modulation to shift the frequency spectrum of a signal is known as **carrier communication**. In this mode, one of the basic parameters (amplitude, frequency,

or phase) of a **sinusoidal carrier** of high frequency  $\omega_c$  is varied in proportion to the baseband signal  $m(t)$ . This results in amplitude modulation (AM), frequency modulation (FM), or phase modulation (PM), respectively. The latter two types of modulation are similar, and belong to the class of modulation known as **angle modulation**. Modulation is used to transmit analog as well as digital baseband signals.

A comment about pulse-modulated signals [pulse amplitude modulation (PAM), pulse width modulation (PWM), pulse position modulation (PPM), pulse code modulation (PCM), and delta modulation (DM)] is in order here. Despite the term modulation, these signals are baseband signals. The term modulation is used here in another sense. Pulse-modulation schemes are really baseband coding schemes, and they yield baseband signals. These signals must still modulate a carrier in order to shift their spectra.

## 4.2 AMPLITUDE MODULATION: DOUBLE SIDEBAND (DSB)

Amplitude modulation is characterized by the fact that the amplitude  $A$  of the **carrier**  $A \cos(\omega_c t + \theta_c)$  is varied in proportion to the baseband (message) signal  $m(t)$ , the **modulating signal**. The frequency  $\omega_c$  and the phase  $\theta_c$  are constant. We can assume  $\theta_c = 0$  without a loss of generality. If the carrier amplitude  $A$  is made directly proportional to the modulating signal  $m(t)$ , the **modulated signal** is  $m(t) \cos \omega_c t$  (Fig. 4.1). As was seen earlier [Eq. (3.35)], this type of modulation simply shifts the spectrum of  $m(t)$  to the carrier frequency (Fig. 4.1a). Thus, if

$$m(t) \iff M(\omega)$$

then

$$m(t) \cos \omega_c t \iff \frac{1}{2}[M(\omega + \omega_c) + M(\omega - \omega_c)] \quad (4.1)$$

Recall that  $M(\omega - \omega_c)$  is  $M(\omega)$  shifted to the right by  $\omega_c$  and  $M(\omega + \omega_c)$  is  $M(\omega)$  shifted to the left by  $\omega_c$ . Thus, the process of modulation shifts the spectrum of the modulating signal to the left and the right by  $\omega_c$ . Note also that if the bandwidth of  $m(t)$  is  $B$  Hz, then, as seen from Fig. 4.1c, the bandwidth of the modulated signal is  $2B$  Hz. We also observe that the modulated signal spectrum centered at  $\omega_c$  is composed of two parts: a portion that lies above  $\omega_c$ , known as the **upper sideband (USB)**, and a portion that lies below  $\omega_c$ , known as the **lower sideband (LSB)**. Similarly, the spectrum centered at  $-\omega_c$  has upper and lower sidebands. Hence, this is a modulation scheme with double sidebands. We shall see a little later that the modulated signal in this scheme does not contain a discrete component of the carrier frequency  $\omega_c$ . For this reason it is called **double-sideband suppressed carrier (DSB-SC) modulation**.

The relationship of  $B$  to  $\omega_c$  is of interest. Figure 4.1c shows that  $\omega_c \geq 2\pi B$  in order to avoid the overlap of the spectra centered at  $\omega_c$  and  $-\omega_c$ . If  $\omega_c < 2\pi B$ , these spectra overlap and the information of  $m(t)$  is lost in the process of modulation, which makes it impossible to get back  $m(t)$  from the modulated signal  $m(t) \cos \omega_c t$ .\*

\* Practical factors may impose additional restrictions on  $\omega_c$ . For instance, in the case of broadcast applications, a radiating antenna can radiate only a narrow band without distortion. This means that to avoid distortion caused by the radiating antenna,  $\omega_c/2\pi B \gg 1$ . The broadcast band AM radio, for instance, with  $B = 5$  kHz and the band of 550 to 1600 kHz for the carrier frequency give a ratio of  $\omega_c/2\pi B$  roughly in the range of 100 to 300.



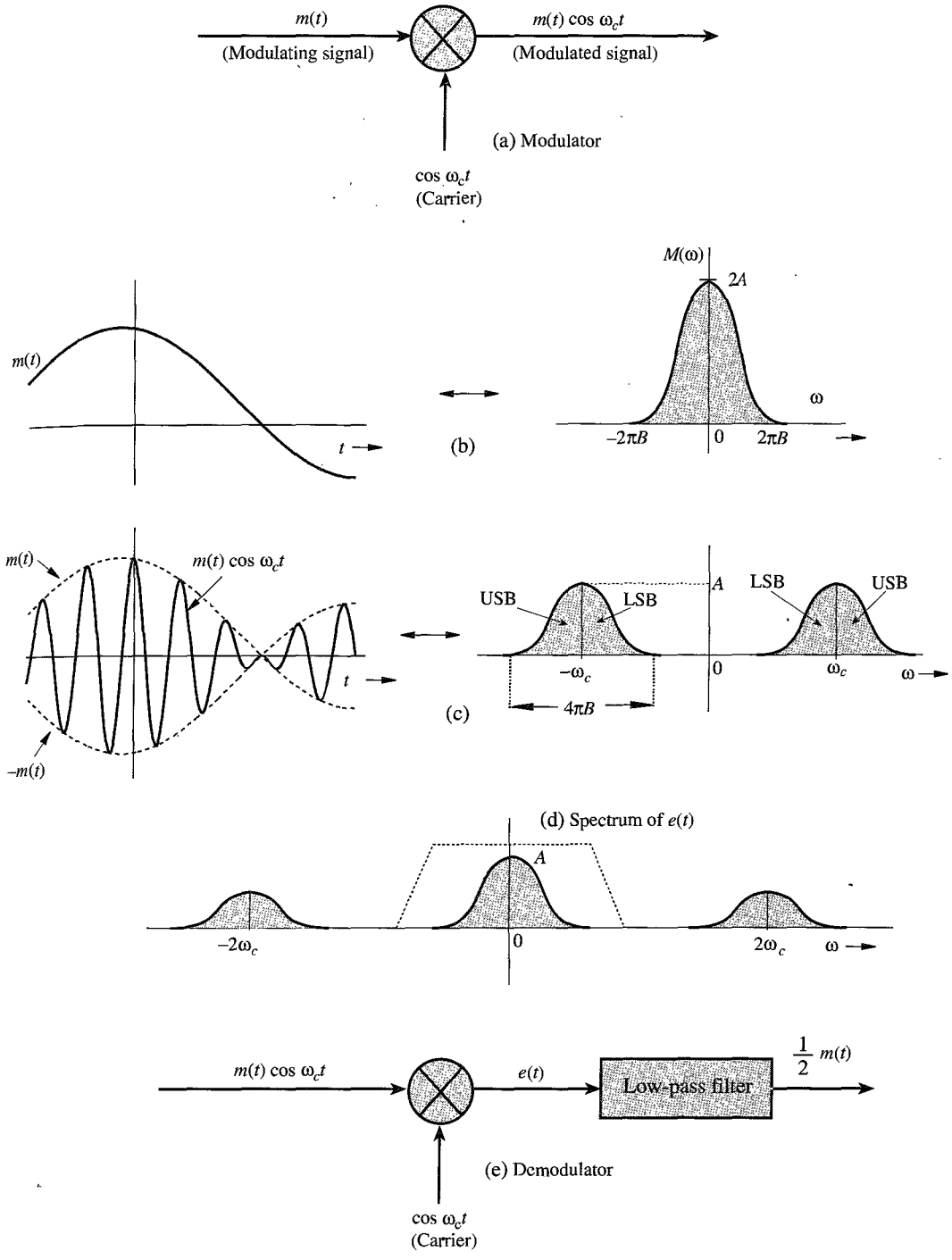


Figure 4.1 DSB-SC modulation and demodulation.

### Demodulation

The DSB-SC modulation translates or shifts the frequency spectrum to the left and the right by  $\omega_c$  (that is, at  $+\omega_c$  and  $-\omega_c$ ), as seen from Eq. (4.1). To recover the original signal  $m(t)$  from the modulated signal, it is necessary to retranslate the spectrum to its original position. The process of recovering the signal from the modulated signal (retranslating the spectrum to its original position) is referred to as **demodulation**, or **detection**. Observe that if the modulated signal spectrum in Fig. 4.1c is shifted to the left and to the right by  $\omega_c$  (and multiplied by one-half), we obtain the spectrum shown in Fig. 4.1d, which contains the desired baseband spectrum plus an unwanted spectrum at  $\pm 2\omega_c$ . The latter can be suppressed by a low pass filter. Thus, demodulation, which is almost identical to modulation, consists of multiplication of the incoming modulated signal  $m(t) \cos \omega_c t$  by a carrier  $\cos \omega_c t$  followed by a low pass filter, as shown in Fig. 4.1e. We can verify this conclusion directly in the time domain by observing that the signal  $e(t)$  in Fig. 4.1e is

$$\begin{aligned} e(t) &= m(t) \cos^2 \omega_c t \\ &= \frac{1}{2} [m(t) + m(t) \cos 2\omega_c t] \end{aligned} \quad (4.2a)$$

Therefore, the Fourier transform of the signal  $e(t)$  is

$$E(\omega) = \frac{1}{2} M(\omega) + \frac{1}{4} [M(\omega + 2\omega_c) + M(\omega - 2\omega_c)] \quad (4.2b)$$

This shows that the signal  $e(t)$  consists of two components  $(1/2)m(t)$  and  $(1/2)m(t) \cos 2\omega_c t$ , with their spectra as shown in Fig. 4.1d. The spectrum of the second component, being a modulated signal with carrier frequency  $2\omega_c$ , is centered at  $\pm 2\omega_c$ . Hence, this component is suppressed by the low pass filter in Fig. 4.1e. The desired component  $(1/2)M(\omega)$ , being a low pass spectrum (centered at  $\omega = 0$ ), passes through the filter unharmed, resulting in the output  $(1/2)m(t)$ . We can get rid of the inconvenient fraction  $1/2$  in the output by using a carrier  $2 \cos \omega_c t$  instead of  $\cos \omega_c t$ . In fact, in future, we shall often use this strategy, which does not affect general conclusions.

A possible form of low pass filter characteristics is shown (dotted) in Fig. 4.1d. This method of recovering the baseband signal is called **synchronous detection**, or **coherent detection**, where we use a carrier of exactly the same frequency (and phase) as the carrier used for modulation. Thus, for demodulation, we need to generate a local carrier at the receiver in frequency and phase coherence (synchronism) with the carrier used at the modulator.

---

**EXAMPLE 4.1** For a baseband signal  $m(t) = \cos \omega_m t$ , find the DSB-SC signal, and sketch its spectrum. Identify the USB and LSB. Verify that the DSB-SC modulated signal can be demodulated by the demodulator in Fig. 4.1e.

The case in this example is referred to as **tone modulation** because the modulating signal is a pure sinusoid, or tone,  $\cos \omega_m t$ . We shall work this problem in the frequency domain as well as the time domain in order to clarify the basic concepts of DSB-SC modulation. In the frequency domain approach, we work with the signal spectra. The spectrum of the baseband signal  $m(t) = \cos \omega_m t$  is given by

$$M(\omega) = \pi [\delta(\omega - \omega_m) + \delta(\omega + \omega_m)]$$

The spectrum consists of two impulses located at  $\pm\omega_m$ , as shown in Fig. 4.2a. The DSB-SC (modulated) spectrum, as seen from Eq. (4.1), is the baseband spectrum in Fig. 4.2a shifted to the right and the left by  $\omega_c$  (times one-half), as shown in Fig. 4.2b. This spectrum consists of impulses at  $\pm(\omega_c - \omega_m)$  and  $\pm(\omega_c + \omega_m)$ . The spectrum beyond  $\omega_c$  is the USB, and the one below  $\omega_c$  is the LSB. Observe that the DSB-SC spectrum does not have the component of the carrier frequency  $\omega_c$ . This is why it is called **suppressed carrier**.

In the time-domain approach, we work directly with signals in the time domain. For the baseband signal  $m(t) = \cos \omega_m t$ , the DSB-SC signal  $\varphi_{\text{DSB-SC}}(t)$  is

$$\begin{aligned}\varphi_{\text{DSB-SC}}(t) &= m(t) \cos \omega_c t \\ &= \cos \omega_m t \cos \omega_c t \\ &= \frac{1}{2} [\cos (\omega_c + \omega_m)t + \cos (\omega_c - \omega_m)t]\end{aligned}$$

This shows that when the baseband (message) signal is a single sinusoid of frequency  $\omega_m$ , the modulated signal consists of two sinusoids: the component of frequency  $\omega_c + \omega_m$  (the USB) and the component of frequency  $\omega_c - \omega_m$  (the LSB). Figure 4.2b shows precisely the spectrum of  $\varphi_{\text{DSB-SC}}(t)$ . Thus, each component of frequency  $\omega_m$  in the modulating signal results into two components of frequencies  $\omega_c + \omega_m$  and  $\omega_c - \omega_m$  in the modulated signal. Note the curious fact that there is no component of the carrier frequency  $\omega_c$  on the right-hand side of the preceding equation. As mentioned, this is why it is called double sideband-suppressed carrier (DSB-SC) modulation.\*

We now verify that the modulated signal  $\varphi_{\text{DSB-SC}}(t) = \cos \omega_m t \cos \omega_c t$ , when applied to the input of the demodulator in Fig. 4.1e, yields the output proportional to the desired baseband signal  $\cos \omega_m t$ . The signal  $e(t)$  in Fig. 4.1e is given by

$$\begin{aligned}e(t) &= \cos \omega_m t \cos^2 \omega_c t \\ &= \frac{1}{2} \cos \omega_m t (1 + \cos 2\omega_c t)\end{aligned}$$

The spectrum of the term  $\cos \omega_m t \cos 2\omega_c t$  is centered at  $2\omega_c$ , and will be suppressed by the low-pass filter, yielding  $\frac{1}{2} \cos \omega_m t$  as the output. We can also derive this result in the frequency domain. Demodulation causes the spectrum in Fig. 4.2b to shift left and right by  $\omega_c$  (and multiplies by one-half). This results in the spectrum shown in Fig. 4.2c. The low-pass filter suppresses the spectrum centered at  $\pm 2\omega_c$ , yielding the spectrum  $\frac{1}{2}M(\omega)$ .

## Modulators

Modulation can be achieved in several ways. We shall discuss here some important categories of modulators.

