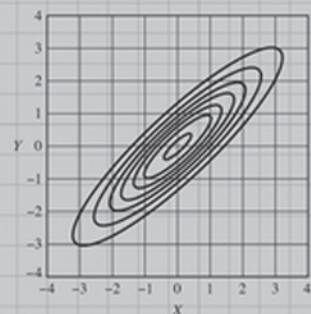
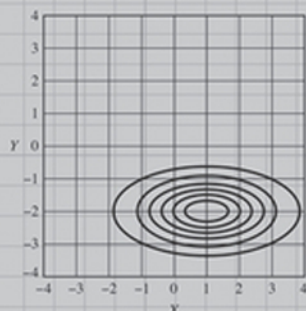
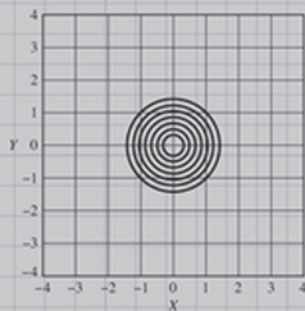
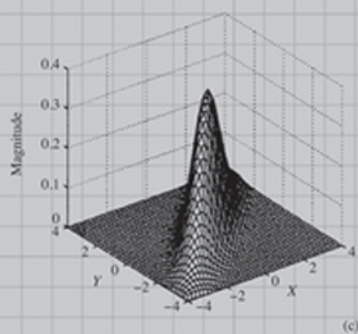
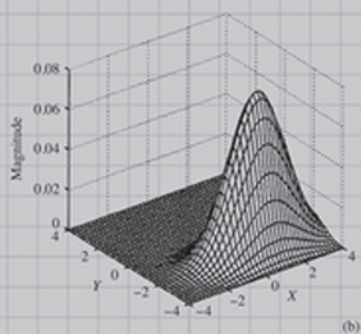
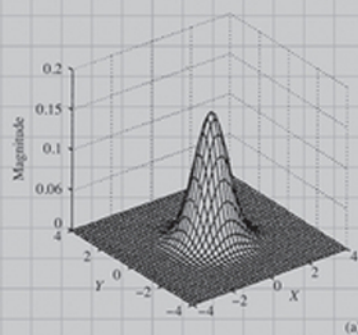


Principles of COMMUNICATIONS Systems, Modulation, and Noise

RODGER E. ZIEMER • WILLIAM H. TRANTER



WILEY

PRINCIPLES OF COMMUNICATIONS

Systems, Modulation,
and Noise

SEVENTH EDITION

RODGER E. ZIEMER

University of Colorado at Colorado Springs

WILLIAM H. TRANTER

Virginia Polytechnic Institute and State University

WILEY

VP & PUBLISHER:	Don Fowley
EXECUTIVE EDITOR:	Dan Sayre
SPONSORING EDITOR:	Mary O'Sullivan
PROJECT EDITOR:	Ellen Keohane
COVER DESIGNER:	Kenji Ngieng
ASSOCIATE PRODUCTION MANAGER:	Joyce Poh
SENIOR PRODUCTION EDITOR:	Jolene Ling
PRODUCTION MANAGEMENT SERVICES:	Thomson Digital
COVER ILLUSTRATION CREDITS:	© Rodger E. Ziemer, William H. Tranter

This book was set by Thomson Digital.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

Copyright © 2015, 2009, 2002, 1995 John Wiley & Sons, Inc. 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, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc. 222 Rosewood Drive, Danvers, MA 01923, website www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, (201)748-6011, fax (201)748-6008, website <http://www.wiley.com/go/permissions>.

Evaluation copies are provided to qualified academics and professionals for review purposes only, for use in their courses during the next academic year. These copies are licensed and may not be sold or transferred to a third party. Upon completion of the review period, please return the evaluation copy to Wiley. Return instructions and a free of charge return mailing label are available at www.wiley.com/go/returnlabel. If you have chosen to adopt this textbook for use in your course, please accept this book as your complimentary desk copy. Outside of the United States, please contact your local sales representative.

Library of Congress Cataloging-in-Publication Data:

Ziemer, Rodger E.

Principles of communication : systems, modulation, and noise / Rodger E. Ziemer,
William H. Tranter. — Seventh edition.

pages cm

Includes bibliographical references and index.

ISBN 978-1-118-07891-4 (paper)

1. Telecommunication. 2. Signal theory (Telecommunication) I. Tranter,

William H. II. Title.

TK5105.Z54 2014

621.382'2—dc23

2013034294

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

The first edition of this book was published in 1976, less than a decade after Neil Armstrong became the first man to walk on the moon in 1969. The programs that lead to the first moon landing gave birth to many advances in science and technology. A number of these advances, especially those in microelectronics and digital signal processing (DSP), became enabling technologies for advances in communications. For example, prior to 1969, essentially all commercial communication systems, including radio, telephones, and television, were analog. Enabling technologies gave rise to the internet and the World Wide Web, digital radio and television, satellite communications, Global Positioning Systems, cellular communications for voice and data, and a host of other applications that impact our daily lives. A number of books have been written that provide an in-depth study of these applications. In this book we have chosen not to cover application areas in detail but, rather, to focus on basic theory and fundamental techniques. A firm understanding of basic theory prepares the student to pursue study of higher-level theoretical concepts and applications.

True to this philosophy, we continue to resist the temptation to include a variety of new applications and technologies in this edition and believe that application examples and specific technologies, which often have short lifetimes, are best treated in subsequent courses after students have mastered the basic theory and analysis techniques. Reactions to previous editions have shown that emphasizing fundamentals, as opposed to specific technologies, serve the user well while keeping the length of the book reasonable. This strategy appears to have worked well for advanced undergraduates, for new graduate students who may have forgotten some of the fundamentals, and for the working engineer who may use the book as a reference or who may be taking a course after-hours. New developments that appear to be fundamental, such as multiple-input multiple-output (MIMO) systems and capacity-approaching codes, are covered in appropriate detail.

The two most obvious changes to the seventh edition of this book are the addition of drill problems to the Problems section at the end of each chapter and the division of chapter three into two chapters. The drill problems provide the student problem-solving practice with relatively simple problems. While the solutions to these problems are straightforward, the complete set of drill problems covers the important concepts of each chapter. Chapter 3, as it appeared in previous editions, is now divided into two chapters mainly due to length. Chapter 3 now focuses on linear analog modulation and simple discrete-time modulation techniques that are direct applications of the sampling theorem. Chapter 4 now focuses on nonlinear modulation techniques. A number of new or revised end-of-chapter problems are included in all chapters.

In addition to these obvious changes, a number of other changes have been made in edition seven. An example on signal space was deleted from Chapter 2 since it is really not necessary at this point in the book. (Chapter 11 deals more fully with the concepts of signal space.) Chapter 3, as described in the previous paragraph, now deals with linear analog modulation techniques. A section on measuring the modulation index of AM signals and measuring transmitter linearity has been added. The section on analog television has been deleted from Chapter 3 since it is no longer relevant. Finally, the section on adaptive delta modulation has been deleted. Chapter 4 now deals with non-linear analog modulation techniques. Except for the problems, no significant additions or deletions have been made to Chapter 5. The same is true of Chapters 6 and 7, which treat probability and random processes, respectively. A section on signal-to-noise ratio measurement has been added to Chapter 8, which treats noise effects in modulation systems. More detail on basic channel

models for fading channels has been added in Chapter 9 along with simulation results for bit error rate (BER) performance of a minimum mean-square error (MMSE) equalizer with optimum weights and an additional example of the MMSE equalizer with adaptive weights. Several changes have been made to Chapter 10. Satellite communications was reluctantly deleted because it would have required adding several additional pages to do it justice. A section was added on MIMO systems using the Alamouti approach, which concludes with a BER curve comparing performances of 2-transmit 1-receive Alamouti signaling in a Rayleigh fading channel with a 2-transmit 2-receive diversity system. A short discussion was also added to Chapter 10 illustrating the features of 4G cellular communications as compared with 2G and 3G systems. With the exception of the Problems, no changes have been made to Chapter 11. A “Quick Overview” section has been added to Chapter 12 discussing capacity-approaching codes, run-length codes, and digital television.

A feature of the later editions of *Principles of Communications* was the inclusion of several computer examples within each chapter. (MATLAB was chosen for these examples because of its widespread use in both academic and industrial settings, as well as for MATLAB’s rich graphics library.) These computer examples, which range from programs for computing performance curves to simulation programs for certain types of communication systems and algorithms, allow the student to observe the behavior of more complex systems without the need for extensive computations. These examples also expose the student to modern computational tools for analysis and simulation in the context of communication systems. Even though we have limited the amount of this material in order to ensure that the character of the book is not changed, the number of computer examples has been increased for the seventh edition. In addition to the in-chapter computer examples, a number of “computer exercises” are included at the end of each chapter. The number of these has also been increased in the seventh edition. These exercises follow the end-of-chapter problems and are designed to make use of the computer in order to illustrate basic principles and to provide the student with additional insight. A number of new problems have been included at the end of each chapter in addition to a number of problems that were revised from the previous edition.

The publisher maintains a web site from which the source code for all in-chapter computer examples can be downloaded. Also included on the web site are Appendix G (answers to the drill problems) and the bibliography. The URL is

www.wiley.com/college/ziemer

We recommend that, although MATLAB code is included in the text, students download MATLAB code of interest from the publisher website. The code in the text is subject to printing and other types of errors and is included to give the student insight into the computational techniques used for the illustrative examples. In addition, the MATLAB code on the publisher website is periodically updated as need justifies. This web site also contains complete solutions for the end-of-chapter problems and computer exercises. (The solutions manual is password protected and is intended only for course instructors.)

We wish to thank the many persons who have contributed to the development of this textbook and who have suggested improvements for this and previous editions of this book. We also express our thanks to the many colleagues who have offered suggestions to us by correspondence or verbally as well as the industries and agencies that have supported our research. We especially thank our colleagues and students at the University of Colorado at Colorado Springs, the Missouri University of Science and Technology, and Virginia Tech for their comments and suggestions. It is to our students that we dedicate this book. We have worked with many people over the past forty years and many of them have helped shape our teaching and research philosophy. We thank them all.

Finally, our families deserve much more than a simple thanks for the patience and support that they have given us throughout forty years of seemingly endless writing projects.

Rodger E. Ziemer
William H. Tranter

CONTENTS

CHAPTER 1

INTRODUCTION 1

- 1.1 The Block Diagram of a Communication System 4
- 1.2 Channel Characteristics 5
 - 1.2.1 Noise Sources 5
 - 1.2.2 Types of Transmission Channels 7
- 1.3 Summary of Systems-Analysis Techniques 13
 - 1.3.1 Time and Frequency-Domain Analyses 13
 - 1.3.2 Modulation and Communication Theories 13
- 1.4 Probabilistic Approaches to System Optimization 14
 - 1.4.1 Statistical Signal Detection and Estimation Theory 14
 - 1.4.2 Information Theory and Coding 15
 - 1.4.3 Recent Advances 16
- 1.5 Preview of This Book 16
- Further Reading 16

CHAPTER 2

SIGNAL AND LINEAR SYSTEM ANALYSIS 17

- 2.1 Signal Models 17
 - 2.1.1 Deterministic and Random Signals 17
 - 2.1.2 Periodic and Aperiodic Signals 18
 - 2.1.3 Phasor Signals and Spectra 18
 - 2.1.4 Singularity Functions 21
- 2.2 Signal Classifications 24
- 2.3 Fourier Series 26
 - 2.3.1 Complex Exponential Fourier Series 26
 - 2.3.2 Symmetry Properties of the Fourier Coefficients 27
 - 2.3.3 Trigonometric Form of the Fourier Series 28
 - 2.3.4 Parseval's Theorem 28
 - 2.3.5 Examples of Fourier Series 29
 - 2.3.6 Line Spectra 30
- 2.4 The Fourier Transform 34
 - 2.4.1 Amplitude and Phase Spectra 35
 - 2.4.2 Symmetry Properties 36
 - 2.4.3 Energy Spectral Density 37

- 2.4.4 Convolution 38
- 2.4.5 Transform Theorems: Proofs and Applications 40
- 2.4.6 Fourier Transforms of Periodic Signals 48
- 2.4.7 Poisson Sum Formula 50
- 2.5 Power Spectral Density and Correlation 50
 - 2.5.1 The Time-Average Autocorrelation Function 51
 - 2.5.2 Properties of $R(\tau)$ 52
- 2.6 Signals and Linear Systems 55
 - 2.6.1 Definition of a Linear Time-Invariant System 56
 - 2.6.2 Impulse Response and the Superposition Integral 56
 - 2.6.3 Stability 58
 - 2.6.4 Transfer (Frequency Response) Function 58
 - 2.6.5 Causality 58
 - 2.6.6 Symmetry Properties of $H(f)$ 59
 - 2.6.7 Input-Output Relationships for Spectral Densities 62
 - 2.6.8 Response to Periodic Inputs 62
 - 2.6.9 Distortionless Transmission 64
 - 2.6.10 Group and Phase Delay 64
 - 2.6.11 Nonlinear Distortion 67
 - 2.6.12 Ideal Filters 68
 - 2.6.13 Approximation of Ideal Lowpass Filters by Realizable Filters 70
 - 2.6.14 Relationship of Pulse Resolution and Risettime to Bandwidth 75
- 2.7 Sampling Theory 78
- 2.8 The Hilbert Transform 82
 - 2.8.1 Definition 82
 - 2.8.2 Properties 83
 - 2.8.3 Analytic Signals 85
 - 2.8.4 Complex Envelope Representation of Bandpass Signals 87
 - 2.8.5 Complex Envelope Representation of Bandpass Systems 89
- 2.9 The Discrete Fourier Transform and Fast Fourier Transform 91
- Further Reading 95

vi Contents

Summary	95
Drill Problems	98
Problems	100
Computer Exercises	111

CHAPTER 3**LINEAR MODULATION TECHNIQUES 112**

3.1 Double-Sideband Modulation	113
3.2 Amplitude Modulation (AM)	116
3.2.1 Envelope Detection	118
3.2.2 The Modulation Trapezoid	122
3.3 Single-Sideband (SSB) Modulation	124
3.4 Vestigial-Sideband (VSB) Modulation	133
3.5 Frequency Translation and Mixing	136
3.6 Interference in Linear Modulation	139
3.7 Pulse Amplitude Modulation—PAM	142
3.8 Digital Pulse Modulation	144
3.8.1 Delta Modulation	144
3.8.2 Pulse-Code Modulation	146
3.8.3 Time-Division Multiplexing	147
3.8.4 An Example: The Digital Telephone System	149

Further Reading 150**Summary 150****Drill Problems 151****Problems 152****Computer Exercises 155****CHAPTER 4****ANGLE MODULATION AND MULTIPLEXING 156**

4.1 Phase and Frequency Modulation Defined	156
4.1.1 Narrowband Angle Modulation	157
4.1.2 Spectrum of an Angle-Modulated Signal	161
4.1.3 Power in an Angle-Modulated Signal	168
4.1.4 Bandwidth of Angle-Modulated Signals	168
4.1.5 Narrowband-to-Wideband Conversion	173
4.2 Demodulation of Angle-Modulated Signals	175
4.3 Feedback Demodulators: The Phase-Locked Loop	181
4.3.1 Phase-Locked Loops for FM and PM Demodulation	181
4.3.2 Phase-Locked Loop Operation in the Tracking Mode: The Linear Model	184
4.3.3 Phase-Locked Loop Operation in the Acquisition Mode	189
4.3.4 Costas PLLs	194
4.3.5 Frequency Multiplication and Frequency Division	195
4.4 Interference in Angle Modulation	196

4.5 Analog Pulse Modulation 201

4.5.1 Pulse-Width Modulation (PWM)	201
4.5.2 Pulse-Position Modulation (PPM)	203

4.6 Multiplexing 204

4.6.1 Frequency-Division Multiplexing	204
4.6.2 Example of FDM: Stereophonic FM Broadcasting	205
4.6.3 Quadrature Multiplexing	206
4.6.4 Comparison of Multiplexing Schemes	207

Further Reading 208**Summary 208****Drill Problems 209****Problems 210****Computer Exercises 213****CHAPTER 5****PRINCIPLES OF BASEBAND DIGITAL DATA TRANSMISSION 215**

5.1 Baseband Digital Data Transmission Systems	215
5.2 Line Codes and Their Power Spectra	216
5.2.1 Description of Line Codes	216
5.2.2 Power Spectra for Line-Coded Data	218
5.3 Effects of Filtering of Digital Data—ISI	225
5.4 Pulse Shaping: Nyquist's Criterion for Zero ISI	227
5.4.1 Pulses Having the Zero ISI Property	228
5.4.2 Nyquist's Pulse-Shaping Criterion	229
5.4.3 Transmitter and Receiver Filters for Zero ISI	231
5.5 Zero-Forcing Equalization	233
5.6 Eye Diagrams	237
5.7 Synchronization	239
5.8 Carrier Modulation of Baseband Digital Signals	243
Further Reading	244
Summary	244
Drill Problems	245
Problems	246
Computer Exercises	249

CHAPTER 6**OVERVIEW OF PROBABILITY AND RANDOM VARIABLES 250**

6.1 What is Probability?	250
6.1.1 Equally Likely Outcomes	250
6.1.2 Relative Frequency	251
6.1.3 Sample Spaces and the Axioms of Probability	252
6.1.4 Venn Diagrams	253
6.1.5 Some Useful Probability Relationships	253

6.1.6	Tree Diagrams	257
6.1.7	Some More General Relationships	259
6.2	Random Variables and Related Functions	260
6.2.1	Random Variables	260
6.2.2	Probability (Cumulative) Distribution Functions	262
6.2.3	Probability-Density Function	263
6.2.4	Joint cdfs and pdfs	265
6.2.5	Transformation of Random Variables	270
6.3	Statistical Averages	274
6.3.1	Average of a Discrete Random Variable	274
6.3.2	Average of a Continuous Random Variable	275
6.3.3	Average of a Function of a Random Variable	275
6.3.4	Average of a Function of More Than One Random Variable	277
6.3.5	Variance of a Random Variable	279
6.3.6	Average of a Linear Combination of N Random Variables	280
6.3.7	Variance of a Linear Combination of Independent Random Variables	281
6.3.8	Another Special Average—The Characteristic Function	282
6.3.9	The pdf of the Sum of Two Independent Random Variables	283
6.3.10	Covariance and the Correlation Coefficient	285
6.4	Some Useful pdfs	286
6.4.1	Binomial Distribution	286
6.4.2	Laplace Approximation to the Binomial Distribution	288
6.4.3	Poisson Distribution and Poisson Approximation to the Binomial Distribution	289
6.4.4	Geometric Distribution	290
6.4.5	Gaussian Distribution	291
6.4.6	Gaussian Q -Function	295
6.4.7	Chebyshev's Inequality	296
6.4.8	Collection of Probability Functions and Their Means and Variances	296
	Further Reading	298
	Summary	298
	Drill Problems	300
	Problems	301
	Computer Exercises	307
	7.2.2	Description of Random Processes in Terms of Joint pdfs
	7.2.2	311
	7.2.3	Stationarity
	7.2.3	311
	7.2.4	Partial Description of Random Processes: Ergodicity
	7.2.4	312
	7.2.5	Meanings of Various Averages for Ergodic Processes
	7.2.5	315
7.3	Correlation and Power Spectral Density	316
7.3.1	Power Spectral Density	316
7.3.2	The Wiener–Khinchine Theorem	318
7.3.3	Properties of the Autocorrelation Function	320
7.3.4	Autocorrelation Functions for Random Pulse Trains	321
7.3.5	Cross-Correlation Function and Cross-Power Spectral Density	324
7.4	Linear Systems and Random Processes	325
7.4.1	Input-Output Relationships	325
7.4.2	Filtered Gaussian Processes	327
7.4.3	Noise-Equivalent Bandwidth	329
7.5	Narrowband Noise	333
7.5.1	Quadrature-Component and Envelope-Phase Representation	333
7.5.2	The Power Spectral Density Function of $n_c(t)$ and $n_s(t)$	335
7.5.3	Ricean Probability Density Function	338
	Further Reading	340
	Summary	340
	Drill Problems	341
	Problems	342
	Computer Exercises	348
	CHAPTER 8	
	NOISE IN MODULATION SYSTEMS	349
8.1	Signal-to-Noise Ratios	350
8.1.1	Baseband Systems	350
8.1.2	Double-Sideband Systems	351
8.1.3	Single-Sideband Systems	353
8.1.4	Amplitude Modulation Systems	355
8.1.5	An Estimator for Signal-to-Noise Ratios	361
8.2	Noise and Phase Errors in Coherent Systems	366
8.3	Noise in Angle Modulation	370
8.3.1	The Effect of Noise on the Receiver Input	370
8.3.2	Demodulation of PM	371
8.3.3	Demodulation of FM: Above Threshold Operation	372
8.3.4	Performance Enhancement through the Use of De-emphasis	374
8.4	Threshold Effect in FM Demodulation	376
8.4.1	Threshold Effects in FM Demodulators	376

CHAPTER 7**RANDOM SIGNALS AND NOISE 308**

7.1	A Relative-Frequency Description of Random Processes	308
7.2	Some Terminology of Random Processes	310
7.2.1	Sample Functions and Ensembles	310

viii Contents**8.5 Noise in Pulse-Code Modulation 384**

8.5.1 Postdetection SNR 384

8.5.2 Companding 387

Further Reading 389**Summary 389****Drill Problems 391****Problems 391****Computer Exercises 394****CHAPTER 9****PRINCIPLES OF DIGITAL DATA TRANSMISSION IN NOISE 396****9.1 Baseband Data Transmission in White Gaussian Noise 398****9.2 Binary Synchronous Data Transmission with Arbitrary Signal Shapes 404**

9.2.1 Receiver Structure and Error Probability 404

9.2.2 The Matched Filter 407

9.2.3 Error Probability for the Matched-Filter Receiver 410

9.2.4 Correlator Implementation of the Matched-Filter Receiver 413

9.2.5 Optimum Threshold 414

9.2.6 Nonwhite (Colored) Noise Backgrounds 414

9.2.7 Receiver Implementation Imperfections 415

9.2.8 Error Probabilities for Coherent Binary Signaling 415

9.3 Modulation Schemes not Requiring Coherent References 421

9.3.1 Differential Phase-Shift Keying (DPSK) 422

9.3.2 Differential Encoding and Decoding of Data 427

9.3.3 Noncoherent FSK 429

9.4 *M*-ary Pulse-Amplitude Modulation (PAM) 431**9.5 Comparison of Digital Modulation Systems 435****9.6 Noise Performance of Zero-ISI Digital Data Transmission Systems 438****9.7 Multipath Interference 443****9.8 Fading Channels 449**

9.8.1 Basic Channel Models 449

9.8.2 Flat-Fading Channel Statistics and Error Probabilities 450

9.9 Equalization 455

9.9.1 Equalization by Zero-Forcing 455

9.9.2 Equalization by MMSE 459

9.9.3 Tap Weight Adjustment 463

Further Reading 466**Summary 466****Drill Problems 468****Problems 469****Computer Exercises 476****CHAPTER 10****ADVANCED DATA COMMUNICATIONS TOPICS 477****10.1 *M*-ary Data Communications Systems 477**10.1.1 *M*-ary Schemes Based on Quadrature Multiplexing 477

10.1.2 OQPSK Systems 481

10.1.3 MSK Systems 482

10.1.4 *M*-ary Data Transmission in Terms of Signal Space 489

10.1.5 QPSK in Terms of Signal Space 491

10.1.6 *M*-ary Phase-Shift Keying 493

10.1.7 Quadrature-Amplitude Modulation (QAM) 495

10.1.8 Coherent FSK 497

10.1.9 Noncoherent FSK 498

10.1.10 Differentially Coherent Phase-Shift Keying 502

10.1.11 Bit Error Probability from Symbol Error Probability 503

10.1.12 Comparison of *M*-ary Communications Systems on the Basis of Bit Error Probability 50510.1.13 Comparison of *M*-ary Communications Systems on the Basis of Bandwidth Efficiency 508**10.2 Power Spectra for Digital Modulation 510**

10.2.1 Quadrature Modulation Techniques 510

10.2.2 FSK Modulation 514

10.2.3 Summary 516

10.3 Synchronization 516

10.3.1 Carrier Synchronization 517

10.3.2 Symbol Synchronization 520

10.3.3 Word Synchronization 521

10.3.4 Pseudo-Noise (PN) Sequences 524

10.4 Spread-Spectrum Communication Systems 528

10.4.1 Direct-Sequence Spread Spectrum 530

10.4.2 Performance of DSSS in CW Interference Environments 532

10.4.3 Performance of Spread Spectrum in Multiple User Environments 533

10.4.4 Frequency-Hop Spread Spectrum 536

10.4.5 Code Synchronization 537

10.4.6 Conclusion 539

10.5 Multicarrier Modulation and Orthogonal Frequency-Division Multiplexing 540**10.6 Cellular Radio Communication Systems 545**