**Multiplier Modulators:** Here modulation is achieved directly by multiplying  $m(t)$  by  $\cos \omega_c t$  using an analog multiplier whose output is proportional to the product of two input

\* The term suppressed carrier does not necessarily mean absence of the spectrum at the carrier frequency. It means that there is no discrete component of the carrier frequency. This implies that the spectrum of the DSB-SC does not have impulses at  $\pm\omega_c$ , which also implies that the modulated signal  $m(t) \cos \omega_c t$  does not contain a term of the form  $k \cos \omega_c t$  [assuming that  $m(t)$  has a zero mean value].

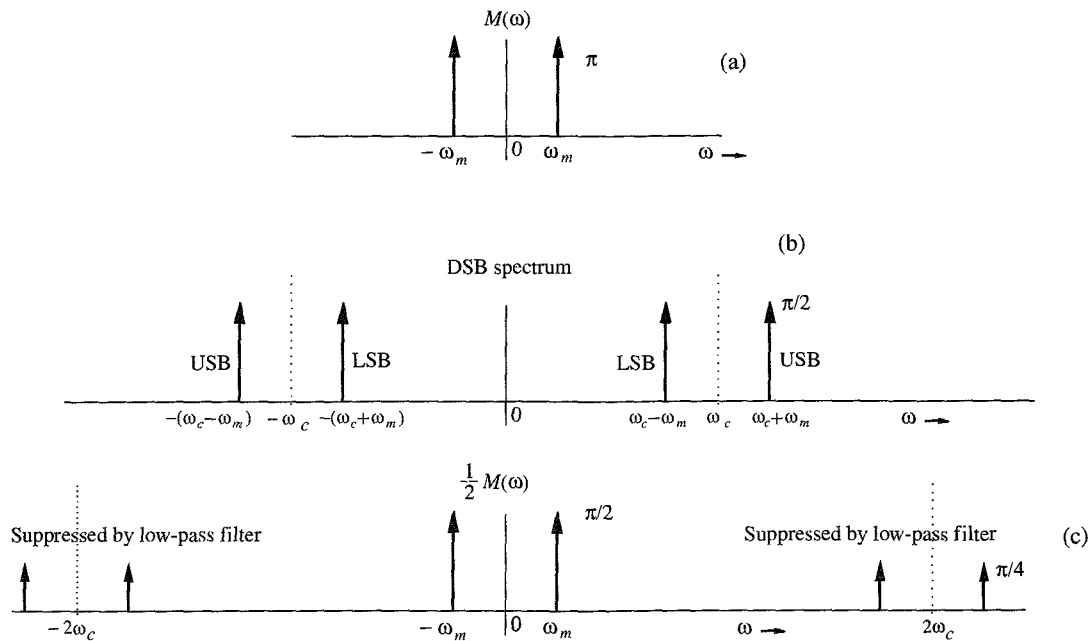


Figure 4.2 Example of DSB-SC modulation.

signals.\* It is rather difficult to maintain linearity in this kind of amplifier, and they tend to be rather expensive. It is best to avoid them if possible. For practical implementation of such modulators, see Sheingold.<sup>1</sup>

**Nonlinear Modulators:** Modulation can also be achieved by using nonlinear devices, such as a semiconductor diode or a transistor. Figure 4.3 shows one possible scheme, which uses two identical nonlinear elements shown by boxes marked NL.

Let the input-output characteristics of either of the nonlinear elements be approximated by a power series:

$$y(t) = ax(t) + bx^2(t) \quad (4.3)$$

where  $x(t)$  and  $y(t)$  are the input and the output, respectively, of the nonlinear element. The summer output  $z(t)$  in Fig. 4.3 is given by

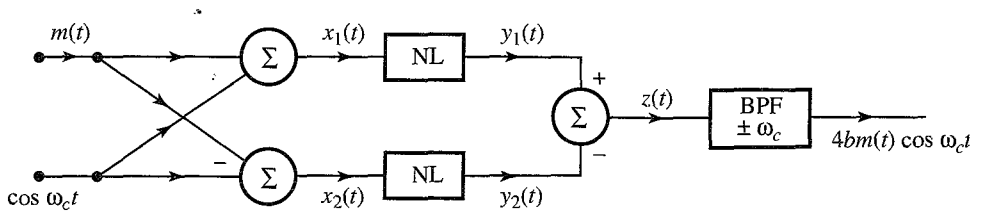
$$z(t) = y_1(t) - y_2(t) = [ax_1(t) + bx_1^2(t)] - [ax_2(t) + bx_2^2(t)]$$

Substituting the two inputs  $x_1(t) = \cos \omega_c t + m(t)$  and  $x_2(t) = \cos \omega_c t - m(t)$  in this equation yields

$$z(t) = 2am(t) + 4bm(t) \cos \omega_c t$$

\* Such a multiplier may be obtained from a variable-gain amplifier in which the gain parameter (such as the  $\beta$  of a transistor) is controlled by one of the signals, say,  $m(t)$ . When the signal  $\cos \omega_c t$  is applied at the input of this amplifier, the output is proportional to  $m(t) \cos \omega_c t$ .

Another way to multiply two signals is through logarithmic amplifiers. Here, the basic components are a logarithmic and an antilogarithmic amplifier with outputs proportional to the log and antilog of their inputs, respectively. Using two logarithmic amplifiers, we generate and add the logarithms of the two signals to be multiplied. The sum is then applied to an antilogarithmic amplifier to obtain the desired product.



**Figure 4.3** Nonlinear DSB-SC modulator.

The spectrum of  $m(t)$  is centered at the origin, whereas the spectrum of  $m(t) \cos \omega_c t$  is centered at  $\pm\omega_c$ . Consequently, when  $z(t)$  is passed through a bandpass filter tuned to  $\omega_c$ , the signal  $am(t)$  is suppressed and the desired modulated signal  $4bm(t) \cos \omega_c t$  passes through unharmed.

In this circuit there are two inputs:  $m(t)$  and  $\cos \omega_c t$ . The summer output  $z(t)$  does not contain one of the inputs, the carrier signal  $\cos \omega_c t$ . Consequently, the carrier signal does not appear at the input of the final bandpass filter. The circuit acts as a balanced bridge for one of the inputs (the carrier). Circuits which have this characteristic are called **balanced circuits**. The nonlinear modulator in Fig. 4.3 is an example of a class of modulators known as **balanced modulators**. This circuit is balanced with respect to only one input (the carrier); the other input  $m(t)$  still appears at the final bandpass filter, which must reject it. For this reason, it is called a **single balanced modulator**. A circuit balanced with respect to both inputs is called a **double balanced modulator**, of which the ring modulator (see Fig. 4.6) is an example.

**Switching Modulators:** The multiplication operation required for modulation can be replaced by a simpler switching operation if we realize that a modulated signal can be obtained by multiplying  $m(t)$  not only by a pure sinusoid but by any periodic signal  $\phi(t)$  of the fundamental radian frequency  $\omega_c$ . Such a periodic signal can be expressed by a trigonometric Fourier series as

$$\phi(t) = \sum_{n=0}^{\infty} C_n \cos(n\omega_c t + \theta_n) \quad (4.4a)$$

Hence,

$$m(t)\phi(t) = \sum_{n=0}^{\infty} C_n m(t) \cos(n\omega_c t + \theta_n) \quad (4.4b)$$

This shows that the spectrum of the product  $m(t)\phi(t)$  is the spectrum  $M(\omega)$  shifted to  $\pm\omega_c, \pm 2\omega_c, \dots, \pm n\omega_c, \dots$ . If this signal is passed through a bandpass filter of bandwidth  $2B$  Hz and tuned to  $\omega_c$ , then we get the desired modulated signal  $c_1 m(t) \cos(\omega_c t + \theta_1)$ .\*

The square pulse train  $w(t)$  in Fig. 4.4b is a periodic signal whose Fourier series was found earlier [Eq. (2.75)] as

$$w(t) = \frac{1}{2} + \frac{2}{\pi} \left( \cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \frac{1}{5} \cos 5\omega_c t - \dots \right) \quad (4.5)$$

The signal  $m(t)w(t)$  is given by

$$m(t)w(t) = \frac{1}{2}m(t) + \frac{2}{\pi} \left[ m(t) \cos \omega_c t - \frac{1}{3}m(t) \cos 3\omega_c t + \frac{1}{5}m(t) \cos 5\omega_c t - \dots \right] \quad (4.6)$$

\* The phase  $\theta_1$  is not important.

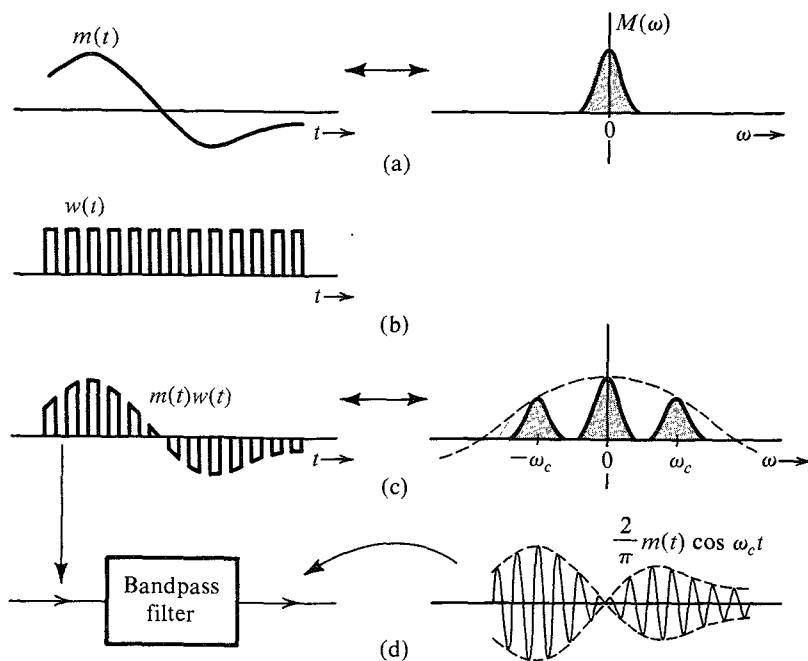


Figure 4.4 Switching modulator for DSB-SC.

The signal  $m(t)w(t)$  consists not only of the component  $m(t)$  but also of an infinite number of modulated signals with carrier frequencies  $\omega_c, 3\omega_c, 5\omega_c, \dots$ . Therefore, the spectrum of  $m(t)w(t)$  consists of the spectrum  $M(\omega)$  and  $M(\omega)$  shifted to  $\pm\omega_c, \pm3\omega_c, \pm5\omega_c, \dots$  (with decreasing relative weights), as shown in Fig. 4.4c. We are interested in the modulated component  $m(t) \cos \omega_c t$  only. To separate this component from the rest of the crowd, we pass the signal  $m(t)w(t)$  through a bandpass filter of bandwidth  $2B$  Hz, centered at the frequency  $\pm\omega_c$ . This will suppress all the spectral components not centered at  $\pm\omega_c$  to yield the desired modulated signal  $(2/\pi)m(t) \cos \omega_c t$  (Fig. 4.4d).

We now see the real payoff of this method. Multiplication of a signal by a square pulse train is in reality a switching operation. It involves switching the signal  $m(t)$  on and off periodically and can be accomplished by simple switching elements controlled by  $w(t)$ . Figure 4.5a shows one such electronic switch, the **diode-bridge modulator**, driven by a sinusoid  $A \cos \omega_c t$  to produce the switching action. Diodes  $D_1, D_2$  and  $D_3, D_4$  are matched pairs. When the signal  $\cos \omega_c t$  is of a polarity that will make terminal  $c$  positive with respect to  $d$ , all the diodes conduct. Because diodes  $D_1$  and  $D_2$  are matched, terminals  $a$  and  $b$  have the same potential and are effectively shorted. During the next half-cycle, terminal  $d$  is positive with respect to  $c$ , and all four diodes open, thus, opening the terminals  $a$  and  $b$ . The diode bridge in Fig. 4.5a, therefore, serves as a desired electronic switch, where the terminals  $a$  and  $b$  open and close periodically with the carrier frequency  $f_c$  when a sinusoid  $A \cos \omega_c t$  is applied across the terminals  $cd$ . To obtain the signal  $m(t)w(t)$ , we may place this electronic switch (terminals  $ab$ ) in series (Fig. 4.5b) or across (in parallel)  $m(t)$ , as shown in Fig. 4.5c. These modulators

are known as the **series-bridge diode modulator** and the **shunt-bridge diode modulator**, respectively. This switching on and off of  $m(t)$  repeats for each cycle of the carrier, resulting in the switched signal  $m(t)w(t)$ , which when bandpass filtered, yields the desired modulated signal  $(2/\pi)m(t) \cos \omega_c t$ .

Another switching modulator, known as the **ring modulator**, is shown in Fig. 4.6a. During the positive half-cycles of the carrier, diodes  $D_1$  and  $D_3$  conduct, and  $D_2$  and  $D_4$  are open. Hence, terminal  $a$  is connected to  $c$ , and terminal  $b$  is connected to  $d$ . During the negative half-cycles of the carrier, diodes  $D_1$  and  $D_3$  are open, and  $D_2$  and  $D_4$  are conducting, thus connecting terminal  $a$  to  $d$  and terminal  $b$  to  $c$ . Hence, the output is proportional to  $m(t)$  during the positive half-cycle and to  $-m(t)$  during the negative half-cycle. In effect,  $m(t)$  is multiplied by a square pulse train  $w_0(t)$ , shown in Fig. 4.6b. The Fourier series for  $w_0(t)$  as found in Eq. (2.76) is

$$w_0(t) = \frac{4}{\pi} \left( \cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \frac{1}{5} \cos 5\omega_c t - \dots \right) \quad (4.7a)$$

and

$$v_i(t) = m(t)w_0(t) = \frac{4}{\pi} \left[ m(t) \cos \omega_c t - \frac{1}{3} m(t) \cos 3\omega_c t + \frac{1}{5} m(t) \cos 5\omega_c t - \dots \right] \quad (4.7b)$$

The signal  $m(t)w_0(t)$  is shown in Fig. 4.6d. When this waveform is passed through a bandpass filter tuned to  $\omega_c$  (Fig. 4.6a), the filter output will be the desired signal  $(4/\pi)m(t) \cos \omega_c t$ .

In this circuit there are two inputs:  $m(t)$  and  $\cos \omega_c t$ . The input to the final bandpass filter does not contain either of these inputs. Consequently, this circuit is an example of a **double balanced modulator**.