10.6.1 Basic Principles of Cellular Radio 546

10.6.2 Channel Perturbations in Cellular Radio 550

10.6.3 Multiple-Input Multiple-Output (MIMO) Systems—Protection Against Fading 551

10.6.4 Characteristics of 1G and 2G Cellular Systems 553

- 10.6.5 Characteristics of cdma2000 and W-CDMA 553
- 10.6.6 Migration to 4G 555

Further Reading 556

Summary 556

Drill Problems 557

Problems 558

Computer Exercises 563

CHAPTER 11

OPTIMUM RECEIVERS AND SIGNAL-SPACE CONCEPTS 564

11.1 Bayes Optimization 564

- 11.1.1 Signal Detection versus Estimation 564
- 11.1.2 Optimization Criteria 565
- 11.1.3 Bayes Detectors 565
- 11.1.4 Performance of Bayes Detectors 569
- 11.1.5 The Neyman-Pearson Detector 572
- 11.1.6 Minimum Probability of Error Detectors 573
- 11.1.7 The Maximum *a Posteriori* (MAP) Detector 573
- 11.1.8 Minimax Detectors 573
- 11.1.9 The *M*-ary Hypothesis Case 573
- 11.1.10 Decisions Based on Vector Observations 574

11.2 Vector Space Representation of Signals 574

- 11.2.1 Structure of Signal Space 575
- 11.2.2 Scalar Product 575
- 11.2.3 Norm 576
- 11.2.4 Schwarz's Inequality 576
- 11.2.5 Scalar Product of Two Signals in Terms of Fourier Coefficients 578
- 11.2.6 Choice of Basis Function Sets—The Gram-Schmidt Procedure 579
- 11.2.7 Signal Dimensionality as a Function of Signal Duration 581

11.3 Map Receiver for Digital Data Transmission 583

- 11.3.1 Decision Criteria for Coherent Systems in Terms of Signal Space 583
- 11.3.2 Sufficient Statistics 589
- 11.3.3 Detection of *M*-ary Orthogonal Signals 590
- 11.3.4 A Noncoherent Case 592

11.4 Estimation Theory 596

- 11.4.1 Bayes Estimation 596
- 11.4.2 Maximum-Likelihood Estimation 598
- 11.4.3 Estimates Based on Multiple Observations 599
- 11.4.4 Other Properties of ML Estimates 601
- 11.4.5 Asymptotic Qualities of ML Estimates 602

11.5 Applications of Estimation Theory to Communications 602

- 11.5.1 Pulse-Amplitude Modulation (PAM) 603

- 11.5.2 Estimation of Signal Phase: The PLL Revisited 604

Further Reading 606

Summary 607

Drill Problems 607

Problems 608

Computer Exercises 614

CHAPTER 12

INFORMATION THEORY AND CODING 615

12.1 Basic Concepts 616

- 12.1.1 Information 616
- 12.1.2 Entropy 617
- 12.1.3 Discrete Channel Models 618
- 12.1.4 Joint and Conditional Entropy 621
- 12.1.5 Channel Capacity 622

12.2 Source Coding 626

- 12.2.1 An Example of Source Coding 627
- 12.2.2 Several Definitions 630
- 12.2.3 Entropy of an Extended Binary Source 631
- 12.2.4 Shannon-Fano Source Coding 632
- 12.2.5 Huffman Source Coding 632

12.3 Communication in Noisy Environments: Basic Ideas 634

12.4 Communication in Noisy Channels: Block Codes 636

- 12.4.1 Hamming Distances and Error Correction 637
- 12.4.2 Single-Parity-Check Codes 638
- 12.4.3 Repetition Codes 639
- 12.4.4 Parity-Check Codes for Single Error Correction 640
- 12.4.5 Hamming Codes 644
- 12.4.6 Cyclic Codes 645
- 12.4.7 The Golay Code 647
- 12.4.8 Bose-Chaudhuri-Hocquenghem (BCH) Codes and Reed Solomon Codes 648
- 12.4.9 Performance Comparison Techniques 648
- 12.4.10 Block Code Examples 650

12.5 Communication in Noisy Channels: Convolutional Codes 657

- 12.5.1 Tree and Trellis Diagrams 659
- 12.5.2 The Viterbi Algorithm 661
- 12.5.3 Performance Comparisons for Convolutional Codes 664

12.6 Bandwidth and Power Efficient Modulation (TCM) 668

12.7 Feedback Channels 672

12.8 Modulation and Bandwidth Efficiency 676

- 12.8.1 Bandwidth and SNR 677
- 12.8.2 Comparison of Modulation Systems 678

x Contents

- 12.9 Quick Overviews 679**
 - 12.9.1 Interleaving and Burst-Error Correction 679
 - 12.9.2 Turbo Coding 681
 - 12.9.3 Source Coding Examples 683
 - 12.9.4 Digital Television 685

Further Reading 686

Summary 686

Drill Problems 688

Problems 688

Computer Exercises 692

APPENDIX A

PHYSICAL NOISE SOURCES 693

A.1 Physical Noise Sources 693

- A.1.1 Thermal Noise 693
- A.1.2 Nyquist's Formula 695
- A.1.3 Shot Noise 695
- A.1.4 Other Noise Sources 696
- A.1.5 Available Power 696
- A.1.6 Frequency Dependence 697
- A.1.7 Quantum Noise 697

A.2 Characterization of Noise in Systems 698

- A.2.1 Noise Figure of a System 699
- A.2.2 Measurement of Noise Figure 700
- A.2.3 Noise Temperature 701
- A.2.4 Effective Noise Temperature 702
- A.2.5 Cascade of Subsystems 702
- A.2.6 Attenuator Noise Temperature and Noise Figure 704

A.3 Free-Space Propagation Example 705

Further Reading 708

Problems 708

APPENDIX B

JOINTLY GAUSSIAN RANDOM VARIABLES 710

B.1 The pdf 710

B.2 The Characteristic Function 711

B.3 Linear Transformations 711

APPENDIX C

PROOF OF THE NARROWBAND NOISE MODEL 712

APPENDIX D

ZERO-CROSSING AND ORIGIN ENCIRCLEMENT STATISTICS 714

D.1 The Zero-Crossing Problem 714

D.2 Average Rate of Zero Crossings 716

Problems 719

APPENDIX E

CHI-SQUARE STATISTICS 720

APPENDIX F

MATHEMATICAL AND NUMERICAL TABLES 722

F.1 The Gaussian Q -Function 722

F.2 Trigonometric Identities 724

F.3 Series Expansions 724

F.4 Integrals 725

F.4.1 Indefinite 725

F.4.2 Definite 726

F.5 Fourier-Transform Pairs 727

F.6 Fourier-Transform Theorems 727

APPENDIX G

ANSWERS TO DRILL PROBLEMS

www.wiley.com/college/ziemer

BIBLIOGRAPHY

www.wiley.com/college/ziemer

INDEX 728

We are said to live in an era called the intangible economy, driven not by the physical flow of material goods but rather by the flow of information. If we are thinking about making a major purchase, for example, chances are we will gather information about the product by an Internet search. Such information gathering is made feasible by virtually instantaneous access to a myriad of facts about the product, thereby making our selection of a particular brand more informed. When one considers the technological developments that make such instantaneous information access possible, two main ingredients surface—a reliable, fast means of communication and a means of storing the information for ready access, sometimes referred to as the *convergence* of communications and computing.

This book is concerned with the theory of systems for the conveyance of information. A *system* is a combination of circuits and/or devices that is assembled to accomplish a desired task, such as the transmission of intelligence from one point to another. Many means for the transmission of information have been used down through the ages ranging from the use of sunlight reflected from mirrors by the Romans to our modern era of electrical communications that began with the invention of the telegraph in the 1800s. It almost goes without saying that we are concerned about the theory of systems for *electrical* communications in this book.

A characteristic of electrical communication systems is the presence of uncertainty. This uncertainty is due in part to the inevitable presence in any system of unwanted signal perturbations, broadly referred to as *noise*, and in part to the unpredictable nature of information itself. Systems analysis in the presence of such uncertainty requires the use of probabilistic techniques.

Noise has been an ever-present problem since the early days of electrical communication, but it was not until the 1940s that probabilistic systems analysis procedures were used to analyze and optimize communication systems operating in its presence [Wiener 1949; Rice 1944, 1945].¹ It is also somewhat surprising that the unpredictable nature of information was not widely recognized until the publication of Claude Shannon's mathematical theory of communications [Shannon 1948] in the late 1940s. This work was the beginning of the science of information theory, a topic that will be considered in some detail later.

Major historical facts related to the development of electrical communications are given in Table 1.1. It provides an appreciation for the accelerating development of communications-related inventions and events down through the years.

¹References in brackets [] refer to Historical References in the Bibliography.

Table 1.1 Major Events and Inventions in the Development of Electrical Communications

Year	Event
1791	Alessandro Volta invents the galvanic cell, or battery
1826	Georg Simon Ohm establishes a law on the voltage-current relationship in resistors
1838	Samuel F. B. Morse demonstrates the telegraph
1864	James C. Maxwell predicts electromagnetic radiation
1876	Alexander Graham Bell patents the telephone
1887	Heinrich Hertz verifies Maxwell's theory
1897	Guglielmo Marconi patents a complete wireless telegraph system
1904	John Fleming patents the thermionic diode
1905	Reginald Fessenden transmits speech signals via radio
1906	Lee De Forest invents the triode amplifier
1915	The Bell System completes a U.S. transcontinental telephone line
1918	B. H. Armstrong perfects the superheterodyne radio receiver
1920	J. R. Carson applies sampling to communications
1925–27	First television broadcasts in England and the United States
1931	Teletypewriter service is initialized
1933	Edwin Armstrong invents frequency modulation
1936	Regular television broadcasting begun by the BBC
1937	Alec Reeves conceives pulse-code modulation (PCM)
WWII	Radar and microwave systems are developed; Statistical methods are applied to signal extraction problems
1944	Computers put into public service (government owned)
1948	The transistor is invented by W. Brattain, J. Bardeen, & W. Shockley
1948	Claude Shannon's "A Mathematical Theory of Communications" is published
1950	Time-division multiplexing is applied to telephony
1956	First successful transoceanic telephone cable
1959	Jack Kilby patents the "Solid Circuit"—precursor to the integrated circuit
1960	First working laser demonstrated by T. H. Maiman of Hughes Research Labs (patent awarded to G. Gould after 20-year dispute with Bell Labs)
1962	First communications satellite, Telstar I, launched
1966	First successful FAX (facsimile) machine
1967	U.S. Supreme Court Carterfone decision opens door for modem development
1968	Live television coverage of the moon exploration
1969	First Internet started—ARPANET
1970	Low-loss optic fiber developed
1971	Microprocessor invented
1975	Ethernet patent filed
1976	Apple I home computer invented
1977	Live telephone traffic carried by fiber-optic cable system
1977	Interplanetary grand tour launched; Jupiter, Saturn, Uranus, and Neptune
1979	First cellular telephone network started in Japan
1981	IBM personal computer developed and sold to public
1981	Hayes Smartmodem marketed (automatic dial-up allowing computer control)
1982	Compact disk (CD) audio based on 16-bit PCM developed
1983	First 16-bit programmable digital signal processors sold
1984	Divestiture of AT&T's local operations into seven Regional Bell Operating Companies

Table 1.1 (Continued)

Year	Event
1985	Desktop publishing programs first sold; Ethernet developed
1988	First commercially available flash memory (later applied in cellular phones, etc.)
1988	ADSL (asymmetric digital subscriber lines) developed
1990s	Very small aperture satellites (VSATs) become popular
1991	Application of echo cancellation results in low-cost 14,400 bits/s modems
1993	Invention of turbo coding allows approach to Shannon limit
mid-1990s	Second-generation (2G) cellular systems fielded
1995	Global Positioning System reaches full operational capability
1996	All-digital phone systems result in modems with 56 kbps download speeds
late-1990s	Widespread personal and commercial applications of the Internet High-definition TV becomes mainstream Apple iPod first sold (October); 100 million sold by April 2007
2001	Fielding of 3G cellular telephone systems begins; WiFi and WiMAX allow wireless access to the Internet and electronic devices wherever mobility is desired
2000s	Wireless sensor networks, originally conceived for military applications, find civilian applications such as environment monitoring, healthcare applications, home automation, and traffic control as well
2010s	Introduction of fourth-generation cellular radio. Technological convergence of communications-related devices—e.g., cell phones, television, personal digital assistants, etc.

It is an interesting fact that the first electrical communication system, the telegraph, was digital—that is, it conveyed information from point to point by means of a digital code consisting of words composed of dots and dashes.² The subsequent invention of the telephone 38 years after the telegraph, wherein voice waves are conveyed by an analog current, swung the pendulum in favor of this more convenient means of word communication for about 75 years.³

One may rightly ask, in view of this history, why the almost complete domination by digital formatting in today's world? There are several reasons, among which are: (1) Media integrity—a digital format suffers much less deterioration in reproduction than does an analog record; (2) Media integration—whether a sound, picture, or naturally digital data such as a word file, all are treated the same when in digital format; (3) Flexible interaction—the digital domain is much more convenient for supporting anything from one-on-one to many-to-many interactions; (4) Editing—whether text, sound, images, or video, all are conveniently and easily edited when in digital format.

With this brief introduction and history, we now look in more detail at the various components that make up a typical communication system.

²In the actual physical telegraph system, a dot was conveyed by a short double-click by closing and opening of the circuit with the telegrapher's key (a switch), while a dash was conveyed by a longer double click by an extended closing of the circuit by means of the telegrapher's key.

³See B. Oliver, J. Pierce, and C. Shannon, "The Philosophy of PCM," *Proc. IRE*, Vol. 16, pp. 1324–1331, November 1948.

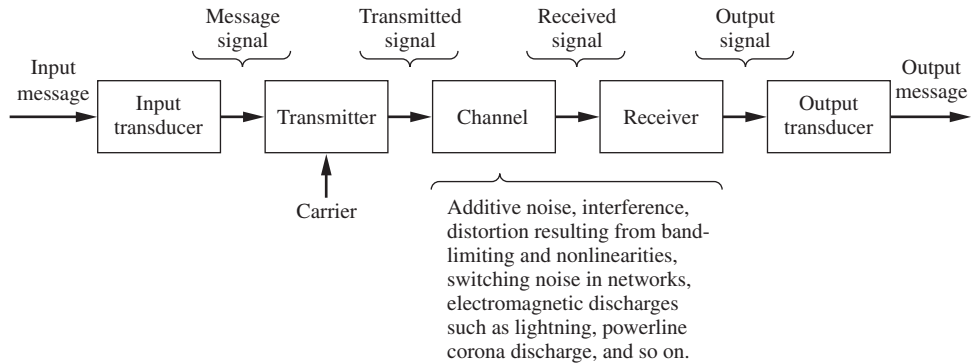


Figure 1.1
The Block Diagram of a Communication System.

1.1 THE BLOCK DIAGRAM OF A COMMUNICATION SYSTEM

Figure 1.1 shows a commonly used model for a single-link communication system.⁴ Although it suggests a system for communication between two remotely located points, this block diagram is also applicable to remote sensing systems, such as radar or sonar, in which the system input and output may be located at the same site. Regardless of the particular application and configuration, all information transmission systems invariably involve three major subsystems—a transmitter, the channel, and a receiver. In this book we will usually be thinking in terms of systems for transfer of information between remotely located points. It is emphasized, however, that the techniques of systems analysis developed are not limited to such systems.

We will now discuss in more detail each functional element shown in Figure 1.1.

Input Transducer The wide variety of possible sources of information results in many different forms for messages. Regardless of their exact form, however, messages may be categorized as *analog* or *digital*. The former may be modeled as functions of a continuous-time variable (for example, pressure, temperature, speech, music), whereas the latter consist of discrete symbols (for example, written text or a sampled/quantized analog signal such as speech). Almost invariably, the message produced by a source must be converted by a transducer to a form suitable for the particular type of communication system employed. For example, in electrical communications, speech waves are converted by a microphone to voltage variations. Such a converted message is referred to as the *message signal*. In this book, therefore, a *signal* can be interpreted as the variation of a quantity, often a voltage or current, with time.

⁴More complex communications systems are the rule rather than the exception: a broadcast system, such as television or commercial radio, is a one-to-many type of situation composed of several sinks receiving the same information from a single source; a multiple-access communication system is where many users share the same channel and is typified by satellite communications systems; a many-to-many type of communications scenario is the most complex and is illustrated by examples such as the telephone system and the Internet, both of which allow communication between any pair out of a multitude of users. For the most part, we consider only the simplest situation in this book of a single sender to a single receiver, although means for sharing a communication resource will be dealt with under the topics of multiplexing and multiple access.

Transmitter The purpose of the transmitter is to couple the message to the channel. Although it is not uncommon to find the input transducer directly coupled to the transmission medium, as for example in some intercom systems, it is often necessary to *modulate* a carrier wave with the signal from the input transducer. *Modulation* is the systematic variation of some attribute of the carrier, such as amplitude, phase, or frequency, in accordance with a function of the message signal. There are several reasons for using a carrier and modulating it. Important ones are (1) for ease of radiation, (2) to reduce noise and interference, (3) for channel assignment, (4) for multiplexing or transmission of several messages over a single channel, and (5) to overcome equipment limitations. Several of these reasons are self-explanatory; others, such as the second, will become more meaningful later.

In addition to modulation, other primary functions performed by the transmitter are filtering, amplification, and coupling the modulated signal to the channel (for example, through an antenna or other appropriate device).

Channel The channel can have many different forms; the most familiar, perhaps, is the channel that exists between the transmitting antenna of a commercial radio station and the receiving antenna of a radio. In this channel, the transmitted signal propagates through the atmosphere, or free space, to the receiving antenna. However, it is not uncommon to find the transmitter hard-wired to the receiver, as in most local telephone systems. This channel is vastly different from the radio example. However, all channels have one thing in common: the signal undergoes degradation from transmitter to receiver. Although this degradation may occur at any point of the communication system block diagram, it is customarily associated with the channel alone. This degradation often results from noise and other undesired signals or interference but also may include other distortion effects as well, such as fading signal levels, multiple transmission paths, and filtering. More about these unwanted perturbations will be presented shortly.

Receiver The receiver's function is to extract the desired message from the received signal at the channel output and to convert it to a form suitable for the output transducer. Although amplification may be one of the first operations performed by the receiver, especially in radio communications, where the received signal may be extremely weak, the main function of the receiver is to *demodulate* the received signal. Often it is desired that the receiver output be a scaled, possibly delayed, version of the message signal at the modulator input, although in some cases a more general function of the input message is desired. However, as a result of the presence of noise and distortion, this operation is less than ideal. Ways of approaching the ideal case of perfect recovery will be discussed as we proceed.

Output Transducer The output transducer completes the communication system. This device converts the electric signal at its input into the form desired by the system user. Perhaps the most common output transducer is a loudspeaker or ear phone.

■ 1.2 CHANNEL CHARACTERISTICS

1.2.1 Noise Sources

Noise in a communication system can be classified into two broad categories, depending on its source. Noise generated by components within a communication system, such as resistors and

solid-state active devices is referred to as *internal noise*. The second category, *external noise*, results from sources outside a communication system, including atmospheric, man-made, and extraterrestrial sources.

Atmospheric noise results primarily from spurious radio waves generated by the natural electrical discharges within the atmosphere associated with thunderstorms. It is commonly referred to as *static* or *spherics*. Below about 100 MHz, the field strength of such radio waves is inversely proportional to frequency. Atmospheric noise is characterized in the time domain by large-amplitude, short-duration bursts and is one of the prime examples of noise referred to as *impulsive*. Because of its inverse dependence on frequency, atmospheric noise affects commercial AM broadcast radio, which occupies the frequency range from 540 kHz to 1.6 MHz, more than it affects television and FM radio, which operate in frequency bands above 50 MHz.

Man-made noise sources include high-voltage powerline corona discharge, commutator-generated noise in electrical motors, automobile and aircraft ignition noise, and switching-gear noise. Ignition noise and switching noise, like atmospheric noise, are impulsive in character. Impulse noise is the predominant type of noise in switched wireline channels, such as telephone channels. For applications such as voice transmission, impulse noise is only an irritation factor; however, it can be a serious source of error in applications involving transmission of digital data.

Yet another important source of man-made noise is radio-frequency transmitters other than the one of interest. Noise due to interfering transmitters is commonly referred to as *radio-frequency interference* (RFI). RFI is particularly troublesome in situations in which a receiving antenna is subject to a high-density transmitter environment, as in mobile communications in a large city.

Extraterrestrial noise sources include our sun and other hot heavenly bodies, such as stars. Owing to its high temperature (6000°C) and relatively close proximity to the earth, the sun is an intense, but fortunately localized source of radio energy that extends over a broad frequency spectrum. Similarly, the stars are sources of wideband radio energy. Although much more distant and hence less intense than the sun, nevertheless they are collectively an important source of noise because of their vast numbers. Radio stars such as quasars and pulsars are also intense sources of radio energy. Considered a signal source by radio astronomers, such stars are viewed as another noise source by communications engineers. The frequency range of solar and cosmic noise extends from a few megahertz to a few gigahertz.

Another source of interference in communication systems is multiple transmission paths. These can result from reflection off buildings, the earth, airplanes, and ships or from refraction by stratifications in the transmission medium. If the scattering mechanism results in numerous reflected components, the received multipath signal is noiselike and is termed *diffuse*. If the multipath signal component is composed of only one or two strong reflected rays, it is termed *specular*. Finally, signal degradation in a communication system can occur because of random changes in attenuation within the transmission medium. Such signal perturbations are referred to as *fading*, although it should be noted that specular multipath also results in fading due to the constructive and destructive interference of the received multiple signals.

Internal noise results from the random motion of charge carriers in electronic components. It can be of three general types: the first is referred to as *thermal noise*, which is caused by the random motion of free electrons in a conductor or semiconductor excited by thermal agitation; the second is called *shot noise* and is caused by the random arrival of discrete charge carriers in such devices as thermionic tubes or semiconductor junction devices; the third, known as *flicker noise*, is produced in semiconductors by a mechanism not well understood and is more

severe the lower the frequency. The first type of noise source, *thermal noise*, is modeled analytically in Appendix A, and examples of system characterization using this model are given there.

1.2.2 Types of Transmission Channels

There are many types of transmission channels. We will discuss the characteristics, advantages, and disadvantages of three common types: electromagnetic-wave propagation channels, guided electromagnetic-wave channels, and optical channels. The characteristics of all three may be explained on the basis of electromagnetic-wave propagation phenomena. However, the characteristics and applications of each are different enough to warrant our considering them separately.

Electromagnetic-Wave Propagation Channels

The possibility of the propagation of electromagnetic waves was predicted in 1864 by James Clerk Maxwell (1831–1879), a Scottish mathematician who based his theory on the experimental work of Michael Faraday. Heinrich Hertz (1857–1894), a German physicist, carried out experiments between 1886 and 1888 using a rapidly oscillating spark to produce electromagnetic waves, thereby experimentally proving Maxwell's predictions. Therefore, by the latter part of the nineteenth century, the physical basis for many modern inventions utilizing electromagnetic-wave propagation—such as radio, television, and radar—was already established.

The basic physical principle involved is the coupling of electromagnetic energy into a propagation medium, which can be free space or the atmosphere, by means of a radiation element referred to as an *antenna*. Many different propagation modes are possible, depending on the physical configuration of the antenna and the characteristics of the propagation medium. The simplest case—which never occurs in practice—is propagation from a point source in a medium that is infinite in extent. The propagating wave fronts (surfaces of constant phase) in this case would be concentric spheres. Such a model might be used for the propagation of electromagnetic energy from a distant spacecraft to earth. Another idealized model, which approximates the propagation of radio waves from a commercial broadcast antenna, is that of a conducting line perpendicular to an infinite conducting plane. These and other idealized cases have been analyzed in books on electromagnetic theory. Our purpose is not to summarize all the idealized models, but to point out basic aspects of propagation phenomena in practical channels.

Except for the case of propagation between two spacecraft in outer space, the intermediate medium between transmitter and receiver is never well approximated by free space. Depending on the distance involved and the frequency of the radiated waveform, a terrestrial communication link may depend on line-of-sight, ground-wave, or ionospheric skip-wave propagation (see Figure 1.2). Table 1.2 lists frequency bands from 3 kHz to 10^7 GHz, along with letter designations for microwave bands used in radar among other applications. Note that the frequency bands are given in decades; the VHF band has 10 times as much frequency space as the HF band. Table 1.3 shows some bands of particular interest.

General application allocations are arrived at by international agreement. The present system of frequency allocations is administered by the International Telecommunications Union (ITU), which is responsible for the periodic convening of Administrative Radio Conferences

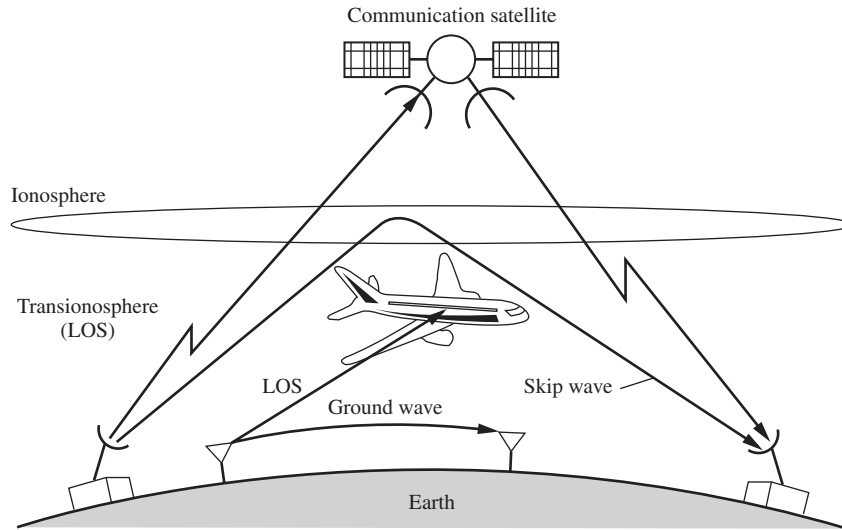


Figure 1.2

The various propagation modes for electromagnetic waves (LOS stands for line of sight).