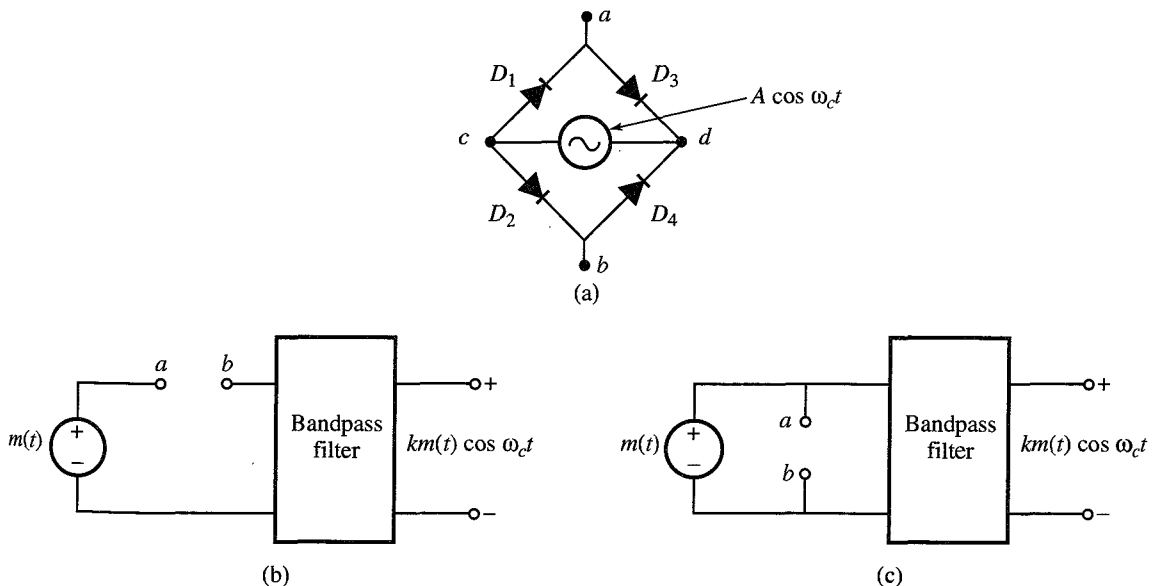


Figure 4.5 (a) Diode-bridge electronic switch. (b) Series-bridge diode modulator. (c) Shunt-bridge diode modulator.

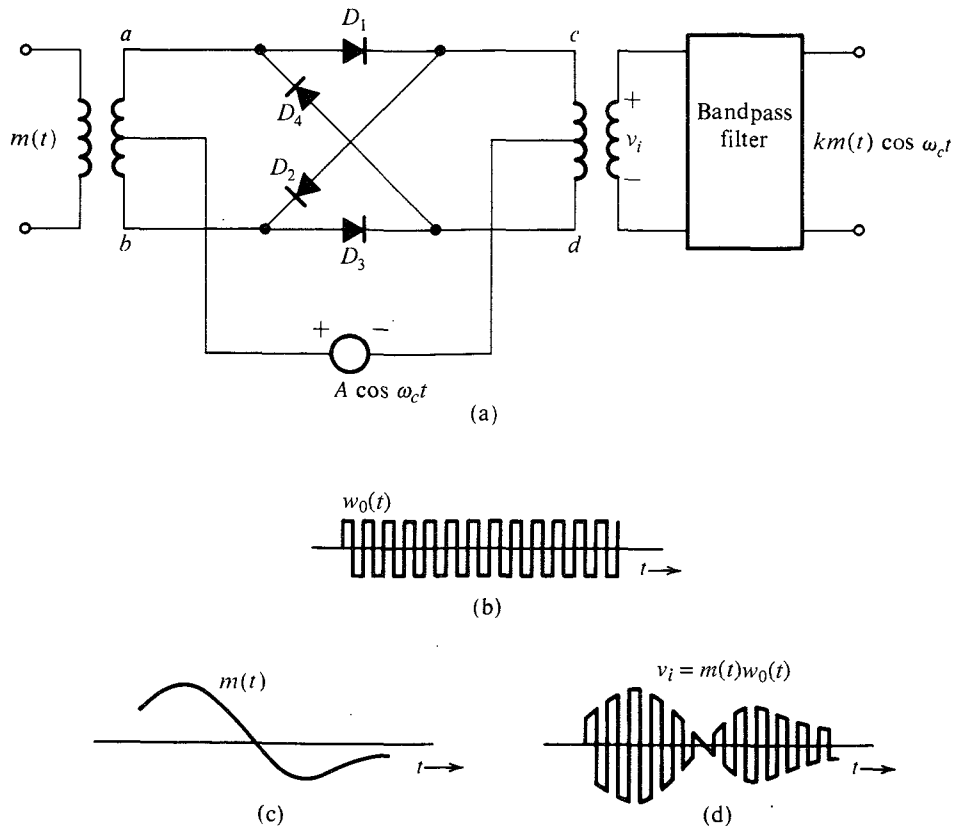


Figure 4.6 Ring modulator.

### EXAMPLE 4.2 Frequency Mixer or Converter

We shall analyze a frequency mixer, or frequency converter, used to change the carrier frequency of a modulated signal  $m(t) \cos \omega_c t$  from  $\omega_c$  to some other frequency  $\omega_I$ .

This can be done by multiplying  $m(t) \cos \omega_c t$  by  $2 \cos \omega_{\text{mix}} t$ , where  $\omega_{\text{mix}} = \omega_c + \omega_I$  or  $\omega_c - \omega_I$ , and then bandpass-filtering the product, as shown in Fig. 4.7a.

The product  $x(t)$  is

$$\begin{aligned} x(t) &= 2m(t) \cos \omega_c t \cos \omega_{\text{mix}} t \\ &= m(t) [\cos (\omega_c - \omega_{\text{mix}})t + \cos (\omega_c + \omega_{\text{mix}})t] \end{aligned}$$

If we select  $\omega_{\text{mix}} = \omega_c - \omega_I$ ,

$$x(t) = m(t) [\cos \omega_I t + \cos (2\omega_c - \omega_I)t]$$

If we select  $\omega_{\text{mix}} = \omega_c + \omega_I$ ,

$$x(t) = m(t) [\cos \omega_I t + \cos (2\omega_c + \omega_I)t]$$



In either case, a bandpass filter at the output, tuned to  $\omega_I$ , will pass the term  $m(t) \cos \omega_I t$  and suppress the other term, yielding the output  $m(t) \cos \omega_I t$ .<sup>\*</sup> Thus, the carrier frequency has been translated to  $\omega_I$  from  $\omega_c$ .

The operation of frequency mixing, or frequency conversion (also known as heterodyning), is identical to the operation of modulation with a modulating carrier frequency (the mixer oscillator frequency  $\omega_{\text{mix}}$ ) that differs from the incoming carrier frequency by  $\omega_I$ . Any one of the modulators discussed earlier can be used for frequency mixing. When we select the local carrier frequency  $\omega_{\text{mix}} = \omega_c + \omega_I$ , the operation is called **up-conversion**, and when we select  $\omega_{\text{mix}} = \omega_c - \omega_I$ , the operation is **down-conversion**.

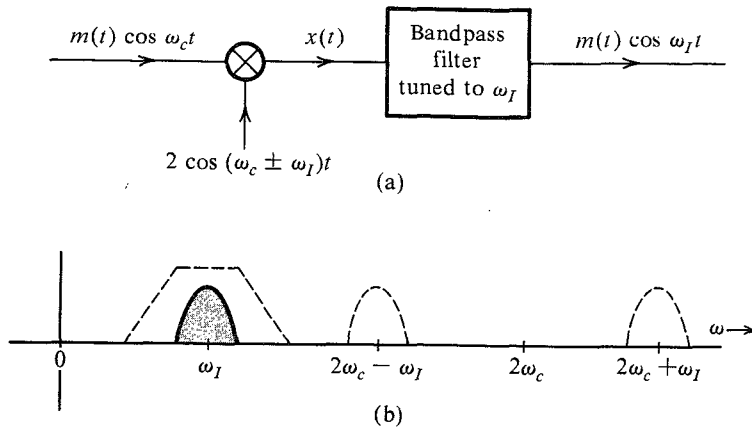


Figure 4.7 Frequency mixer or converter.

### Demodulation of DSB-SC Signals

As discussed earlier, demodulation of a DSB-SC signal is identical to modulation (see Fig. 4.1). At the receiver, we multiply the incoming signal by a local carrier of frequency and phase in synchronism with the carrier used at the modulator. The product is then passed through a low-pass filter. The only difference between the modulator and the demodulator is the output filter. In the modulator, the multiplier output is passed through a bandpass filter tuned to  $\omega_c$ , whereas in the demodulator, the multiplier output is passed through a low-pass filter. Therefore, all the modulators discussed earlier can also be used as demodulators, provided the bandpass filters at the output are replaced by low-pass filters of bandwidth  $B$ .

For demodulation, the receiver must generate a carrier in phase and frequency synchronism with the incoming carrier. These demodulators are called **synchronous** or **coherent** (also **homodyne**) demodulators.<sup>†</sup>

**EXAMPLE 4.3** Analyze the switching demodulator that uses the electronic switch (diode bridge) in Fig. 4.5 as a switch (either in series or in parallel).

<sup>\*</sup> Assuming that  $\omega_c - \omega_I \geq 2\pi B$  and  $\omega_I \geq 2\pi B$  so that various spectra in Fig. 4.7b do not overlap.

<sup>†</sup> The terms synchronous, coherent, and homodyne mean the same thing. The term homodyne is used in contrast to heterodyne where a different carrier frequency is used for the purpose of translating the spectrum (see Example 4.2).

The input signal is  $m(t) \cos \omega_c t$ . The carrier causes the periodic switching on and off of the input signal. Therefore, the output is  $m(t) \cos \omega_c t \times w(t)$ . Using the identity  $\cos x \cos y = 0.5[\cos(x + y) + \cos(x - y)]$ , we obtain

$$\begin{aligned} m(t) \cos \omega_c t \times w(t) &= m(t) \cos \omega_c t \left[ \frac{1}{2} + \frac{2}{\pi} \left( \cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \dots \right) \right] \\ &= \frac{2}{\pi} m(t) \cos^2 \omega_c t + \text{terms of the form } m(t) \cos n\omega_c t \\ &= \frac{1}{\pi} m(t) + \frac{1}{\pi} m(t) \cos 2\omega_c t + \text{terms of the form } m(t) \cos n\omega_c t \end{aligned}$$

Spectra of the terms of the form  $m(t) \cos n\omega_c t$  are centered at  $\pm n\omega_c$  and are filtered out by the low-pass filter yielding the output  $(1/\pi)m(t)$ . It is left as an exercise for the reader to show that the output of the ring demodulator in Fig. 4.6a (with the low-pass filter at the output) is  $(2/\pi)m(t)$  (twice that of the switching demodulator in this example).

### 4.3 AMPLITUDE MODULATION (AM)

For the suppressed carrier scheme discussed in the last section, a receiver must generate a carrier in frequency and phase synchronism with the carrier at the transmitter that may be located hundreds or thousands of miles away. This calls for a sophisticated receiver and could be quite costly. The other alternative is for the transmitter to transmit a carrier  $A \cos \omega_c t$  [along with the modulated signal  $m(t) \cos \omega_c t$ ] so that there is no need to generate a carrier at the receiver. In this case the transmitter needs to transmit much larger power, which makes it rather expensive. In point-to-point communications, where there is one transmitter for each receiver, substantial complexity in the receiver system can be justified, provided it results in a large enough saving in expensive high-power transmitting equipment. On the other hand, for a broadcast system with a multitude of receivers for each transmitter, it is more economical to have one expensive high-power transmitter and simpler, less expensive receivers. The second option (transmitting a carrier along with the modulated signal) is the obvious choice for this case. This is the so-called AM (amplitude modulation), in which the transmitted signal  $\varphi_{AM}(t)$  is given by

$$\varphi_{AM}(t) = A \cos \omega_c t + m(t) \cos \omega_c t \quad (4.8a)$$

$$= [A + m(t)] \cos \omega_c t \quad (4.8b)$$

The spectrum of  $\varphi_{AM}(t)$  is the same as that of  $m(t) \cos \omega_c t$  plus two additional impulses at  $\pm\omega_c$ ,

$$\varphi_{AM}(t) \iff \frac{1}{2}[M(\omega + \omega_c) + M(\omega - \omega_c)] + \pi A[\delta(\omega + \omega_c) + \delta(\omega - \omega_c)] \quad (4.8c)$$

Recall that the DSB-SC signal is  $m(t) \cos \omega_c t$ . From Eq. (4.8b) it follows that the AM signal is identical to the DSB-SC signal with  $A + m(t)$  as the modulating signal [instead of  $m(t)$ ]. Therefore, to sketch  $\varphi_{AM}(t)$ , we sketch  $A + m(t)$  and  $-[A + m(t)]$  and fill in between with the sinusoid of the carrier frequency. Two cases are considered in Fig. 4.8. In the first case,  $A$  is

large enough so that  $A + m(t) \geq 0$  (is nonnegative) for all values of  $t$ . In the second case,  $A$  is not large enough to satisfy this condition. In the first case, the envelope has the same shape as  $m(t)$  (although riding on a dc of magnitude  $A$ ). In the second case, the envelope shape is not  $m(t)$  because some parts get rectified. This means we can detect the desired signal  $m(t)$  by detecting the envelope in the first case. Such a detection is not possible in the second case. We shall see that the envelope detection is an extremely simple and inexpensive operation, which does not require generation of a local carrier for the demodulation. But as seen above the envelope of AM has the information about  $m(t)$  only if the AM signal  $[A + m(t)] \cos \omega_c t$  satisfies the condition  $A + m(t) > 0$  for all  $t$ .