Table 1.2 Frequency Bands with Designations

Frequency band	Name	Microwave band (GHz)	Letter designation	
3–30 kHz	Very low frequency (VLF)			
30–300 kHz	Low frequency (LF)			
300–3000 kHz	Medium frequency (MF)			
3–30 MHz	High frequency (HF)			
30–300 MHz	Very high frequency (VHF)			
0.3–3 GHz	Ultrahigh frequency (UHF)	1.0–2.0	L	
		2.0–3.0	S	
		3.0–4.0	S	
		4.0–6.0	C	
		6.0–8.0	C	
		8.0–10.0	X	
		10.0–12.4	X	
3–30 GHz	Superhigh frequency (SHF)	12.4–18.0	Ku	
		18.0–20.0	K	
		20.0–26.5	K	
		26.5–40.0	Ka	
		30–300 GHz	Extremely high frequency (EHF)	
		43–430 THz	Infrared (0.7–7 μm)	
		430–750 THz	Visible light (0.4–0.7 μm)	
750–3000 THz	Ultraviolet (0.1–0.4 μm)			

Note: kHz = kilohertz = $\times 10^3$; MHz = megahertz = $\times 10^6$; GHz = gigahertz = $\times 10^9$; THz = terahertz = $\times 10^{12}$; μm = micrometers = $\times 10^{-6}$ meters.

Table 1.3 Selected Frequency Bands for Public Use and Military Communications⁵

Use		Frequency
Radio navigation		6–14 kHz; 90–110 kHz
Loran C navigation		100 kHz
Standard (AM) broadcast		540–1600 kHz
ISM band	Industrial heaters; welders	40.66–40.7 MHz
Television:	Channels 2–4	54–72 MHz
	Channels 5–6	76–88 MHz
FM broadcast		88–108 MHz
Television	Channels 7–13	174–216 MHz
	Channels 14–83	420–890 MHz
	(In the United States, channels 2–36 and 38–51 are used for digital TV broadcast; others were reallocated.)	
Cellular mobile radio	AMPS, D-AMPS (1G, 2G)	800 MHz bands
	IS-95 (2G)	824–844 MHz/1.8–2 GHz
	GSM (2G)	850/900/1800/1900 MHz
	3G (UMTS, cdma-2000)	1.8/2.5 GHz bands
Wi-Fi (IEEE 802.11)		2.4/5 GHz
Wi-MAX (IEEE 802.16)		2–11 GHz
ISM band	Microwave ovens; medical	902–928 MHz
Global Positioning System		1227.6, 1575.4 MHz
Point-to-point microwave		2.11–2.13 GHz
Point-to-point microwave	Interconnecting base stations	2.16–2.18 GHz
ISM band	Microwave ovens; unlicensed spread spectrum; medical	2.4–2.4835 GHz
		23.6–24 GHz
		122–123 GHz
		244–246 GHz

on a regional or a worldwide basis (WARC before 1995; WRC 1995 and after, standing for World Radiocommunication Conference).⁶ The responsibility of the WRCs is the drafting, revision, and adoption of the *Radio Regulations*, which is an instrument for the international management of the radio spectrum.⁷

In the United States, the Federal Communications Commission (FCC) awards specific applications within a band as well as licenses for their use. The FCC is directed by five commissioners appointed to five-year terms by the President and confirmed by the Senate. One commissioner is appointed as chairperson by the President.⁸

At lower frequencies, or long wavelengths, propagating radio waves tend to follow the earth's surface. At higher frequencies, or short wavelengths, radio waves propagate in straight

⁵Bennet Z. Kobb, *Spectrum Guide*, 3rd ed., Falls Church, VA: New Signals Press, 1996. Bennet Z. Kobb, *Wireless Spectrum Finder*, New York: McGraw Hill, 2001.

⁶WARC-79, WARC-84, and WARC-92, all held in Geneva, Switzerland, were the last three held under the WARC designation; WRC-95, WRC-97, WRC-00, WRC-03, WRC-07, and WRC-12 are those held under the WRC designation. The next one to be held is WRC-15 and includes four informal working groups: Maritime, Aeronautical and Radar Services; Terrestrial Services; Space Services; and Regulatory Issues.

⁷Available on the Radio Regulations website: <http://www.itu.int/pub/R-REG-RR-2004/en>

⁸<http://www.fcc.gov/>

lines. Another phenomenon that occurs at lower frequencies is reflection (or refraction) of radio waves by the ionosphere (a series of layers of charged particles at altitudes between 30 and 250 miles above the earth's surface). Thus, for frequencies below about 100 MHz, it is possible to have skip-wave propagation. At night, when lower ionospheric layers disappear due to less ionization from the sun (the E , F_1 , and F_2 layers coalesce into one layer—the F layer), longer skip-wave propagation occurs as a result of reflection from the higher, single reflecting layer of the ionosphere.

Above about 300 MHz, propagation of radio waves is by line of sight, because the ionosphere will not bend radio waves in this frequency region sufficiently to reflect them back to the earth. At still higher frequencies, say above 1 or 2 GHz, atmospheric gases (mainly oxygen), water vapor, and precipitation absorb and scatter radio waves. This phenomenon manifests itself as attenuation of the received signal, with the attenuation generally being more severe the higher the frequency (there are resonance regions for absorption by gases that peak at certain frequencies). Figure 1.3 shows specific attenuation curves as a function of frequency⁹ for oxygen, water vapor, and rain [recall that 1 decibel (dB) is ten times the logarithm to the base 10 of a power ratio]. One must account for the possible attenuation by such atmospheric constituents in the design of microwave links, which are used, for example, in transcontinental telephone links and ground-to-satellite communications links.

At about 23 GHz, the first absorption resonance due to water vapor occurs, and at about 62 GHz a second one occurs due to oxygen absorption. These frequencies should be avoided in transmission of desired signals through the atmosphere, or undue power will be expended (one might, for example, use 62 GHz as a signal for cross-linking between two satellites, where atmospheric absorption is no problem, and thereby prevent an enemy on the ground from listening in). Another absorption frequency for oxygen occurs at 120 GHz, and two other absorption frequencies for water vapor occur at 180 and 350 GHz.

Communication at millimeter-wave frequencies (that is, at 30 GHz and higher) is becoming more important now that there is so much congestion at lower frequencies (the Advanced Technology Satellite, launched in the mid-1990s, employs an uplink frequency band around 20 GHz and a downlink frequency band at about 30 GHz). Communication at millimeter-wave frequencies is becoming more feasible because of technological advances in components and systems. Two bands at 30 and 60 GHz, the LMDS (Local Multipoint Distribution System) and MMDS (Multichannel Multipoint Distribution System) bands, have been identified for terrestrial transmission of wideband signals. Great care must be taken to design systems using these bands because of the high atmospheric and rain absorption as well as blockage by objects such as trees and buildings. To a great extent, use of these bands has been obsoleted by more recent standards such as WiMAX (Worldwide Interoperability for Microwave Access), sometimes referred to as Wi-Fi on steroids.¹⁰

Somewhere above 1 THz (1000 GHz), the propagation of radio waves becomes optical in character. At a wavelength of 10 μm (0.00001 m), the carbon dioxide laser provides a source of coherent radiation, and visible-light lasers (for example, helium-neon) radiate in the wavelength region of 1 μm and shorter. Terrestrial communications systems employing such frequencies experience considerable attenuation on cloudy days, and laser communications over terrestrial links are restricted to optical fibers for the most part. Analyses have been carried out for the employment of laser communications cross-links between satellites.

⁹Data from Louis J. Ippolito, Jr., *Radiowave Propagation in Satellite Communications*, New York: Van Nostrand Reinhold, 1986, Chapters 3 and 4.

¹⁰See Wikipedia under LMDS, MMDS, WiMAX, or Wi-Fi for more information on these terms.

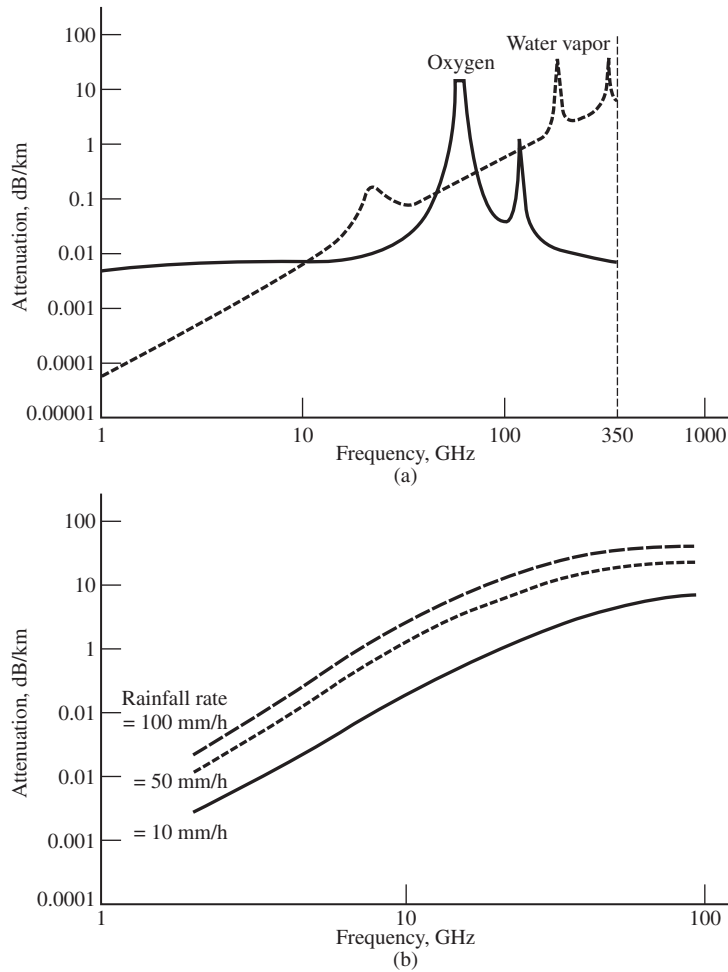


Figure 1.3

Specific attenuation for atmospheric gases and rain. (a) Specific attenuation due to oxygen and water vapor (concentration of 7.5 g/m^3). (b) Specific attenuation due to rainfall at rates of 10, 50, and 100 mm/h.

Guided Electromagnetic-Wave Channels

Up until the last part of the twentieth century, the most extensive example of guided electromagnetic-wave channels is the part of the long-distance telephone network that uses wire lines, but this has almost exclusively been replaced by optical fiber.¹¹ Communication between persons a continent apart was first achieved by means of voice frequency transmission (below 10,000 Hz) over open wire. Quality of transmission was rather poor. By 1952, use of the types of modulation known as double-sideband and single-sideband on high-frequency carriers was established. Communication over predominantly multipair and coaxial-cable lines

¹¹For a summary of guided transmission systems as applied to telephone systems, see F. T. Andrews, Jr., "Communications Technology: 25 Years in Retrospect. Part III, Guided Transmission Systems: 1952–1973." *IEEE Communications Society Magazine*, Vol. 16, pp. 4–10, January 1978.

produced transmission of much better quality. With the completion of the first trans-Atlantic cable in 1956, intercontinental telephone communication was no longer dependent on high-frequency radio, and the quality of intercontinental telephone service improved significantly.