Recall also that the envelope of a signal  $E(t) \cos \omega_c t$  is  $E(t)$  provided  $E(t) \geq 0$  for all  $t$ .<sup>\*</sup> This means [see Eq. (4.8b)] that  $A + m(t)$  is the envelope of  $\varphi_{AM}(t)$  only if  $A + m(t) \geq 0$  for all  $t$ . This conclusion is readily verified from Fig. 4.8d and e. In Fig. 4.8d, where  $A + m(t) \geq 0$ ,  $A + m(t)$  is indeed the envelope, and  $m(t)$  can be recovered from this envelope. In Fig. 4.8e, where  $A + m(t)$  is not always positive, the envelope is not  $A + m(t)$ , but rectified  $A + m(t)$ , and  $m(t)$  cannot be recovered from the envelope. Consequently, demodulation of  $\varphi_{AM}(t)$  in

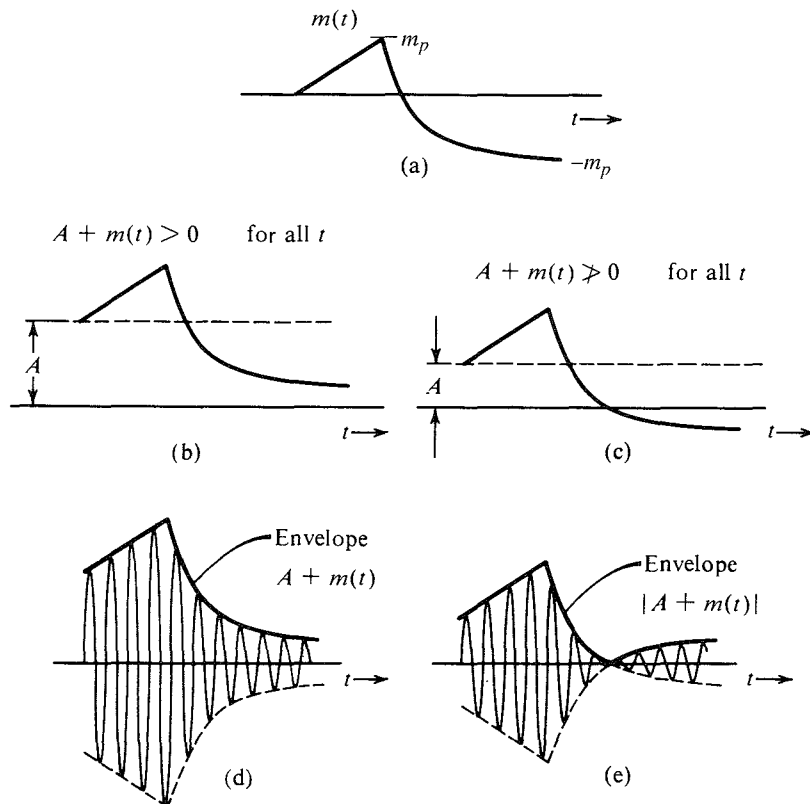


Figure 4.8 AM signal and its envelope.

\*  $E(t)$  must also be a slowly varying signal as compared to  $\cos \omega_c t$ .

Fig. 4.8d amounts to simple envelope detection. Thus, the condition for envelope detection of an AM signal is

$$A + m(t) \geq 0 \quad \text{for all } t \quad (4.9a)$$

If  $m(t) \geq 0$  for all  $t$ , then  $A = 0$  also satisfies the condition (4.9a). In this case there is no need to add any carrier because the envelope of the DSB-SC signal  $m(t) \cos \omega_c t$  is  $m(t)$  and such a DSB-SC signal can be detected by envelope detection. In the following discussion we assume that  $m(t) \not\geq 0$  for all  $t$ , that is,  $m(t)$  does take on negative values over some range of  $t$ .

Let  $m_p$  be the peak amplitude (positive or negative) of  $m(t)$  (see Fig. 4.8). This means that  $m(t) \geq -m_p$ . Hence, the condition (4.9a) is equivalent to\*

$$A \geq m_p \quad (4.9b)$$

Thus, the minimum carrier amplitude required for the viability of envelope detection is  $m_p$ . This is quite clear from Fig. 4.8.

We define the modulation index  $\mu$  as

$$\mu = \frac{m_p}{A} \quad (4.10a)$$

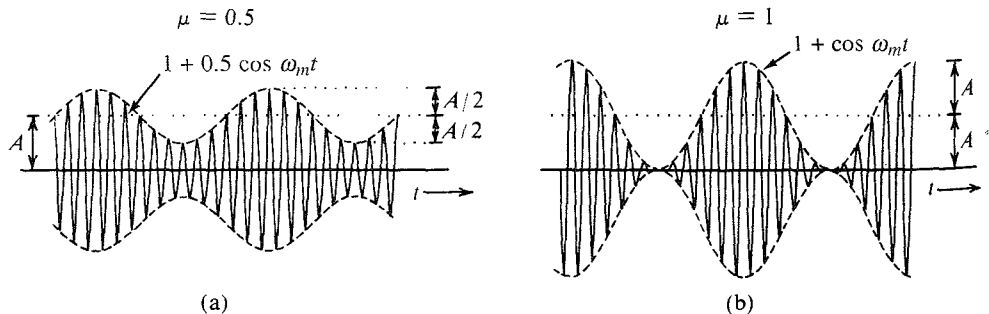
where  $A$  is the carrier amplitude. Note that  $m_p$  is a constant of the signal  $m(t)$ . Because  $A \geq m_p$  and because there is no upper bound on  $A$ , it follows that

$$0 \leq \mu \leq 1 \quad (4.10b)$$

as the required condition for the viability of demodulation of AM by an envelope detector.

When  $A < m_p$ , Eq. (4.10a) shows that  $\mu > 1$  (overmodulation). In this case, the option of envelope detection is no longer viable. We then need to use synchronous demodulation. Note that synchronous demodulation can be used for any value of  $\mu$  (see Prob. 4.3-1). The envelope detector, which is considerably simpler and less expensive than the synchronous detector, can be used only for  $\mu \leq 1$ .

**EXAMPLE 4.4** Sketch  $\varphi_{AM}(t)$  for modulation indices of  $\mu = 0.5$  and  $\mu = 1$ , when  $m(t) = B \cos \omega_m t$ . This case is referred to as **tone modulation** because the modulating signal is a pure sinusoid (or tone).



**Figure 4.9** Tone-modulated AM. (a)  $\mu = 0.5$ . (b)  $\mu = 1$ .

\* In case the negative and the positive peak amplitudes are not identical,  $m_p$  in condition (4.9b) is the absolute negative peak amplitude.

In this case,  $m_p = B$  and the modulation index according to Eq. (4.10a) is

$$\mu = \frac{B}{A}$$

Hence,  $B = \mu A$  and

$$m(t) = B \cos \omega_m t = \mu A \cos \omega_m t$$

Therefore,

$$\varphi_{AM}(t) = [A + m(t)] \cos \omega_c t = A[1 + \mu \cos \omega_m t] \cos \omega_c t \quad (4.11)$$

Figure 4.9 shows the modulated signals corresponding to  $\mu = 0.5$  and  $\mu = 1$ , respectively.

### Sideband and Carrier Power

The advantage of envelope detection in AM has its price. In AM, the carrier term does not carry any information, and hence, the carrier power is wasted,

$$\varphi_{AM}(t) = \underbrace{A \cos \omega_c t}_{\text{carrier}} + \underbrace{m(t) \cos \omega_c t}_{\text{sidebands}}$$

The carrier power  $P_c$  is the mean square value of  $A \cos \omega_c t$ , which is  $A^2/2$ . The sideband power  $P_s$  is the power of  $m(t) \cos \omega_c t$ , which is  $0.5 \overline{m^2(t)}$  [see Eq. (3.70)]. Hence,

$$P_c = \frac{A^2}{2} \quad \text{and} \quad P_s = \frac{1}{2} \overline{m^2(t)}$$

The sideband power is the useful power and the carrier power is the power wasted for convenience. The total power is the sum of the carrier (wasted) power and the sideband (useful) power. Hence,  $\eta$ , the power efficiency, is

$$\eta = \frac{\text{useful power}}{\text{total power}} = \frac{P_s}{P_c + P_s} = \frac{\overline{m^2(t)}}{A^2 + \overline{m^2(t)}} 100\%$$

For the special case of tone modulation,

$$m(t) = \mu A \cos \omega_m t \quad \text{and} \quad \overline{m^2(t)} = \frac{(\mu A)^2}{2}$$

Hence

$$\eta = \frac{\mu^2}{2 + \mu^2} 100\%$$

with the condition that  $0 \leq \mu \leq 1$ . It can be seen that  $\eta$  increases monotonically with  $\mu$ , and  $\eta_{\max}$  occurs at  $\mu = 1$ , for which

$$\eta_{\max} = 33\%$$

Thus, for tone modulation, under best conditions ( $\mu = 1$ ), only one-third of the transmitted power is used for carrying message. For practical signals, the efficiency is even worse—the order of 25% or lower—compared to that of the DSB-SC case. The best condition implies

$\mu = 1$ . Smaller values of  $\mu$  degrade efficiency further. For this reason volume compression and peak limiting are commonly used in AM to ensure that full modulation ( $\mu = 1$ ) is maintained most of the time.

**EXAMPLE 4.5** Determine  $\eta$  and the percentage of the total power carried by the sidebands of the AM wave for tone modulation when (a)  $\mu = 0.5$  and (b)  $\mu = 0.3$ .

For  $\mu = 0.5$ ,

$$\eta = \frac{\mu^2}{2 + \mu^2} 100\% = \frac{(0.5)^2}{2 + (0.5)^2} 100\% = 11.11\%$$

Hence, only about 11% of the total power is in the sidebands. For  $\mu = 0.3$ ,

$$\eta = \frac{(0.3)^2}{2 + (0.3)^2} 100\% = 4.3\%$$

Hence, only 4.3% of the total power is the useful power (power in sidebands).

### Generation of AM Signals

AM signals can be generated by any DSB-SC modulators discussed in Sec. 4.2 if the modulating signal is  $A + m(t)$  instead of just  $m(t)$ . But because there is no need to suppress the carrier in the output, the modulating circuits do not have to be balanced. This results in considerably simpler modulators for AM. Figure 4.10 shows a switching modulator, where the switching action is provided by a single diode (instead of a diode bridge as in Fig. 4.5). The input is  $c \cos \omega_c t + m(t)$  with  $c \gg m(t)$ , so that the switching action of the diode is controlled by  $c \cos \omega_c t$ . The diode opens and shorts periodically with  $\cos \omega_c t$ , in effect multiplying the input signal  $[c \cos \omega_c t + m(t)]$  by  $w(t)$ . The voltage across terminals  $bb'$  is

$$\begin{aligned} v_{bb'}(t) &= [c \cos \omega_c t + m(t)]w(t) \\ &= [c \cos \omega_c t + m(t)] \left[ \frac{1}{2} + \frac{2}{\pi} \left( \cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \frac{1}{5} \cos 5\omega_c t - \dots \right) \right] \\ &= \underbrace{\frac{c}{2} \cos \omega_c t + \frac{2}{\pi} m(t) \cos \omega_c t}_{\text{AM}} + \underbrace{\text{other terms}}_{\text{suppressed by bandpass filter}} \end{aligned}$$

The bandpass filter tuned to  $\omega_c$  suppresses all the other terms, yielding the desired AM signal at the output.

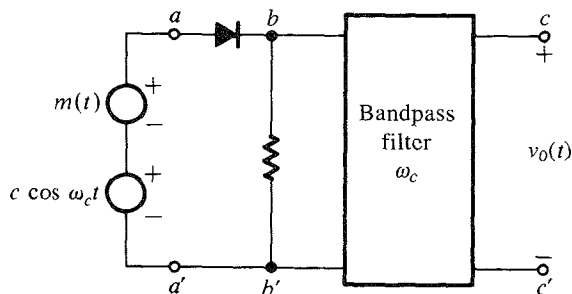


Figure 4.10 AM generator.

## Demodulation of AM Signals

The AM signal can be demodulated coherently by a locally generated carrier (see Prob. 4.3-1). However, coherent, or synchronous, demodulation of AM\* will defeat the very purpose of AM and, hence, is rarely used in practice. We shall consider here two noncoherent methods of AM demodulation: (1) rectifier detection, and (2) envelope detection.

**Rectifier Detector:** If an AM signal is applied to a diode and a resistor circuit (Fig. 4.11), the negative part of the AM wave will be suppressed. The output across the resistor is a half-wave rectified version of the AM signal. In essence, the AM signal is multiplied by  $w(t)$ . Hence, the rectified output  $v_R$  is

$$\begin{aligned} v_R &= \{[A + m(t)] \cos \omega_c t\} w(t) \\ &= [A + m(t)] \cos \omega_c t \left[ \frac{1}{2} + \frac{2}{\pi} \left( \cos \omega_c t - \frac{1}{3} \cos 3\omega_c t + \frac{1}{5} \cos 5\omega_c t - \dots \right) \right] \\ &= \frac{1}{\pi} [A + m(t)] + \text{other terms of higher frequencies} \end{aligned}$$

When  $v_R$  is applied to a low-pass filter of cutoff  $B$  Hz, the output is  $[A + m(t)]/\pi$ , and all the other terms in  $v_R$  of frequencies higher than  $B$  Hz are suppressed. The dc term  $A/\pi$  may be blocked by a capacitor (Fig. 4.11) to give the desired output  $m(t)/\pi$ . The output can be doubled by using a full-wave rectifier.

It is interesting to note that rectifier detection is in effect synchronous detection performed without using a local carrier. The high carrier content in AM ensures that its zero crossings are periodic and the information about frequency and phase of the carrier at the transmitter is built in to the AM signal itself.

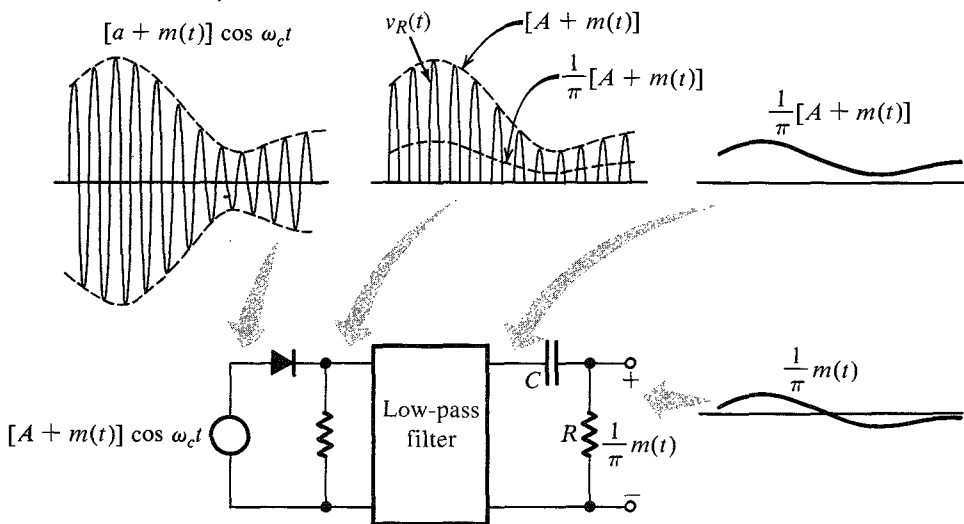


Figure 4.11 Rectifier detector for AM.

\* By AM, we mean the case  $\mu \leq 1$ .