Bandwidths on coaxial-cable links are a few megahertz. The need for greater bandwidth initiated the development of millimeter-wave waveguide transmission systems. However, with the development of low-loss optical fibers, efforts to improve millimeter-wave systems to achieve greater bandwidth ceased. The development of optical fibers, in fact, has made the concept of a wired city—wherein digital data and video can be piped to any residence or business within a city—nearly a reality.¹² Modern coaxial-cable systems can carry only 13,000 voice channels per cable, but optical links are capable of carrying several times this number (the limiting factor being the current driver for the light source).¹³

Optical Links The use of optical links was, until recently, limited to short and intermediate distances. With the installation of trans-Pacific and trans-Atlantic optical cables in 1988 and early 1989, this is no longer true.¹⁴ The technological breakthroughs that preceded the widespread use of light waves for communication were the development of small coherent light sources (semiconductor lasers), low-loss optical fibers or waveguides, and low-noise detectors.¹⁵

A typical fiber-optic communication system has a light source, which may be either a light-emitting diode or a semiconductor laser, in which the intensity of the light is varied by the message source. The output of this modulator is the input to a light-conducting fiber. The receiver, or light sensor, typically consists of a photodiode. In a photodiode, an average current flows that is proportional to the optical power of the incident light. However, the exact number of charge carriers (that is, electrons) is random. The output of the detector is the sum of the average current that is proportional to the modulation and a noise component. This noise component differs from the thermal noise generated by the receiver electronics in that it is “bursty” in character. It is referred to as shot noise, in analogy to the noise made by shot hitting a metal plate. Another source of degradation is the dispersion of the optical fiber

¹²The limiting factor here is expense—stringing anything under city streets is a very expensive proposition although there are many potential customers to bear the expense. Providing access to the home in the country is relatively easy from the standpoint of stringing cables or optical fiber, but the number of potential users is small so that the cost per customer goes up. As for cable versus fiber, the “last mile” is in favor of cable again because of expense. Many solutions have been proposed for this “last-mile problem” as it is sometimes referred to, including special modulation schemes to give higher data rates over telephone lines (see ADSL in Table 1.1), making cable TV access two-way (plenty of bandwidth but attenuation a problem), satellite (in remote locations), optical fiber (for those who want wideband and are willing/able to pay for it), and wireless or radio access (see the earlier reference to Wi-MAX). A universal solution for all situations is most likely not possible. For more on this intriguing topic, see Wikipedia.

¹³Wavelength division multiplexing (WDM) is the latest development in the relatively short existence of optical fiber delivery of information. The idea here is that different wavelength bands (“colors”), provided by different laser light sources, are sent in parallel through an optical fiber to vastly increase the bandwidth—several gigahertz of bandwidth is possible. See, for example, *The IEEE Communications Magazine*, February 1999 (issue on “Optical Networks, Communication Systems, and Devices”), October 1999 (issue on “Broadband Technologies and Trials”), February 2000 (issue on “Optical Networks Come of Age”), and June 2000 (“Intelligent Networks for the New Millennium”).

¹⁴See Wikipedia, “Fiber-optic communications.”

¹⁵For an overview on the use of signal-processing methods to improve optical communications, see J. H. Winters, R. D. Gitlin, and S. Kasturia, “Reducing the Effects of Transmission Impairments in Digital Fiber Optic Systems,” *IEEE Communications Magazine*, Vol. 31, pp. 68–76, June 1993.

itself. For example, pulse-type signals sent into the fiber are observed as “smeared out” at the receiver. Losses also occur as a result of the connections between cable pieces and between cable and system components.

Finally, it should be mentioned that optical communications can take place through free space.¹⁶

■ 1.3 SUMMARY OF SYSTEMS-ANALYSIS TECHNIQUES

Having identified and discussed the main subsystems in a communication system and certain characteristics of transmission media, let us now look at the techniques at our disposal for systems analysis and design.

1.3.1 Time and Frequency-Domain Analyses

From circuits courses or prior courses in linear systems analysis, you are well aware that the electrical engineer lives in the two worlds, so to speak, of time and frequency. Also, you should recall that dual time-frequency analysis techniques are especially valuable for linear systems for which the principle of superposition holds. Although many of the subsystems and operations encountered in communication systems are for the most part linear, many are not. Nevertheless, frequency-domain analysis is an extremely valuable tool to the communications engineer, more so perhaps than to other systems analysts. Since the communications engineer is concerned primarily with signal bandwidths and signal locations in the frequency domain, rather than with transient analysis, the essentially steady-state approach of the Fourier series and transforms is used. Accordingly, we provide an overview of the Fourier series, the Fourier integral, and their role in systems analysis in Chapter 2.

1.3.2 Modulation and Communication Theories

Modulation theory employs time and frequency-domain analyses to analyze and design systems for modulation and demodulation of information-bearing signals. To be specific consider the message signal $m(t)$, which is to be transmitted through a channel using the method of double-sideband modulation. The modulated carrier for double-sideband modulation is of the form $x_c(t) = A_c m(t) \cos \omega_c t$, where ω_c is the carrier frequency in radians per second and A_c is the carrier amplitude. Not only must a modulator be built that can multiply two signals, but amplifiers are required to provide the proper power level of the transmitted signal. The exact design of such amplifiers is not of concern in a systems approach. However, the frequency content of the modulated carrier, for example, is important to their design and therefore must be specified. The dual time-frequency analysis approach is especially helpful in providing such information.

At the other end of the channel, there must be a receiver configuration capable of extracting a replica of $m(t)$ from the modulated signal, and one can again apply time and frequency-domain techniques to good effect.

The analysis of the effect of interfering signals on system performance and the subsequent modifications in design to improve performance in the face of such interfering signals are part of *communication theory*, which, in turn, makes use of modulation theory.

¹⁶See *IEEE Communications Magazine*, Vol. 38, pp. 124–139, August 2000 (section on free space laser communications).

This discussion, although mentioning interfering signals, has not explicitly emphasized the uncertainty aspect of the information-transfer problem. Indeed, much can be done without applying probabilistic methods. However, as pointed out previously, the application of probabilistic methods, coupled with optimization procedures, has been one of the key ingredients of the modern communications era and led to the development during the latter half of the twentieth century of new techniques and systems totally different in concept from those that existed before World War II.

We will now survey several approaches to statistical optimization of communication systems.

■ 1.4 PROBABILISTIC APPROACHES TO SYSTEM OPTIMIZATION

The works of Wiener and Shannon, previously cited, were the beginning of modern statistical communication theory. Both these investigators applied probabilistic methods to the problem of extracting information-bearing signals from noisy backgrounds, but they worked from different standpoints. In this section we briefly examine these two approaches to optimum system design.

1.4.1 Statistical Signal Detection and Estimation Theory

Wiener considered the problem of optimally filtering signals from noise, where “optimum” is used in the sense of minimizing the average squared error between the desired and the actual output. The resulting filter structure is referred to as the *Wiener filter*. This type of approach is most appropriate for analog communication systems in which the demodulated output of the receiver is to be a faithful replica of the message input to the transmitter.

Wiener’s approach is reasonable for analog communications. However, in the early 1940s, [North 1943] provided a more fruitful approach to the digital communications problem, in which the receiver must distinguish between a number of discrete signals in background noise. Actually, North was concerned with radar, which requires only the detection of the presence or absence of a pulse. Since fidelity of the detected signal at the receiver is of no consequence in such signal-detection problems, North sought the filter that would maximize the peak-signal-to-root-mean-square (rms)-noise ratio at its output. The resulting optimum filter is called the *matched filter*, for reasons that will become apparent in Chapter 9, where we consider digital data transmission. Later adaptations of the Wiener and matched-filter ideas to time-varying backgrounds resulted in *adaptive filters*. We will consider a subclass of such filters in Chapter 9 when *equalization* of digital data signals is discussed.

The signal-extraction approaches of Wiener and North, formalized in the language of statistics in the early 1950s by several researchers (see [Middleton 1960], p. 832, for several references), were the beginnings of what is today called *statistical signal detection and estimation theory*. In considering the design of receivers utilizing *all* the information available at the channel output, [Woodward and Davies 1952 and Woodward, 1953] determined that this so-called ideal receiver computes the probabilities of the received waveform given the possible transmitted messages. These computed probabilities are known as *a posteriori* probabilities. The ideal receiver then makes the decision that the transmitted message was the one corresponding to the largest *a posteriori* probability. Although perhaps somewhat vague at

this point, this *maximum a posteriori* (MAP) principle, as it is called, is one of the cornerstones of detection and estimation theory. Another development that had far-reaching consequences in the development of detection theory was the application of generalized vector space ideas ([Kotelnikov 1959] and [Wozencraft and Jacobs 1965]). We will examine these ideas in more detail in Chapters 9 through 11.

1.4.2 Information Theory and Coding

The basic problem that Shannon considered is, “Given a message source, how shall the messages produced be represented so as to maximize the information conveyed through a given channel?” Although Shannon formulated his theory for both discrete and analog sources, we will think here in terms of discrete systems. Clearly, a basic consideration in this theory is a measure of information. Once a suitable measure has been defined (and we will do so in Chapter 12), the next step is to define the information carrying capacity, or simply capacity, of a channel as the maximum rate at which information can be conveyed through it. The obvious question that now arises is, “Given a channel, how closely can we approach the capacity of the channel, and what is the quality of the received message?” A most surprising, and the singularly most important, result of Shannon’s theory is that by suitably restructuring the transmitted signal, we can transmit information through a channel *at any rate less than the channel capacity with arbitrarily small error*, despite the presence of noise, provided we have an arbitrarily long time available for transmission. This is the gist of Shannon’s *second theorem*. Limiting our discussion at this point to binary discrete sources, a proof of Shannon’s second theorem proceeds by selecting codewords at random from the set of 2^n possible binary sequences n digits long at the channel input. The probability of error in receiving a given n -digit sequence, when averaged over all possible code selections, becomes arbitrarily small as n becomes arbitrarily large. Thus, many suitable codes exist, *but we are not told how to find these codes*. Indeed, this has been the dilemma of information theory since its inception and is an area of active research. In recent years, great strides have been made in finding good coding and decoding techniques that are implementable with a reasonable amount of hardware and require only a reasonable amount of time to decode.

Several basic coding techniques will be discussed in Chapter 12.¹⁷ Perhaps the most astounding development in the recent history of coding was the invention of turbo coding and subsequent publication by French researchers in 1993.¹⁸ Their results, which were subsequently verified by several researchers, showed performance to within a fraction of a decibel of the Shannon limit.¹⁹

¹⁷For a good survey on “Shannon Theory” as it is known, see S. Verdú, “Fifty Years of Shannon Theory,” *IEEE Trans. on Infor. Theory*, Vol. 44, pp. 2057–2078, October 1998.

¹⁸C. Berrou, A. Glavieux, and P. Thitimajshima, “Near Shannon Limit Error-Correcting Coding and Decoding: Turbo Codes,” *Proc. 1993 Int. Conf. Commun.*, pp. 1064–1070, Geneva, Switzerland, May 1993.

See also D. J. Costello and G. D. Forney, “Channel Coding: The Road to Channel Capacity,” *Proc. IEEE*, Vol. 95, pp. 1150–1177, June 2007, for an excellent tutorial article on the history of coding theory.

¹⁹Actually low-density parity-check codes, invented and published by Robert Gallager in 1963, were the first codes to allow data transmission rates close to the theoretical limit ([Gallager, 1963]). However, they were impractical to implement in 1963, so were forgotten about until the past 10–20 years whence practical advances in their theory and substantially advanced processors have spurred a resurgence of interest in them.

1.4.3 Recent Advances

There have been great strides made in communications theory and its practical implementation in the past few decades. Some of these will be pointed out later in the book. To capture the gist of these advances at this point would delay the coverage of basic concepts of communications theory, which is the underlying intent of this book. For those wanting additional reading at this point, two recent issues of the *IEEE Proceedings* will provide information in two areas, turbo-information processing (used in decoding turbo codes among other applications)²⁰, and multiple-input multiple-output (MIMO) communications theory, which is expected to have far-reaching impact on wireless local- and wide-area network development.²¹ An appreciation for the broad sweep of developments from the beginnings of modern communications theory to recent times can be gained from a collection of papers put together in a single volume, spanning roughly 50 years, that were judged to be worthy of note by experts in the field.²²

1.5 PREVIEW OF THIS BOOK

From the previous discussion, the importance of probability and noise characterization in analysis of communication systems should be apparent. Accordingly, after presenting basic signal, system, noiseless modulation theory, and basic elements of digital data transmission in Chapters 2, 3, 4, and 5, we briefly discuss probability and noise theory in Chapters 6 and 7. Following this, we apply these tools to the noise analysis of analog communications schemes in Chapter 8. In Chapters 9 and 10, we use probabilistic techniques to find optimum receivers when we consider digital data transmission. Various types of digital modulation schemes are analyzed in terms of error probability. In Chapter 11, we approach optimum signal detection and estimation techniques on a generalized basis and use signal-space techniques to provide insight as to why systems that have been analyzed previously perform as they do. As already mentioned, information theory and coding are the subjects of Chapter 12. This provides us with a means of comparing actual communication systems with the ideal. Such comparisons are then considered in Chapter 12 to provide a basis for selection of systems.

In closing, we must note that large areas of communications technology such as optical, computer, and satellite communications are not touched on in this book. However, one can apply the principles developed in this text in those areas as well.

Further Reading

The references for this chapter were chosen to indicate the historical development of modern communications theory and by and large are not easy reading. They are found in the Historical References section of the Bibliography. You also may consult the introductory chapters of the books listed in the Further Reading sections of Chapters 2, 3, and 4. These books appear in the main portion of the Bibliography.

²⁰*Proceedings of the IEEE*, Vol. 95, no. 6, June 2007. Special issue on turbo-information processing.

²¹*Proceedings of the IEEE*, Vol. 95, no. 7, July 2007. Special issue on multi-user MIMO-OFDM for next-generation wireless.

²²W. H. Tranter, D. P. Taylor, R. E. Ziemer, N. F. Maxemchuk, and J. W. Mark (eds.). *The Best of the Best: Fifty Years of Communications and Networking Research*, John Wiley and IEEE Press, January 2007.

SIGNAL AND LINEAR SYSTEM ANALYSIS

The study of information transmission systems is inherently concerned with the transmission of signals through systems. Recall that in Chapter 1 a *signal* was defined as the time history of some quantity, usually a voltage or current. A *system* is a combination of devices and networks (subsystems) chosen to perform a desired function. Because of the sophistication of modern communication systems, a great deal of analysis and experimentation with trial subsystems occurs before actual building of the desired system. Thus, the communications engineer's tools are mathematical models for signals and systems.

In this chapter, we review techniques useful for modeling and analysis of signals and systems used in communications engineering.¹ Of primary concern will be the dual time-frequency viewpoint for signal representation, and models for linear, time-invariant, two-port systems. It is important to always keep in mind that a model is not the signal or the system, but a mathematical idealization of certain characteristics of it that are most relevant to the problem at hand.

With this brief introduction, we now consider signal classifications and various methods for modeling signals and systems. These include frequency-domain representations for signals via the complex exponential Fourier series and the Fourier transform, followed by linear system models and techniques for analyzing the effects of such systems on signals.

2.1 SIGNAL MODELS

2.1.1 Deterministic and Random Signals

In this book we are concerned with two broad classes of signals, referred to as deterministic and random. *Deterministic signals* can be modeled as completely specified functions of time. For example, the signal

$$x(t) = A \cos(\omega_0 t), \quad -\infty < t < \infty \quad (2.1)$$

where A and ω_0 are constants, is a familiar example of a deterministic signal. Another example of a deterministic signal is the unit rectangular pulse, denoted as $\Pi(t)$ and

¹More complete treatments of these subjects can be found in texts on linear system theory. See the references for this chapter for suggestions.

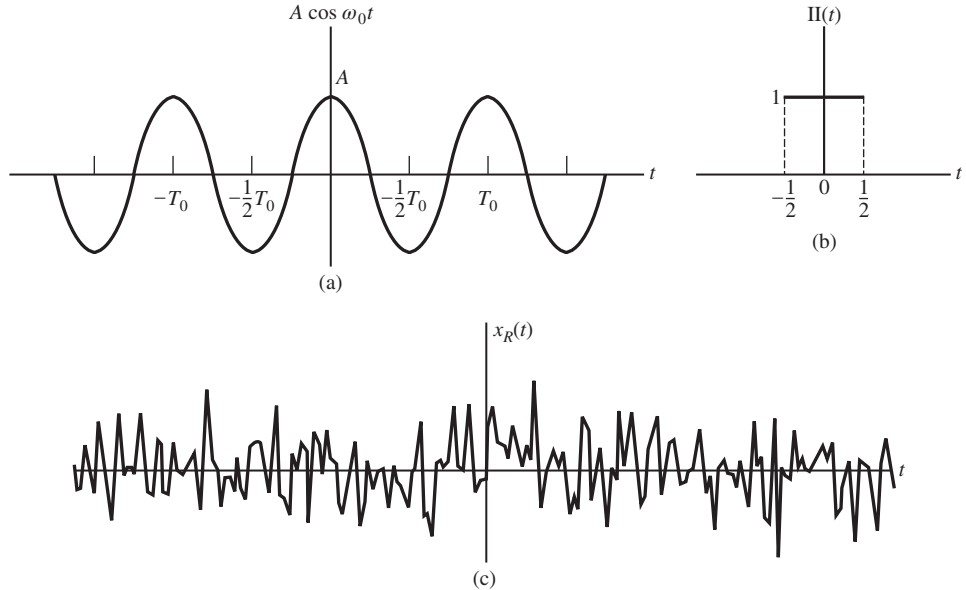


Figure 2.1
Examples of various types of signals. (a) Deterministic (sinusoidal) signal. (b) Unit rectangular pulse signal. (c) Random signal.

defined as

$$\Pi(t) = \begin{cases} 1, & |t| \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases} \quad (2.2)$$

Random signals are signals that take on random values at any given time instant and must be modeled probabilistically. They will be considered in Chapters 6 and 7. Figure 2.1 illustrates the various types of signals just discussed.

2.1.2 Periodic and Aperiodic Signals

The signal defined by (2.1) is an example of a *periodic signal*. A signal $x(t)$ is periodic if and only if

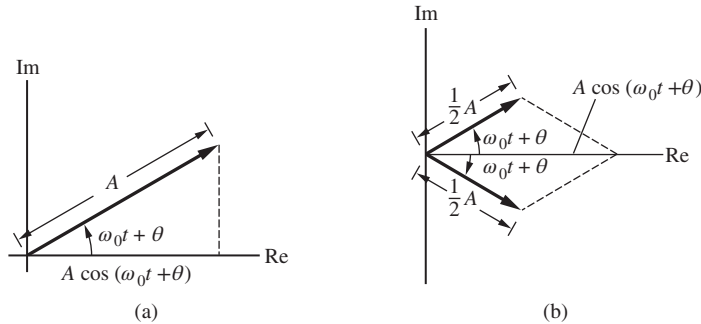
$$x(t + T_0) = x(t), \quad -\infty < t < \infty \quad (2.3)$$

where the constant T_0 is the period. The smallest such number satisfying (2.3) is referred to as the *fundamental period* (the modifier “fundamental” is often excluded). Any signal not satisfying (2.3) is called *aperiodic*.

2.1.3 Phasor Signals and Spectra

A useful periodic signal in system analysis is the signal

$$\tilde{x}(t) = Ae^{j(\omega_0 t + \theta)}, \quad -\infty < t < \infty \quad (2.4)$$

**Figure 2.2**

Two ways of relating a phasor signal to a sinusoidal signal. (a) Projection of a rotating phasor onto the real axis. (b) Addition of complex conjugate rotating phasors.

which is characterized by three parameters: amplitude A , phase θ in radians, and frequency ω_0 in radians per second or $f_0 = \omega_0/2\pi$ hertz. We will refer to $\tilde{x}(t)$ as a *rotating phasor* to distinguish it from the phasor $Ae^{j\theta}$, for which $e^{j\omega_0 t}$ is implicit. Using Euler's theorem,² we may readily show that $\tilde{x}(t) = \tilde{x}(t + T_0)$, where $T_0 = 2\pi/\omega_0$. Thus, $\tilde{x}(t)$ is a periodic signal with period $2\pi/\omega_0$.

The rotating phasor $Ae^{j(\omega_0 t + \theta)}$ can be related to a real, sinusoidal signal $A \cos(\omega_0 t + \theta)$ in two ways. The first is by taking its real part,

$$\begin{aligned} x(t) &= A \cos(\omega_0 t + \theta) = \operatorname{Re} \tilde{x}(t) \\ &= \operatorname{Re} A e^{j(\omega_0 t + \theta)} \end{aligned} \quad (2.5)$$

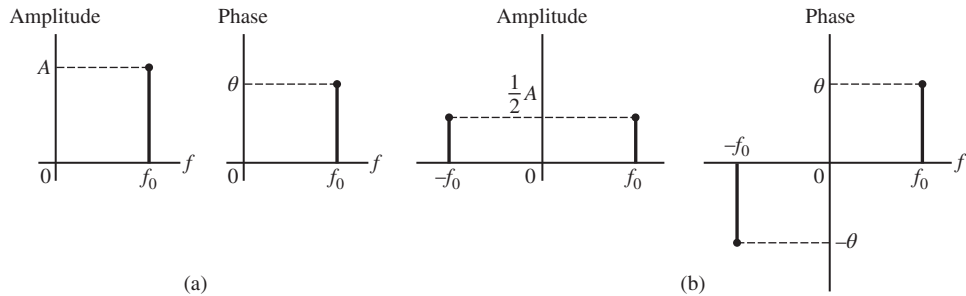
and the second is by taking one-half of the sum of $\tilde{x}(t)$ and its complex conjugate,

$$\begin{aligned} A \cos(\omega_0 t + \theta) &= \frac{1}{2} \tilde{x}(t) + \frac{1}{2} \tilde{x}^*(t) \\ &= \frac{1}{2} A e^{j(\omega_0 t + \theta)} + \frac{1}{2} A e^{-j(\omega_0 t + \theta)} \end{aligned} \quad (2.6)$$

Figure 2.2 illustrates these two procedures graphically.

Equations (2.5) and (2.6), which give alternative representations of the sinusoidal signal $x(t) = A \cos(\omega_0 t + \theta)$ in terms of the rotating phasor $\tilde{x}(t) = A \exp[j(\omega_0 t + \theta)]$, are time-domain representations for $x(t)$. Two equivalent representations of $x(t)$ in the frequency domain may be obtained by noting that the rotating phasor signal is completely specified if the parameters A and θ are given for a particular f_0 . Thus, plots of the magnitude and angle of $Ae^{j\theta}$ versus frequency give sufficient information to characterize $x(t)$ completely. Because $\tilde{x}(t)$ exists only at the single frequency, f_0 , for this case of a single sinusoidal signal, the resulting plots consist of discrete lines and are known as *line spectra*. The resulting plots are referred to as the *amplitude line spectrum* and the *phase line spectrum* for $x(t)$, and are shown in Figure 2.3(a). These are *frequency-domain* representations not only of $\tilde{x}(t)$ but of $x(t)$ as well, by virtue of (2.5). In addition, the plots of Figure 2.3(a) are referred to as the *single-sided amplitude and phase spectra* of $x(t)$ because they exist only for positive frequencies. For a

²Recall that Euler's theorem is $e^{\pm ju} = \cos u \pm j \sin u$. Also recall that $e^{j2\pi} = 1$.

**Figure 2.3**

Amplitude and phase spectra for the signal $A \cos(\omega_0 t + \theta)$ (a) Single-sided. (b) Double-sided.

signal consisting of a sum of sinusoids of differing frequencies, the single-sided spectrum consists of a multiplicity of lines, with one line for each sinusoidal component of the sum.

By plotting the amplitude and phase of the complex conjugate phasors of (2.6) versus frequency, one obtains another frequency-domain representation for $x(t)$, referred to as the *double-sided amplitude and phase spectra*. This representation is shown in Figure 2.3(b). Two important observations may be made from Figure 2.3(b). First, the lines at the *negative* frequency $f = -f_0$ exist precisely because it is necessary to add complex conjugate (or oppositely rotating) phasor signals to obtain the real signal $A \cos(\omega_0 t + \theta)$. Second, we note that the amplitude spectrum has *even* symmetry and that the phase spectrum has *odd* symmetry about $f = 0$. This symmetry is again a consequence of $x(t)$ being a real signal. As in the single-sided case, the two-sided spectrum for a sum of sinusoids consists of a multiplicity of lines, with one pair of lines for each sinusoidal component.

Figures 2.3(a) and 2.3(b) are therefore equivalent spectral representations for the signal $A \cos(\omega_0 t + \theta)$, consisting of lines at the frequency $f = f_0$ (and its negative). For this simple case, the use of spectral plots seems to be an unnecessary complication, but we will find shortly how the Fourier series and Fourier transform lead to spectral representations for more complex signals.

EXAMPLE 2.1

(a) To sketch the single-sided and double-sided spectra of

$$x(t) = 2 \sin \left(10\pi t - \frac{1}{6}\pi \right) \quad (2.7)$$

we note that $x(t)$ can be written as

$$\begin{aligned} x(t) &= 2 \cos \left(10\pi t - \frac{1}{6}\pi - \frac{1}{2}\pi \right) = 2 \cos \left(10\pi t - \frac{2}{3}\pi \right) \\ &= \operatorname{Re} 2e^{j(10\pi t - 2\pi/3)} = e^{j(10\pi t - 2\pi/3)} + e^{-j(10\pi t - 2\pi/3)} \end{aligned} \quad (2.8)$$

Thus, the single-sided and double-sided spectra are as shown in Figure 2.3, with $A = 2$, $\theta = -\frac{2}{3}\pi$ rad, and $f_0 = 5$ Hz.