**Envelope Detector:** In an envelope detector, the output of the detector follows the envelope of the modulated signal. The circuit shown in Fig. 4.12a functions as an envelope detector. On the positive cycle of the input signal, the diode conducts and the capacitor  $C$  charges up to the peak voltage of the input signal. As the input signal falls below this peak value, the diode is cut off, because the capacitor voltage (which is very nearly the peak voltage) is greater than the input signal voltage, thus causing the diode to open. The capacitor now discharges through the resistor  $R$  at a slow rate (with a time constant  $RC$ ). During the next positive cycle, the same drama repeats. When the input signal becomes greater than the capacitor voltage, the diode conducts again. The capacitor again charges to the peak value of this (new) cycle. The capacitor discharges slowly during the cutoff period, thus changing the capacitor voltage very slightly.

During each positive cycle, the capacitor charges up to the peak voltage of the input signal and then decays slowly until the next positive cycle as shown in Fig. 4.12b. The output voltage  $v_C(t)$ , thus, closely follows the envelope of the input. Capacitor discharge between positive peaks causes a ripple signal of frequency  $\omega_c$  in the output. This ripple can be reduced by increasing the time constant  $RC$  so that the capacitor discharges very little between the positive peaks ( $RC \gg 1/\omega_c$ ). Making  $RC$  too large, however, would make it impossible for the capacitor voltage to follow the envelope (see Fig. 4.12b). Thus,  $RC$  should be large

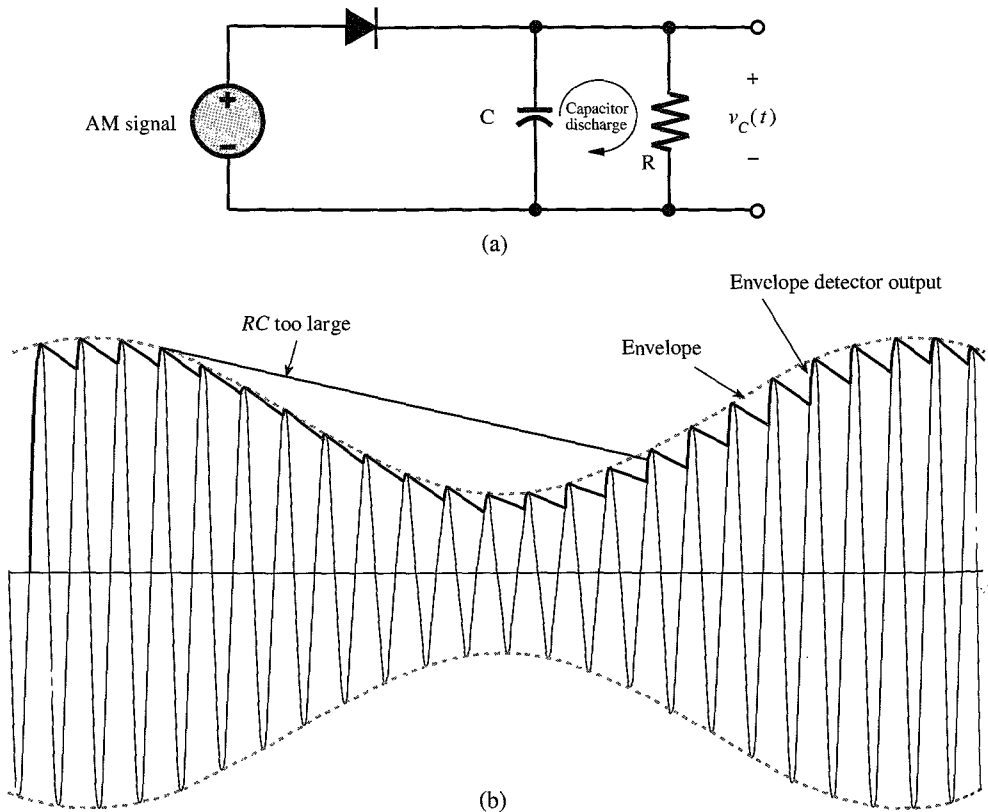


Figure 4.12 Envelope detector for AM.



compared to  $1/\omega_c$  but should be small compared to  $1/2\pi B$ , where  $B$  is the highest frequency in  $m(t)$  (see Example 4.6). This, incidentally, also requires that  $\omega_c \gg 2\pi B$ , a condition that is necessary for a well-defined envelope.

The envelope-detector output is  $v_C(t) = A + m(t)$  with a ripple of frequency  $\omega_c$ . The dc term  $A$  can be blocked out by a capacitor or a simple  $RC$  high-pass filter. The ripple may be reduced further by another (low-pass)  $RC$  filter.

Both the rectifier detector and the envelope detector consist of a half-wave rectifier followed by a low-pass filter. Superficially, these detectors may appear equivalent, but they are distinct and operate on very different principles. The rectifier detector is basically a synchronous demodulator. The envelope detection, on the other hand, is a nonlinear operation. Observe that the low-pass filter in the rectifier detector is designed to separate  $m(t)$  from terms such as  $m(t) \cos n\omega_c t$ ; it does not depend on the value of  $\mu$ . On the other hand, we show in Example 4.6 that the time constant  $RC$  of the low-pass filter for the envelope detector does depend on the value of  $\mu$ .

**EXAMPLE 4.6** For tone modulation (Example 4.4), determine the upper limit of  $RC$  to ensure that the capacitor voltage follows the envelope.

Figure 4.13 shows the envelope and the voltage across the capacitor. The capacitor discharges from the peak value  $E$  starting at some arbitrary instant  $t = 0$ . The voltage  $v_C$  across the capacitor is given by

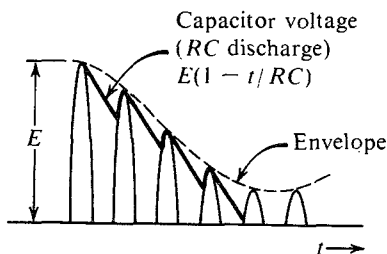
$$v_C = Ee^{-t/RC}$$

Because the time constant is much larger than the interval between the two successive cycles of the carrier ( $RC \gg 1/\omega_c$ ), the capacitor voltage  $v_C$  discharges exponentially for a short time compared to its time constant. Hence, the exponential can be approximated by a straight line obtained from the first two terms in Taylor's series for  $Ee^{-t/RC}$ ,

$$v_C \simeq E \left( 1 - \frac{t}{RC} \right)$$

The slope of the discharge is  $-E/RC$ . In order for the capacitor to follow the envelope  $E(t)$ , the magnitude of the slope of the  $RC$  discharge must be greater than the magnitude of the slope of the envelope  $E(t)$ . Hence,

$$\left| \frac{dv_C}{dt} \right| = \frac{E}{RC} \geq \left| \frac{dE}{dt} \right| \quad (4.12)$$



**Figure 4.13** Capacitor discharge in an envelope detector.

But the envelope  $E(t)$  of a tone-modulated carrier is [Eq. (4.11)]

$$E(t) = A[1 + \mu \cos \omega_m t]$$

$$\frac{dE}{dt} = -\mu A \omega_m \sin \omega_m t$$

Hence, Eq. (4.12) becomes

$$\frac{A(1 + \mu \cos \omega_m t)}{RC} \geq \mu A \omega_m \sin \omega_m t \quad \text{for all } t$$

or

$$RC \leq \frac{1 + \mu \cos \omega_m t}{\mu \omega_m \sin \omega_m t} \quad \text{for all } t$$

The worst possible case occurs when the right-hand side is the minimum. This is found (as usual, by taking the derivative and setting it to zero) to be when  $\cos \omega_m t = -\mu$ . For this case, the right-hand side is  $\sqrt{(1 - \mu^2)}/\mu \omega_m$ . Hence,

$$RC \leq \frac{1}{\omega_m} \left( \frac{\sqrt{1 - \mu^2}}{\mu} \right)$$

## 4.4 QUADRATURE AMPLITUDE MODULATION (QAM)

The DSB signals occupy twice the bandwidth required for the baseband. This disadvantage can be overcome by transmitting two DSB signals using carriers of the same frequency but in phase quadrature, as shown in Fig. 4.14. In this figure, the boxes labeled  $-\pi/2$  are phase shifters, which delay the phase of an input sinusoid by  $-\pi/2$  rad. If the two baseband signals to be transmitted are  $m_1(t)$  and  $m_2(t)$ , the corresponding QAM signal  $\varphi_{\text{QAM}}(t)$ , the sum of the two DSB-modulated signals, is

$$\varphi_{\text{QAM}}(t) = m_1(t) \cos \omega_c t + m_2(t) \sin \omega_c t$$

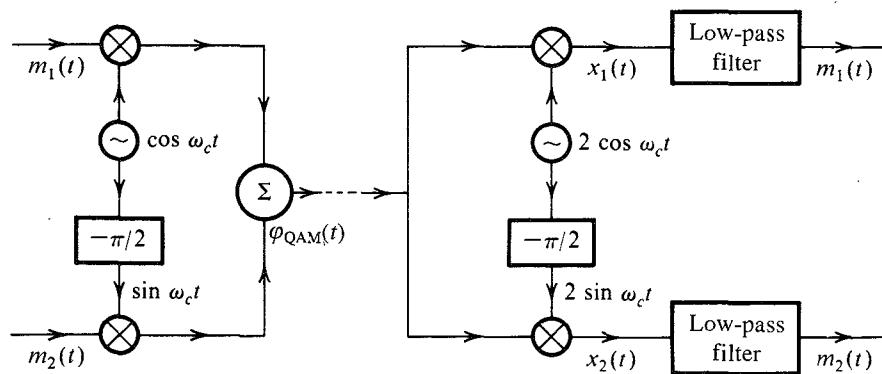


Figure 4.14 Quadrature amplitude multiplexing.

Both modulated signals occupy the same band. Yet two baseband signals can be separated at the receiver by synchronous detection using two local carriers in phase quadrature, as shown in Fig. 4.14. This can be shown by considering the multiplier output  $x_1(t)$  of the upper arm of the receiver (Fig. 4.14):

$$\begin{aligned} x_1(t) &= 2\varphi_{\text{QAM}}(t) \cos \omega_c t = 2[m_1(t) \cos \omega_c t + m_2(t) \sin \omega_c t] \cos \omega_c t \\ &= m_1(t) + m_1(t) \cos 2\omega_c t + m_2(t) \sin 2\omega_c t \end{aligned}$$

The last two terms are suppressed by the low-pass filter, yielding the desired output  $m_1(t)$ . Similarly, the output of the lower receiver branch can be shown to be  $m_2(t)$ . This scheme is known as **quadrature amplitude modulation (QAM)** or **quadrature multiplexing**. Thus, two baseband signals, each of bandwidth  $B$  Hz, can be transmitted simultaneously over a bandwidth  $2B$  by using DSB transmission and quadrature multiplexing. The upper channel is also known as the **in-phase (I)** channel and the lower channel is the **quadrature (Q)** channel.

QAM is somewhat of an exacting scheme. A slight error in the phase or the frequency of the carrier at the demodulator in QAM will not only result in loss and distortion of signals, but will also lead to interference between the two channels. To show this let the carrier at the demodulator be  $2 \cos(\omega_c t + \theta)$ . In this case,

$$\begin{aligned} x_1(t) &= 2[m_1(t) \cos \omega_c t + m_2(t) \sin \omega_c t] \cos(\omega_c t + \theta) \\ &= m_1(t) \cos \theta + m_1(t) \cos(2\omega_c t + \theta) - m_2(t) \sin \theta + m_2(t) \sin(2\omega_c t + \theta) \end{aligned}$$

The low-pass filter suppresses the two signals with frequency  $2\omega_c$ , resulting in the output  $m_1(t) \cos \theta - m_2(t) \sin \theta$ . Thus, in addition to the desired signal  $m_1(t)$ , we also receive signal  $m_2(t)$  in the upper branch. Similar argument shows that in addition to the desired signal  $m_2(t)$ , we receive signal  $m_1(t)$  in the lower branch. This cochannel\* interference is undesirable. Similar difficulties arise when the local frequency is in error (see Prob. 4.4-1). In addition, unequal attenuation of the USB and the LSB during transmission also leads to crosstalk or cochannel interference.

Quadrature multiplexing is used in color television to multiplex the so-called chrominance signals, which carry the information about colors. There the synchronization is achieved by periodic insertion of a short burst of carrier signal (called **color burst** in the transmitted signal, as explained in Sec. 4.9).

## 4.5 AMPLITUDE MODULATION: SINGLE SIDEBAND (SSB)

The DSB spectrum has two sidebands: the upper sideband (USB) and the lower sideband (LSB), both containing the complete information of the baseband signal (Fig. 4.15). A scheme in which only one sideband is transmitted is known as **single-sideband (SSB) transmission**, which requires only one-half the bandwidth of the DSB signal.

An SSB signal can be coherently (synchronously) demodulated. For example, multiplication of a USB signal (Fig. 4.15c) by  $\cos \omega_c t$  shifts its spectrum to the left and right by  $\omega_c$ , yielding the spectrum in Fig. 4.15e. Low-pass filtering of this signal yields the desired baseband signal. The case is similar with LSB signals. Hence, demodulation of SSB signals

\* Cochannel refers to channels having the same carrier frequency.

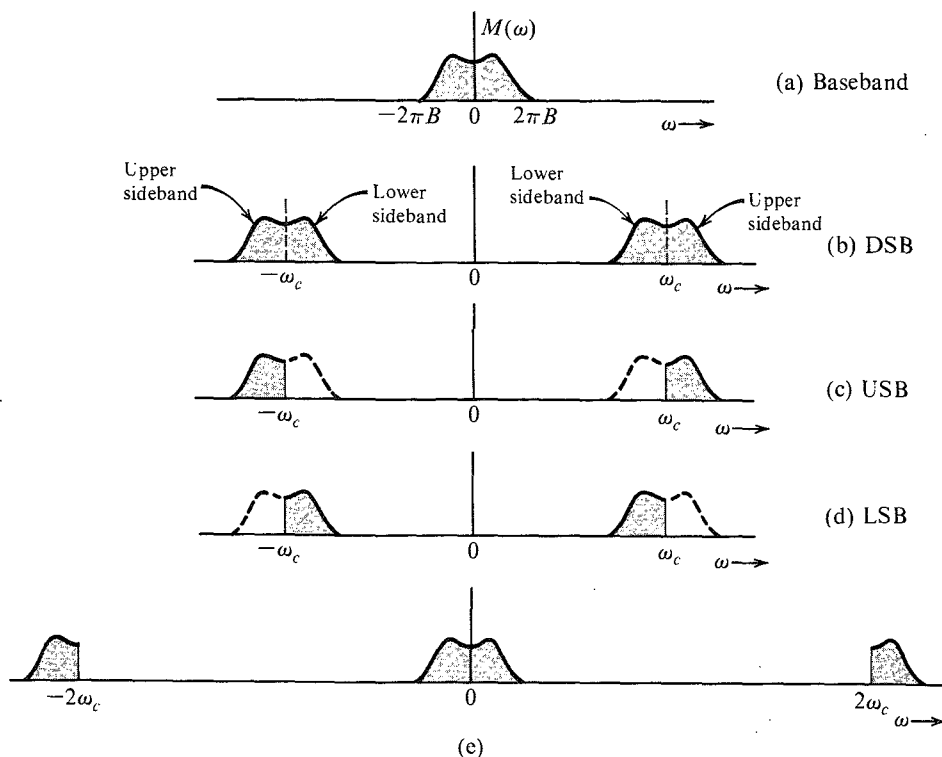


Figure 4.15 SSB spectra.

is identical to that of DSB-SC signals. Note that we are talking of SSB signals without an additional carrier, and, hence, they are suppressed carrier signals (SSB-SC).

### Time-Domain Representation of SSB Signals

Because the building blocks of an SSB signal are the sidebands, we shall first obtain a time-domain expression for each sideband. Figure 4.16a shows the spectrum  $M(\omega)$ . Figure 4.16b shows the USB  $M_+(\omega)$  and Fig. 4.16c shows the LSB  $M_-(\omega)$ . From Fig. 4.16b and c, we observe that  $M_+(\omega) = M(\omega)u(\omega)$  and  $M_-(\omega) = M(\omega)u(-\omega)$ . Let  $m_+(t)$  and  $m_-(t)$  be the inverse Fourier transforms of  $M_+(\omega)$  and  $M_-(\omega)$ , respectively.\* Because the amplitude spectra  $|M_+(\omega)|$  and  $|M_-(\omega)|$  are not even functions of  $\omega$ , the signals  $m_+(t)$  and  $m_-(t)$  cannot be real; they are complex. Moreover,  $M_+(\omega)$  and  $M_-(\omega)$  are the two halves of  $M(\omega)$ . Hence, from Eqs. (3.10), it follows that  $M_+(-\omega)$  and  $M_-(\omega)$  are conjugates. Consequently,  $m_+(t)$  and  $m_-(t)$  are conjugates (see Prob. 3.1-3). Also, because  $m_+(t) + m_-(t) = m(t)$ , we can express

$$m_+(t) = \frac{1}{2}[m(t) + jm_h(t)] \quad (4.13a)$$

and

$$m_-(t) = \frac{1}{2}[m(t) - jm_h(t)] \quad (4.13b)$$

\* In the literature,  $2m_+(t)$  is also known as the preenvelope of  $m(t)$ .

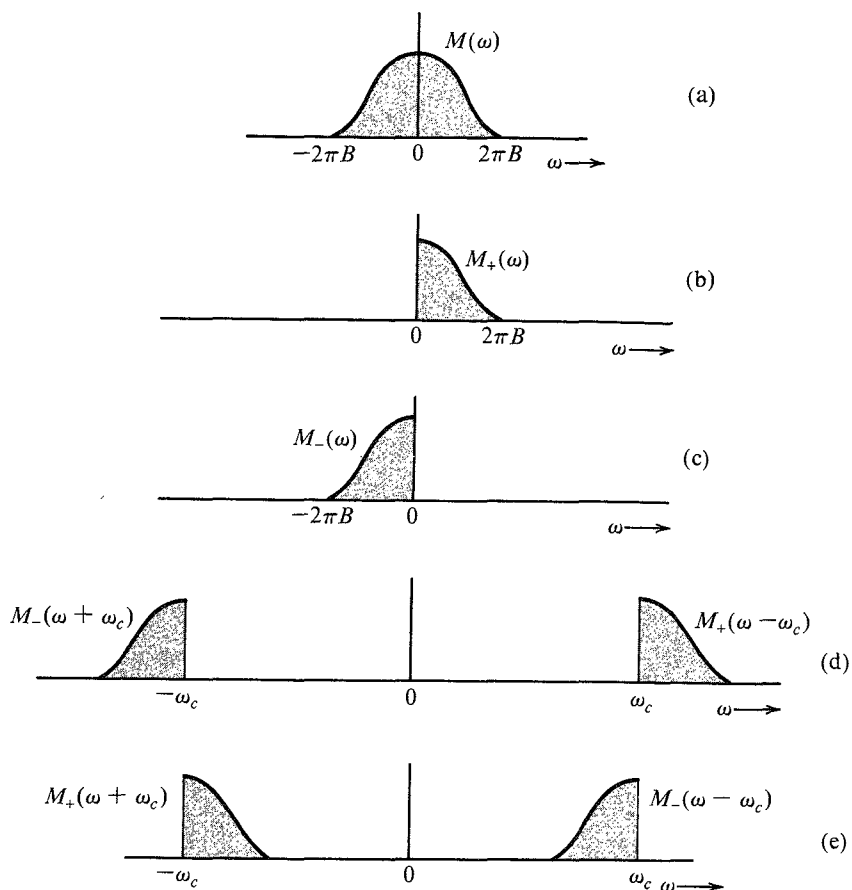


Figure 4.16 Expressing SSB spectra in terms of  $M_+(\omega)$  and  $M_-(\omega)$ .

where  $m_h(t)$  is unknown. To determine  $m_h(t)$ , we note that

$$\begin{aligned}
 M_+(\omega) &= M(\omega)u(\omega) \\
 &= \frac{1}{2}M(\omega)[1 + \operatorname{sgn}(\omega)] \\
 &= \frac{1}{2}M(\omega) + \frac{1}{2}M(\omega)\operatorname{sgn}(\omega)
 \end{aligned} \tag{4.14a}$$

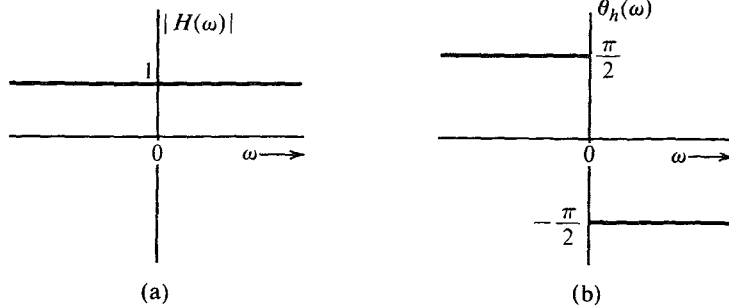
From Eqs. (4.13a) and (4.14a), it follows that  $jm_h(t) \Leftrightarrow M(\omega)\operatorname{sgn}(\omega)$ . Hence,

$$M_h(\omega) = -jM(\omega)\operatorname{sgn}(\omega) \tag{4.14b}$$

Application of the duality property to pair 12 of Table 3.1 yields  $1/\pi t \Leftrightarrow -j\operatorname{sgn}(\omega)$ . Applying this result and the time convolution property to Eq. (4.14b) yields  $m_h(t) = m(t) * 1/\pi t$ , that is,

$$m_h(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{m(\alpha)}{t - \alpha} d\alpha \tag{4.15}$$

The right-hand side of Eq. (4.15) defines the **Hilbert transform** of  $m(t)$ . Thus, the signal  $m_h(t)$



**Figure 4.17** Transfer function of an ideal  $\pi/2$  phase shifter (Hilbert transformer).

is the Hilbert transform of  $m(t)$ . From Eq. (4.14b), it follows that if  $m(t)$  is passed through a transfer function  $H(\omega) = -j \operatorname{sgn}(\omega)$ , then the output is  $m_h(t)$ , the Hilbert transform of  $m(t)$ . Because

$$H(\omega) = -j \operatorname{sgn}(\omega) \quad (4.16a)$$

$$= \begin{cases} -j = 1e^{-j\pi/2} & \omega > 0 \\ j = 1e^{j\pi/2} & \omega < 0 \end{cases} \quad (4.16b)$$

it follows that  $|H(\omega)| = 1$  and that  $\theta_h(\omega) = -\pi/2$  for  $\omega > 0$  and  $\pi/2$  for  $\omega < 0$ , as shown in Fig. 4.17. Thus, if we delay the phase of every component of  $m(t)$  by  $\pi/2$  (without changing its amplitude), the resulting signal is  $m_h(t)$ , the Hilbert transform of  $m(t)$ . Therefore, a Hilbert transformer is an ideal phase shifter that shifts the phase of every spectral component by  $-\pi/2$ . We can now express the SSB signal in terms of  $m(t)$  and  $m_h(t)$ . From Fig. 4.16d it is clear that the USB spectrum  $\Phi_{\text{USB}}(\omega)$  can be expressed as

$$\Phi_{\text{USB}}(\omega) = M_+(\omega - \omega_c) + M_-(\omega + \omega_c)$$

The inverse transform of this equation yields

$$w\varphi_{\text{USB}}(t) = m_+(t)e^{j\omega_c t} + m_-(t)e^{-j\omega_c t}$$

Substituting Eqs. (4.13) in the preceding equation yields

$$\varphi_{\text{USB}}(t) = m(t) \cos \omega_c t - m_h(t) \sin \omega_c t \quad (4.17a)$$

Using a similar argument, we can show that

$$\varphi_{\text{LSB}}(t) = m(t) \cos \omega_c t + m_h(t) \sin \omega_c t \quad (4.17b)$$

Hence, a general SSB signal  $\varphi_{\text{SSB}}(t)$  can be expressed as

$$\varphi_{\text{SSB}}(t) = m(t) \cos \omega_c t \mp m_h(t) \sin \omega_c t \quad (4.17c)$$

where the minus sign applies to USB and the plus sign applies to LSB.

#### EXAMPLE 4.7 Tone Modulation: SSB

Find  $\varphi_{\text{SSB}}(t)$  for a simple case of a tone modulation, that is, when the modulating signal is a sinusoid  $m(t) = \cos \omega_m t$ .

Recall that the Hilbert transform delays the phase of each spectral component by  $\pi/2$ . In the present case, there is only one spectral component of frequency  $\omega_m$ . Delaying the phase of  $m(t)$  by  $\pi/2$  yields

$$m_h(t) = \cos\left(\omega_m t - \frac{\pi}{2}\right) = \sin \omega_m t$$

Hence, from Eq. (4.17c),

$$\begin{aligned}\varphi_{\text{SSB}}(t) &= \cos \omega_m t \cos \omega_c t \mp \sin \omega_m t \sin \omega_c t \\ &= \cos(\omega_c \pm \omega_m)t\end{aligned}$$

Thus,

$$\varphi_{\text{USB}}(t) = \cos(\omega_c + \omega_m)t \quad \varphi_{\text{LSB}}(t) = \cos(\omega_c - \omega_m)t$$

To verify these results, consider the spectrum of  $m(t)$  (Fig. 4.18a) and its DSB-SC (Fig. 4.18b), USB (Fig. 4.18c), and LSB (Fig. 4.18d) spectra. It is evident that the spectra in Fig. 4.18c and d do indeed correspond to the  $\varphi_{\text{USB}}(t)$  and  $\varphi_{\text{LSB}}(t)$  derived earlier.

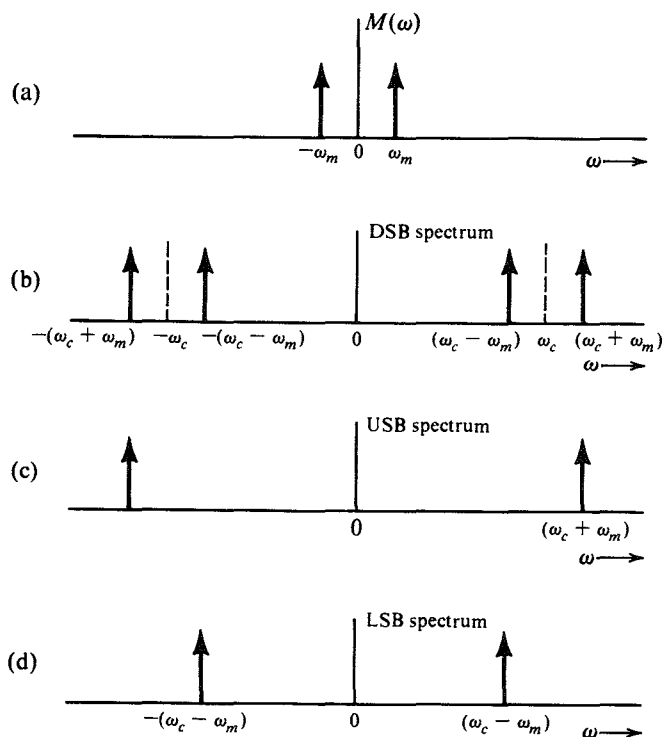


Figure 4.18 SSB spectra for tone modulation.

### Generation of SSB Signals<sup>2</sup>

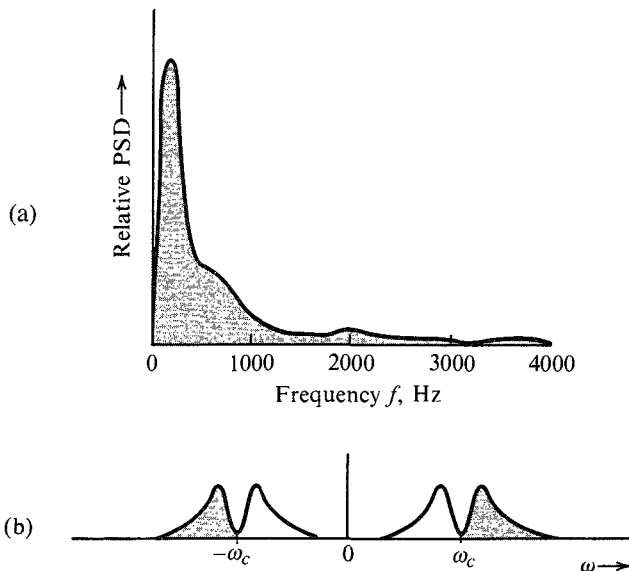
Two methods are commonly used to generate SSB signals. The first method uses sharp cutoff filters to eliminate the undesired sideband, and the second method uses phase-shifting networks

to achieve the same goal. Yet another method, known as Weaver's method,<sup>3</sup> can also be used to generate SSB signals, provided the baseband signal spectrum has little power near the origin.

**Selective-Filtering Method:** This is the most commonly used method of generating SSB signals. In this method, a DSB-SC signal is passed through a sharp cutoff filter to eliminate the undesired sideband.

To obtain the USB, the filter should pass all components above  $\omega_c$  unattenuated and completely suppress all components below  $\omega_c$ . Such an operation requires an ideal filter, which is unrealizable. It can, however, be realized closely if there is some separation between the passband and the stopband. Fortunately, the voice signal provides this condition, because its spectrum shows little power content at the origin (Fig. 4.19a). In addition, articulation tests have shown that for speech signals, frequency components below 300 Hz are not important. In other words, we may suppress all speech components below 300 Hz without affecting the intelligibility appreciably.\* Thus, filtering of the unwanted sideband becomes relatively easy for speech signals because we have a 600-Hz transition region around the cutoff frequency  $\omega_c$ . To minimize adjacent channel interference, the undesired sideband should be attenuated at least 40 dB.†

**Phase-Shift Method:** Equation (4.17) is the basis for this method. Figure 4.20 shows the implementation of Eq. (4.17). The box marked " $-\pi/2$ " is a  $\pi/2$  phase shifter, which delays the phase of every spectral component by  $\pi/2$ . Hence, it is a Hilbert transformer. Note that



**Figure 4.19** Relative power spectrum of speech signal and the corresponding USB spectrum.

\* Similarly, suppression of speech-signal components above 3500 Hz causes no appreciable change in intelligibility.

† For very high carrier frequencies, the ratio of the gap band (600 Hz) to the carrier frequency may be too small, and, thus, a transition of 40 dB in amplitude over 600 Hz may pose a problem. In such a case, the modulation is carried out using a smaller carrier frequency ( $\omega_{c1}$ ) first. The resulting SSB signal effectively widens the gap to  $2\omega_{c1}$  (see Fig. 4.19c). Now, treating this signal as the new baseband signal, it is possible to SSB-modulate the high-frequency carrier.



an ideal phase shifter is also unrealizable. We can, at most, approximate it over a finite band. However, it is possible to realize a filter with two outputs such that both outputs have the same (constant) amplitude spectrum, but their phase spectra differ by  $\pi/2$  rad over a given band of frequencies.\*

In terms of bandwidth requirement, SSB is similar to QAM but less exacting in terms of the carrier frequency and phase or the requirement of a distortionless transmission medium. However, SSB is difficult to generate if the baseband signal has no dc null in its spectrum. It is easy to build a circuit to shift the phase of a single frequency component by  $\pi/2$  rad. But a device to achieve a  $\pi/2$  phase shift of all the spectral components over a band of frequencies is unrealizable. We can, at best, approximate it over a finite band.

### Demodulation of SSB-SC Signals

It was shown earlier that SSB-SC signals can be coherently demodulated. We can readily verify this in another way:

$$\varphi_{SSB}(t) = m(t) \cos \omega_c t \mp m_h(t) \sin \omega_c t$$

Hence,

$$\begin{aligned} \varphi_{SSB}(t) \cos \omega_c t &= \frac{1}{2}m(t)[1 + \cos 2\omega_c t] \mp \frac{1}{2}m_h(t) \sin 2\omega_c t \\ &= \frac{1}{2}m(t) + \frac{1}{2}[m(t) \cos 2\omega_c t \mp m_h(t) \sin 2\omega_c t] \end{aligned}$$

Thus, the product  $\varphi_{SSB}(t) \cos \omega_c t$  yields the baseband signal and another SSB signal with a carrier  $2\omega_c$ . The spectrum in Fig. 4.15e shows precisely this result. A low-pass filter will suppress the unwanted SSB terms, giving the desired baseband signal  $m(t)/2$ . Hence, the demodulator is identical to the synchronous demodulator used for DSB-SC. Thus, any one of the synchronous DSB-SC demodulators discussed in Sec. 4.2 can be used to demodulate an SSB-SC signal.

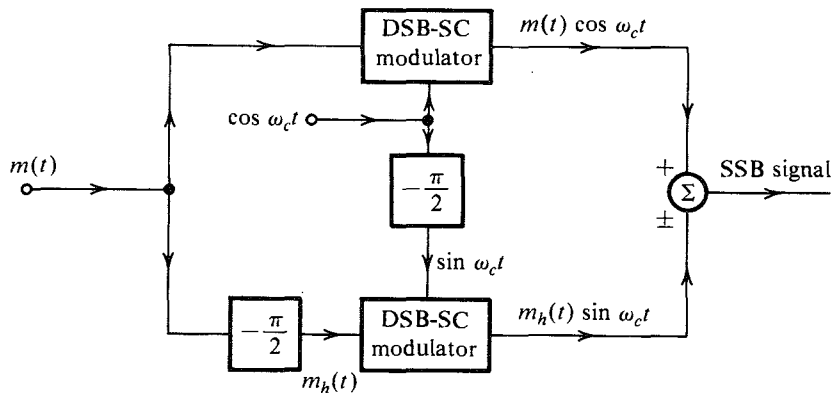


Figure 4.20 SSB generation by phase-shift method.

\* In this case, the phase spectrum of one output may be  $\phi(\omega)$  while that of the other is  $\phi(\omega) - \pi/2$ . The term  $\phi(\omega)$  is an unwanted phase distortion. However, as seen earlier, human ear is not sensitive to this kind of distortion.

**Envelope Detection of SSB Signals with a Carrier (SSB+C):** We now consider SSB signals with an additional carrier (SSB+C). Such a signal can be expressed as

$$\varphi_{\text{SSB+C}} = A \cos \omega_c t + [m(t) \cos \omega_c t + m_h(t) \sin \omega_c t]$$

Although  $m(t)$  can be recovered by synchronous detection [multiplying  $\varphi_{\text{SSB+C}}$  by  $\cos \omega_c t$  if  $A$ , the carrier amplitude, is large enough,  $m(t)$  can also be recovered from  $\varphi_{\text{SSB+C}}$  by envelope or rectifier detection. This can be shown by rewriting  $\varphi_{\text{SSB+C}}$  as

$$\begin{aligned} \varphi_{\text{SSB+C}} &= [A + m(t)] \cos \omega_c t + m_h(t) \sin \omega_c t \\ &= E(t) \cos (\omega_c t + \theta) \end{aligned}$$

where  $E(t)$ , the envelope of  $\varphi_{\text{SSB+C}}$ , is given by [see Eq. (3.39)]

$$\begin{aligned} E(t) &= \{[A + m(t)]^2 + m_h^2(t)\}^{1/2} \\ &= A \left[ 1 + \frac{2m(t)}{A} + \frac{m^2(t)}{A^2} + \frac{m_h^2(t)}{A^2} \right]^{1/2} \end{aligned}$$

If  $A \gg |m(t)|$ , then in general\*  $A \gg |m_h(t)|$ , and the terms  $m^2(t)/A^2$  and  $m_h^2(t)/A^2$  can be ignored. Thus,

$$E(t) \simeq A \left[ 1 + \frac{2m(t)}{A} \right]^{1/2}$$

Using binomial expansion and discarding higher order terms [because  $m(t)/A \ll 1$ ], we get

$$\begin{aligned} E(t) &\simeq A \left[ 1 + \frac{m(t)}{A} \right] \\ &= A + m(t) \end{aligned}$$

It is evident that for a large carrier, the SSB + C can be demodulated by an envelope detector.

In AM, envelope detection requires the condition  $A \geq |m(t)|$ , whereas for SSB+C, the condition is  $A \gg |m(t)|$ . Hence, in SSB case, the required carrier amplitude is much larger than that in AM, and, consequently, the efficiency of SSB+C is pathetically low.

### Telephone-Channel Multiplexing

Until recently, almost all long-haul telephone channels were multiplexed by FDM using SSB signals. This multiplexing technique, standardized by the CCITT, provides considerable flexibility in branching, dropping off, or inserting blocks of channels at points en route.<sup>4</sup> A basic **group** consists of 12 frequency-division multiplexed SSB voice channels, each of bandwidth 4 kHz (first-level multiplexing). A basic group uses LSB spectra and occupies a band of 60 to 108 kHz. An alternate group configuration of 12 USB voice signals, occupying a band of 148 to 196 kHz, is also used.

A basic **supergroup** of 60 channels is formed by multiplexing five basic groups, and it occupies a band of 312 to 552 kHz. An alternate supergroup configuration using USB spectra occupies a band of 60 to 300 kHz.

\* This may not be true for all  $t$ , but it is true for most  $t$ .

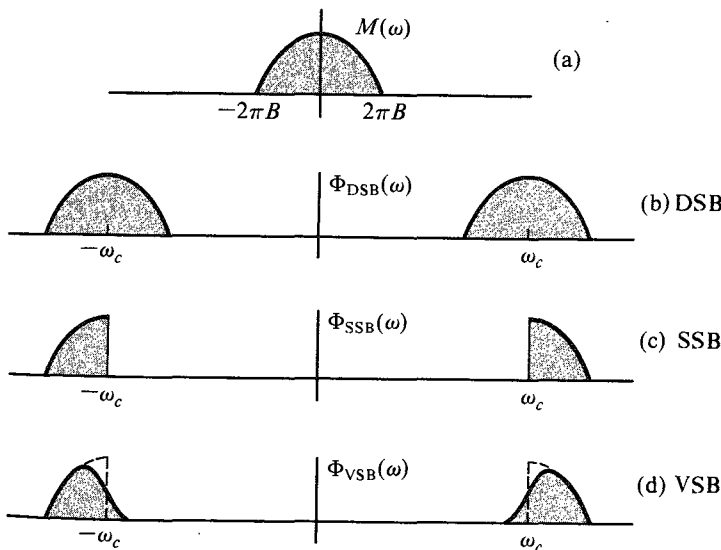
A basic **mastergroup** of 600 channels is formed by multiplexing 10 supergroups.\* There are two standard mastergroup configurations: the L600 and the U600.

Modern broad-band transmission systems can transmit even larger groupings than mastergroups. For the L3 carrier and TH microwave, three mastergroups and one supergroup comprising 1860 message channels are combined. The L4 system utilizes six U600 mastergroups multiplexed to form 3600 channels. The multiplexed signal is fed into the baseband input of a microwave radio channel or directly into a coaxial transmission system.

## 4.6 AMPLITUDE MODULATION: VESTIGIAL SIDEBAND (VSB)

As seen earlier, the generation of SSB signals is rather difficult. The selective-filtering method demands dc null in the modulating signal spectrum. A phase shifter required in the phase-shift method is unrealizable, or realizable only approximately. The generation of DSB signals is much simpler, but requires twice the signal bandwidth. A **vestigial-sideband (VSB)**, also called asymmetric sideband system is a compromise between DSB and SSB. It inherits the advantages of DSB and SSB but avoids their disadvantages at a small cost. VSB signals are relatively easy to generate, and, at the same time, their bandwidth is only (typically 25%) greater than that of SSB signals.

In VSB, instead of rejecting one sideband completely (as in SSB), a gradual cutoff of one sideband, as shown in Fig. 4.21d, is accepted. The baseband signal can be recovered exactly by a synchronous detector in conjunction with an appropriate equalizer filter  $H_o(\omega)$  at the receiver output (Fig. 4.22). If a large carrier is transmitted along with the VSB signal, the baseband signal can be recovered by an envelope (or a rectifier) detector.



**Figure 4.21** Spectra of the modulating signal and corresponding DSB, SSB, and VSB signals.

\* This is true for the North American hierarchy. In the CCITT hierarchy, a basic mastergroup is formed by multiplexing five supergroups (300 voice channels).

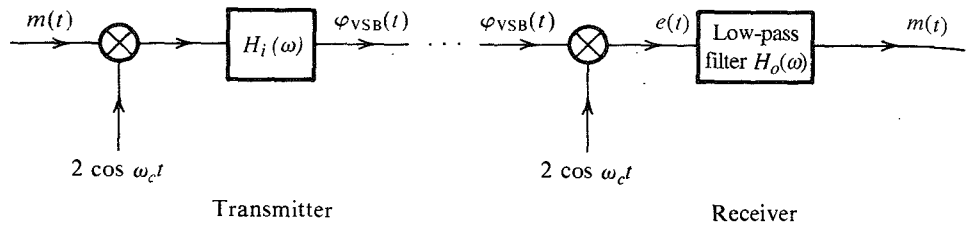


Figure 4.22 VSB modulator and demodulator.

If the vestigial shaping filter that produces VSB from DSB is  $H_i(\omega)$  (Fig. 4.22), then the resulting VSB signal spectrum is

$$\Phi_{\text{VSB}}(\omega) = [M(\omega + \omega_c) + M(\omega - \omega_c)]H_i(\omega) \quad (4.18)$$

This VSB shaping filter  $H_i(\omega)$  allows the transmission of one sideband, but suppresses the other sideband, not completely, but gradually. This makes it easy to realize such a filter, but the transmission bandwidth is now somewhat higher than that of the SSB (where the other sideband is suppressed completely). The bandwidth of the VSB signal is typically 25 to 33% higher than that of the SSB signals.

We require that  $m(t)$  be recoverable from  $\varphi_{\text{VSB}}(t)$  using synchronous demodulation at the receiver. This is done by multiplying the incoming VSB signal  $\varphi_{\text{VSB}}(t)$  by  $2 \cos \omega_c t$ . The product  $e(t)$  is given by

$$e(t) = 2\varphi_{\text{VSB}}(t) \cos \omega_c t \iff [\Phi_{\text{VSB}}(\omega + \omega_c) + \Phi_{\text{VSB}}(\omega - \omega_c)]$$

The signal  $e(t)$  is further passed through the low-pass equalizer filter of transfer function  $H_o(\omega)$ . The output of the equalizer filter is required to be  $m(t)$ . Hence, the output signal spectrum is given by

$$M(\omega) = [\Phi_{\text{VSB}}(\omega + \omega_c) + \Phi_{\text{VSB}}(\omega - \omega_c)]H_o(\omega)$$

Substituting Eq. (4.18) into this equation and eliminating the spectra at  $\pm 2\omega_c$  [suppressed by a low-pass filter  $H_o(\omega)$ ], we obtain

$$M(\omega) = M(\omega)[H_i(\omega + \omega_c) + H_i(\omega - \omega_c)]H_o(\omega) \quad (4.19)$$

Hence\*

$$H_o(\omega) = \frac{1}{H_i(\omega + \omega_c) + H_i(\omega - \omega_c)} \quad |\omega| \leq 2\pi B \quad (4.20)$$

Note that because  $H_i(\omega)$  is a bandpass filter, the terms  $H_i(\omega \pm \omega_c)$  contain low-pass components.

\* If we choose  $H_i(\omega)$  such that

$$H_i(\omega + \omega_c) + H_i(\omega - \omega_c) = 1 \quad |\omega| \leq 2\pi B \quad (4.21a)$$

The output filter is just a simple low-pass filter with transfer function  $H_o(\omega) = 1$  over the baseband  $|\omega| \leq 2\pi B$ , because for a real filter,  $H_i(-\omega) = H_i^*(\omega)$ . Eq. (4.21a) can be expressed as

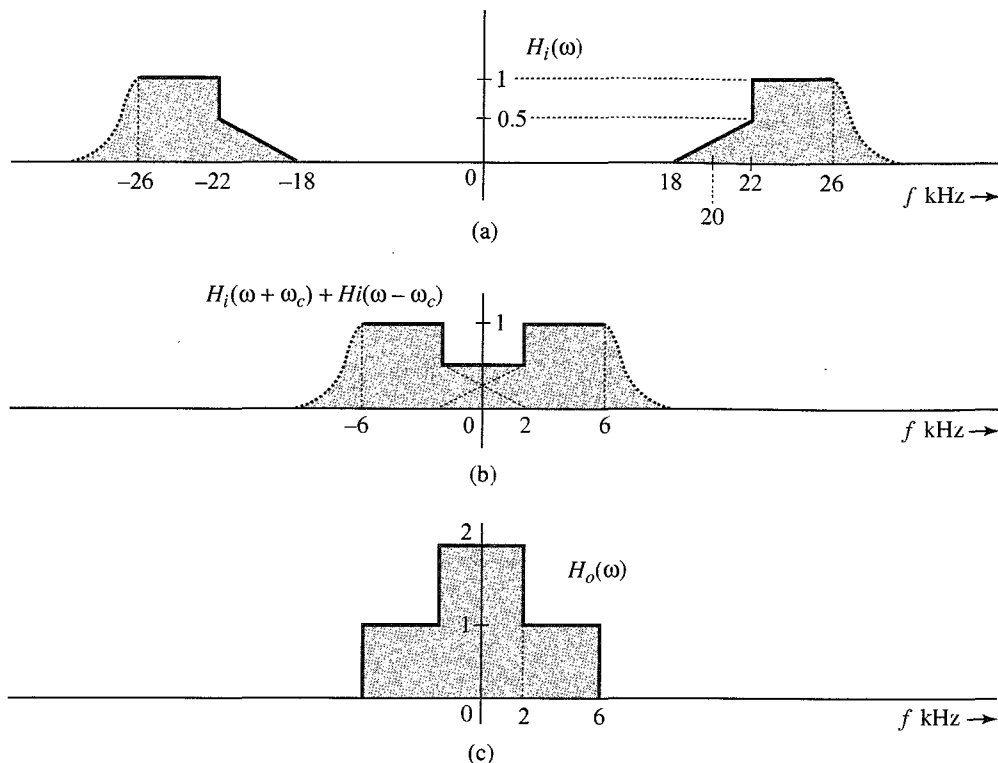
$$H_i(\omega_c + \omega) + H_i^*(\omega_c - \omega) = 1 \quad |\omega| \leq 2\pi B \quad (4.21b)$$

or

$$H_i(\omega_c + x) + H_i^*(\omega_c - x) = 1 \quad |x| \leq 2\pi B \quad (4.21c)$$

**EXAMPLE 4.8**

The carrier frequency of a certain VSB signal is  $\omega_c = 20$  kHz, and the baseband signal bandwidth is 6 kHz. The VSB shaping filter  $H_i(\omega)$  at the input, which cuts off the lower sideband gradually over 2 kHz, is shown in Fig. 4.23a. Find the output filter  $H_o(\omega)$  required for distortionless reception.



**Figure 4.23** VSB out filter.

Figure 4.23b shows the low-pass segments of  $H_i(\omega + \omega_c) + H_i(\omega - \omega_c)$ . We are interested in this spectrum only over the baseband (the remaining undesired portion is suppressed by the output filter). This spectrum is 0.5 over the band of 0 to 2 kHz, and is 1 over 2 to 6 kHz, as shown in Fig. 4.23b. Figure 4.23c shows the desired output filter  $H_o(\omega)$ , which is the reciprocal of the spectrum in Fig. 4.23b [see Eq. (4.20)].

### Envelope Detection of VSB+C Signals

That VSB+C signals can be envelope detected may be proved by using exactly the same argument used in proving the case for SSB+C signals. Both the SSB and the VSB modulated signals have the same form, with  $m_h(t)$  in SSB replaced by some other signal  $m_s(t)$  in VSB. This is because the VSB signal is a bandpass signal, which can be expressed in terms of the quadrature components in Eq. (3.38).

We have shown that VSB+C requires a much larger carrier than DSB+C (AM) for envelope detection. Because VSB+C is an in-between case, the added carrier required in VSB is larger than that in AM, but smaller than that in SSB+C.

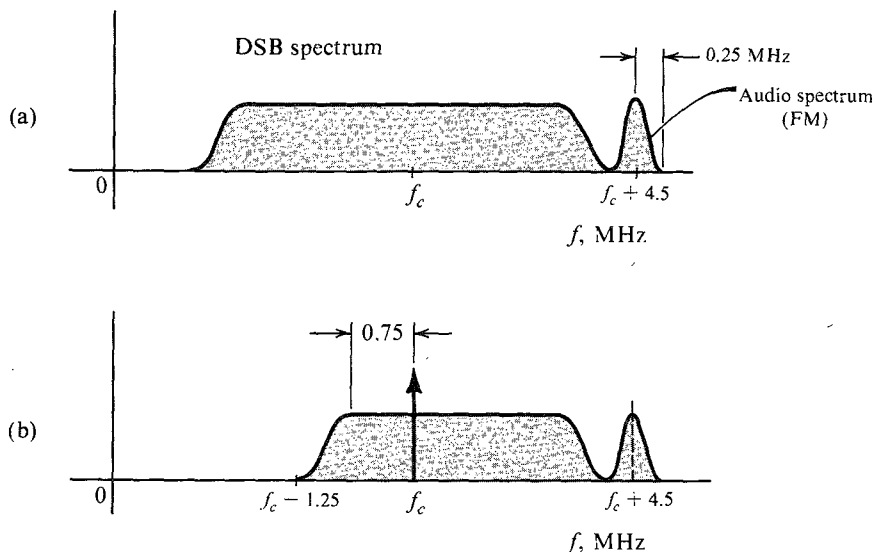
### Use of VSB in Broadcast Television

VSB is a clever compromise between SSB and DSB, which makes it very attractive for television broadcast systems. The baseband video signal of television occupies an enormous bandwidth of 4.5 MHz, and a DSB signal needs a bandwidth of 9 MHz. It would seem desirable to use SSB in order to conserve the bandwidth. Unfortunately, this creates several problems. First, the baseband video signal has sizable power in the low-frequency region, and consequently it is difficult to suppress one sideband completely. Second, for a broadcast receiver, an envelope detector is preferred over a synchronous one in order to reduce the receiver cost. We have seen earlier that SSB+C has a very low power efficiency. Moreover, use of SSB will increase the receiver cost.

The DSB spectrum of a television signal is shown in Fig. 4.24a. The vestigial shaping filter  $H_i(\omega)$  cuts off the lower sideband spectrum gradually starting at 0.75 MHz to 1.25 MHz below the carrier frequency  $f_c$ , as shown in Fig. 4.24b. The receiver output filter  $H_o(\omega)$  is designed according to Eq. (4.20). The resulting VSB spectrum bandwidth is 6 MHz. Compare this with the DSB bandwidth of 9 MHz and the SSB bandwidth of 4.5 MHz.

### Linearity of Amplitude Modulation

In all the types of modulation discussed thus far, the modulated signal (excluding the carrier term) satisfies the principles of superposition. For example, if modulating signals  $m_1(t)$  and  $m_2(t)$  produce modulated signals  $\varphi_1(t)$  and  $\varphi_2(t)$ , respectively, then the modulating signal  $k_1 m_1(t) + k_2 m_2(t)$  produces the modulated signal  $k_1 \varphi_1(t) + k_2 \varphi_2(t)$ . The reader can verify linearity for all types of amplitude modulation (DSB, SSB, AM, and VSB). This property is valuable in analysis. Because any signal can be expressed as a sum (discrete or in continuum)



**Figure 4.24** Television signal spectra. (a) DSB signal. (b) Signal transmitted.

\* Note that we are excluding the carrier term from  $\varphi_1(t)$  and  $\varphi_2(t)$ . In short, superposition applies to the suppressed carrier portion only. For more discussion, see Van Trees.<sup>5</sup>

of sinusoids, the complete description of the modulation system can be expressed in terms of tone modulation. For example, if  $m(t) = \cos \omega_m t$  (tone modulation), the DSB-SC signal is

$$\cos \omega_m t \cos \omega_c t = \frac{1}{2} [\cos (\omega_c - \omega_m)t + \cos (\omega_c + \omega_m)t]$$

This shows that DSB-SC translates a frequency  $\omega_m$  to two frequencies,  $\omega_c - \omega_m$  (LSB) and  $\omega_c + \omega_m$  (USB). We can generalize this result to any nonsinusoidal modulating signal  $m(t)$ . This is precisely the result obtained earlier by using a more general analysis.

## 4.7 CARRIER ACQUISITION

In the suppressed-carrier amplitude-modulated system (DSB-SC, SSB-SC, and VSB-SC), one must generate a local carrier at the receiver for the purpose of synchronous demodulation. Ideally, the local carrier must be in frequency and phase synchronism with the incoming carrier. Any discrepancy in the frequency or phase of the local carrier gives rise to distortion in the detector output.

Consider a DSB-SC case where a received signal is  $m(t) \cos \omega_c t$  and the local carrier is  $2 \cos [(\omega_c + \Delta\omega)t + \delta]$ . The local-carrier frequency and phase errors in this case are  $\Delta\omega$  and  $\delta$ , respectively. The product of the received signal and the local carrier is  $e(t)$ , given by

$$\begin{aligned} e(t) &= 2m(t) \cos \omega_c t \cos [(\omega_c + \Delta\omega)t + \delta] \\ &= m(t) \{ \cos [(\Delta\omega)t + \delta] + \cos [(2\omega_c + \Delta\omega)t + \delta] \} \end{aligned} \quad (4.22)$$

The second term on the right-hand side is filtered out by the low-pass filter, leaving the output  $e_o(t)$  as

$$e_o(t) = m(t) \cos [(\Delta\omega)t + \delta] \quad (4.23)$$

If  $\Delta\omega$  and  $\delta$  are both zero (no frequency or phase error), then

$$e_o(t) = m(t)$$

as expected. Let us consider two special cases. If  $\Delta\omega = 0$ , Eq. (4.23) reduces to

$$e_o(t) = m(t) \cos \delta \quad (4.24a)$$

This output is proportional to  $m(t)$  when  $\delta$  is a constant. The output is maximum when  $\delta = 0$  and minimum (zero) when  $\delta = \pm\pi/2$ . Thus, the phase error in the local carrier causes the attenuation of the output signal without causing any distortion, as long as  $\delta$  is constant. Unfortunately, the phase error  $\delta$  may vary randomly with time. This may occur, for example, because of variations in the propagation path. This causes the gain factor  $\cos \delta$  at the receiver to vary randomly and is undesirable.

Next we consider the case where  $\delta = 0$  and  $\Delta\omega \neq 0$ . In this case, Eq. (4.23) becomes

$$e_o(t) = m(t) \cos (\Delta\omega)t \quad (4.24b)$$

The output here is not merely an attenuated replica of the original signal but is also distorted. Because  $\Delta\omega$  is usually small, the output is the signal  $m(t)$  multiplied by a low-frequency sinusoid. This causes the amplitude of the desired signal  $m(t)$  to vary from maximum to zero

periodically at twice the period of the beat frequency  $\Delta\omega$ . This “beating” effect is catastrophic even for a small frequency difference. The effect of this distortion even for a small frequency mismatch, say  $\Delta f = 1$  Hz, is similar to the output when some restless kid is fiddling with its volume control knob up and down continuously twice a second.

To ensure identical carrier frequencies at the transmitter and the receiver, we can use quartz crystal oscillators, which generally are very stable. Identical crystals are cut to yield the same frequency at the transmitter and the receiver. At very high carrier frequencies, where the crystal dimensions become too small to match exactly, quartz-crystal performance may not be adequate. In such a case, a carrier, or **pilot**, is transmitted at a reduced level (usually about  $-20$  dB) along with the sidebands. The pilot is separated at the receiver by a very narrow-band filter tuned to the pilot frequency. It is amplified and used to synchronize the local oscillator. The phase-locked loop (PLL), which plays an important role in carrier acquisition, will now be discussed.

The nature of the distortion caused by asynchronous carrier in SSB-SC is somewhat different than that in DSB-SC. In SSB-SC, when the carrier at the receiver is  $2 \cos[(\omega_c + \Delta\omega)t]$ , the output is  $m(t)$  with all its spectral components shifted (offset) by  $\Delta\omega$  (see Prob. 4.5-5). Such a shift of every frequency component by a fixed amount  $\Delta\omega$  destroys the harmonic relationship between frequency components. For instance, if  $\Delta f = 10$  Hz, then the components of frequencies 1000 and 2000 Hz will be shifted to frequencies 1010 and 2010. This destroys their harmonic relationship. But unless  $\Delta f$  is very large, such a change does not destroy intelligibility of the output (as the beating effect does in the case of DSB-SC). For audio signals  $\Delta f < 30$  Hz does not significantly affect the signal quality.  $\Delta f > 30$  Hz results in a sound quality similar to that of Donald Duck. But the intelligibility is not completely lost.

When the carrier is  $\cos(\omega_c t + \theta)$ , the output is the signal  $m(t)$  with the phases of all its spectral components shifted by  $\theta$  (see Prob. 4.5-5). The phase distortion in SSB-SC also gives rise to the Donald Duck sound effect. This discussion shows that the problem of carrier synchronization is more critical in DSB-SC than in SSB-SC.

### Phase-Locked Loop (PLL)

The **phase-locked loop (PLL)** can be used to track the phase and the frequency of the carrier component of an incoming signal. It is, therefore, a useful device for synchronous demodulation of AM signals with suppressed carrier or with a little carrier (the pilot). It can also be used for the demodulation of angle-modulated signals, especially under low SNR conditions. For this reason, the PLL is used in such applications as space-vehicle-to-earth data links, where there is a premium on transmitter weight, or where the loss along the transmission path is very large; and, more recently, in commercial FM receivers.

A PLL has three basic components:

1. A voltage-controlled oscillator (VCO)
2. A multiplier, serving as a phase detector (PD) or a phase comparator
3. A loop filter  $H(s)$

The operation of the PLL is similar to that of a feedback system (Fig. 4.25a). In a typical feedback system, the signal fed back tends to follow the input signal. If the signal fed back is not equal to the input signal, the difference (known as the error) will change the signal fed back until it is close to the input signal. A PLL operates on a similar principle, except that the quantity fed back and compared is not the amplitude, but the phase. The VCO adjusts its own