(b) If more than one sinusoidal component is present in a signal, its spectra consist of multiple lines. For example, the signal

$$y(t) = 2 \sin\left(10\pi t - \frac{1}{6}\pi\right) + \cos(20\pi t) \quad (2.9)$$

can be rewritten as

$$\begin{aligned} y(t) &= 2 \cos\left(10\pi t - \frac{2}{3}\pi\right) + \cos(20\pi t) \\ &= \operatorname{Re} [2e^{j(10\pi t - 2\pi/3)} + e^{j20\pi t}] \\ &= e^{j(10\pi t - 2\pi/3)} + e^{-j(10\pi t - 2\pi/3)} + \frac{1}{2}e^{j20\pi t} + \frac{1}{2}e^{-j20\pi t} \end{aligned} \quad (2.10)$$

Its single-sided amplitude spectrum consists of a line of amplitude 2 at $f = 5$ Hz and a line of amplitude 1 at $f = 10$ Hz. Its single-sided phase spectrum consists of a single line of amplitude $-2\pi/3$ radians at $f = 5$ Hz (the phase at 10 Hz is zero). To get the double-sided amplitude spectrum, one simply *halves* the amplitude of the lines in the single-sided amplitude spectrum and takes the mirror image of this result about $f = 0$ (amplitude lines at $f = 0$, if present, remain the same). The double-sided phase spectrum is obtained by taking the mirror image of the single-sided phase spectrum about $f = 0$ and inverting the left-hand (negative frequency) portion. ■

2.1.4 Singularity Functions

An important subclass of aperiodic signals is the singularity functions. In this book we will be concerned with only two: the *unit impulse function* $\delta(t)$ (or *delta function*) and the *unit step function* $u(t)$. The unit impulse function is defined in terms of the integral

$$\int_{-\infty}^{\infty} x(t) \delta(t) dt = x(0) \quad (2.11)$$

where $x(t)$ is any test function that is continuous at $t = 0$. A change of variables and redefinition of $x(t)$ results in the *sifting property*

$$\int_{-\infty}^{\infty} x(t) \delta(t - t_0) dt = x(t_0) \quad (2.12)$$

where $x(t)$ is continuous at $t = t_0$. We will make considerable use of the sifting property in systems analysis. By considering the special case $x(t) = 1$ for $t_1 \leq t \leq t_2$ and $x(t) = 0$ for $t < t_1$ and $t > t_2$ the two properties

$$\int_{t_1}^{t_2} \delta(t - t_0) dt = 1, \quad t_1 < t_0 < t_2 \quad (2.13)$$

and

$$\delta(t - t_0) = 0, \quad t \neq t_0 \quad (2.14)$$

are obtained that provide an alternative definition of the unit impulse. Equation (2.14) allows the integrand in Equation (2.12) to be replaced by $x(t_0)\delta(t - t_0)$, and the sifting property then follows from (2.13).

Other properties of the unit impulse function that can be proved from the definition (2.11) are the following:

1. $\delta(at) = \frac{1}{|a|}\delta(t)$, a is a constant
2. $\delta(-t) = \delta(t)$
3. $\int_{t_1}^{t_2} x(t)\delta(t - t_0)dt = \begin{cases} x(t_0), & t_1 < t_0 < t_2 \\ 0, & \text{otherwise} \end{cases}$ (a generalization of the sifting property)
undefined for $t_0 = t_1$ or t_2
4. $x(t)\delta(t - t_0) = x(t_0)\delta(t - t_0)$ where $x(t)$ is continuous at $t = t_0$
5. $\int_{t_1}^{t_2} x(t)\delta^{(n)}(t - t_0)dt = (-1)^n x^{(n)}(t_0)$, $t_1 < t_0 < t_2$. [In this equation, the superscript (n) denotes the n th derivative; $x(t)$ and its first n derivatives are assumed continuous at $t = t_0$.]
6. If $f(t) = g(t)$, where $f(t) = a_0\delta(t) + a_1\delta^{(1)}(t) + \dots + a_n\delta^{(n)}(t)$ and $g(t) = b_0\delta(t) + b_1\delta^{(1)}(t) + \dots + b_n\delta^{(n)}(t)$, this implies that $a_0 = b_0, a_1 = b_1, \dots, a_n = b_n$

It is reassuring to note that (2.13) and (2.14) correspond to the intuitive notion of a unit impulse function as the limit of a suitably chosen conventional function having unity area in an infinitesimally small width. An example is the signal

$$\delta_\epsilon(t) = \frac{1}{2\epsilon} \Pi\left(\frac{t}{2\epsilon}\right) = \begin{cases} \frac{1}{2\epsilon}, & |t| < \epsilon \\ 0, & \text{otherwise} \end{cases} \quad (2.15)$$

which is shown in Figure 2.4(a) for $\epsilon = 1/4$ and $\epsilon = 1/2$. It seems apparent that any signal having unity area and zero width in the limit as some parameter approaches zero is a suitable representation for $\delta(t)$, for example, the signal

$$\delta_{1\epsilon}(t) = \epsilon \left(\frac{1}{\pi t} \sin \frac{\pi t}{\epsilon} \right)^2 \quad (2.16)$$

which is sketched in Figure 2.4(b).

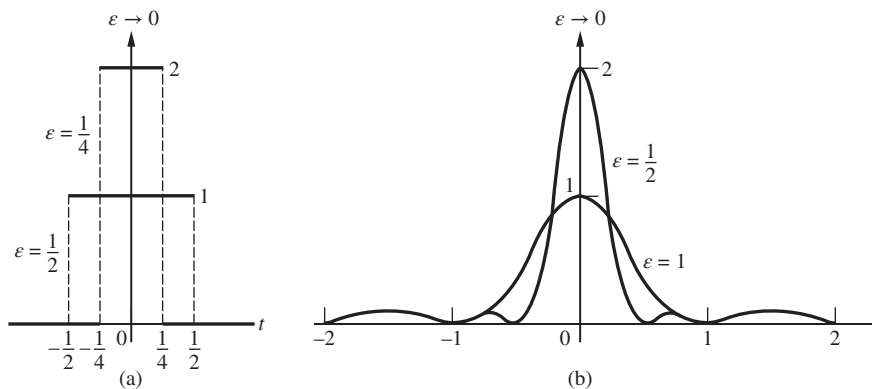


Figure 2.4

Two representations for the unit impulse function in the limit as $\epsilon \rightarrow 0$. (a) $\left(\frac{1}{2\epsilon}\right) \Pi(t/2\epsilon)$.
(b) $\epsilon \left[\left(\frac{1}{\pi t} \sin(\pi t/\epsilon) \right)^2 \right]$.

Other singularity functions may be defined as integrals or derivatives of unit impulses. We will need only the unit step $u(t)$, defined to be the integral of the unit impulse. Thus,

$$u(t) \triangleq \int_{-\infty}^t \delta(\lambda) d\lambda = \begin{cases} 0, & t < 0 \\ 1, & t > 0 \\ \text{undefined,} & t = 0 \end{cases} \quad (2.17)$$

or

$$\delta(t) = \frac{du(t)}{dt} \quad (2.18)$$

(For consistency with the unit pulse function definition, we will define $u(0) = 1$). You are no doubt familiar with the usefulness of the unit step for “turning on” signals of doubly infinite duration and for representing signals of the staircase type. For example, the unit rectangular pulse function defined by (2.2) can be written in terms of unit steps as

$$\Pi(t) = u\left(t + \frac{1}{2}\right) - u\left(t - \frac{1}{2}\right) \quad (2.19)$$

EXAMPLE 2.2

To illustrate calculations with the unit impulse function, consider evaluation of the following expressions:

- $\int_2^5 \cos(3\pi t) \delta(t-1) dt$;
- $\int_0^5 \cos(3\pi t) \delta(t-1) dt$;
- $\int_0^5 \cos(3\pi t) \frac{d\delta(t-1)}{dt} dt$;
- $\int_{-10}^{10} \cos(3\pi t) \delta(2t) dt$;
- $2\delta(t) + 3\frac{d\delta(t)}{dt} = a\delta(t) + b\frac{d\delta(t)}{dt} + c\frac{d^2\delta(t)}{dt^2}$, find a , b , and c ;
- $\frac{d}{dt} [e^{-4t}u(t)]$;

Solution

- This integral evaluates to 0 because the unit impulse function is outside the limits of integration;
- This integral evaluates to $\cos(3\pi t)|_{t=1} = \cos(3\pi) = -1$;
- $\int_0^5 \cos(3\pi t) \frac{d\delta(t-1)}{dt} dt = (-1) \frac{d}{dt} [\cos(3\pi t)]_{t=1} = 3\pi \sin(3\pi) = 0$;
- $\int_{-10}^{10} \cos(3\pi t) \delta(2t) dt = \int_{-20}^{20} \cos(3\pi t) \frac{1}{2} \delta(t) dt = \frac{1}{2} \cos(0) = \frac{1}{2}$ by using property 1 above;
- $2\delta(t) + 3\frac{d\delta(t)}{dt} = a\delta(t) + b\frac{d\delta(t)}{dt} + c\frac{d^2\delta(t)}{dt^2}$ gives $a = 2$, $b = 3$, and $c = 0$ by using property 6 above;
- Using the chain rule for differentiation and $\delta(t) = \frac{du(t)}{dt}$, we get $\frac{d}{dt} [e^{-4t}u(t)] = -4e^{-4t}u(t) + e^{-4t} \frac{du(t)}{dt} = -4e^{-4t}u(t) + e^{-4t} \delta(t) = -4e^{-4t}u(t) + \delta(t)$, where property 4 and (2.18) have been used. ■

We are now ready to consider power and energy signal classifications.

2.2 SIGNAL CLASSIFICATIONS

Because the particular representation used for a signal depends on the type of signal involved, it is useful to pause at this point and introduce signal classifications. In this chapter we will be considering two signal classes, those with finite energy and those with finite power. As a specific example, suppose $e(t)$ is the voltage across a resistance R producing a current $i(t)$. The instantaneous power per ohm is $p(t) = e(t)i(t)/R = i^2(t)$. Integrating over the interval $|t| \leq T$, the total energy and the average power on a per-ohm basis are obtained as the limits

$$E = \lim_{T \rightarrow \infty} \int_{-T}^T i^2(t) dt \quad (2.20)$$

and

$$P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T i^2(t) dt \quad (2.21)$$

respectively.

For an arbitrary signal $x(t)$, which may, in general, be complex, we define total (normalized) energy as

$$E \triangleq \lim_{T \rightarrow \infty} \int_{-T}^T |x(t)|^2 dt = \int_{-\infty}^{\infty} |x(t)|^2 dt \quad (2.22)$$

and (normalized) power as

$$P \triangleq \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |x(t)|^2 dt \quad (2.23)$$

Based on the definitions (2.22) and (2.23), we can define two distinct classes of signals:

1. We say $x(t)$ is an *energy signal* if and only if $0 < E < \infty$, so that $P = 0$.
2. We classify $x(t)$ as a *power signal* if and only if $0 < P < \infty$, thus implying that $E = \infty$.³

EXAMPLE 2.3

As an example of determining the classification of a signal, consider

$$x_1(t) = Ae^{-\alpha t}u(t), \quad \alpha > 0 \quad (2.24)$$

where A and α are positive constants. Using (2.22), we may readily verify that $x_1(t)$ is an *energy signal*, since $E = A^2/2\alpha$ by applying (2.22). Letting $\alpha \rightarrow 0$, we obtain the signal $x_2(t) = Au(t)$, which has infinite energy. Applying (2.23), we find that $P = \frac{1}{2}A^2$ for $Au(t)$, thus verifying that $x_2(t)$ is a *power signal*. ■

³Signals that are neither energy nor power signals are easily found. For example, $x(t) = t^{-1/4}, t \geq t_0 > 0$, and zero otherwise.

EXAMPLE 2.4

Consider the rotating phasor signal given by Equation (2.4). We may verify that $\tilde{x}(t)$ is a power signal, since

$$P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |\tilde{x}(t)|^2 dt = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-\infty}^{\infty} |Ae^{j(\omega_0 t + \theta)}|^2 dt = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T A^2 dt = A^2 \quad (2.25)$$

is finite. ■

We note that there is no need to carry out the limiting operation to find P for a periodic signal, since an average carried out over a single period gives the same result as (2.23); that is, for a periodic signal $x_p(t)$,

$$P = \frac{1}{T_0} \int_{t_0}^{t_0+T_0} |x_p(t)|^2 dt \quad (2.26)$$

where T_0 is the period and t_0 is an arbitrary starting time (chosen for convenience). The proof of (2.26) is left to the problems.

EXAMPLE 2.5

The sinusoidal signal

$$x_p(t) = A \cos(\omega_0 t + \theta) \quad (2.27)$$

has average power

$$\begin{aligned} P &= \frac{1}{T_0} \int_{t_0}^{t_0+T_0} A^2 \cos^2(\omega_0 t + \theta) dt \\ &= \frac{\omega_0}{2\pi} \int_{t_0}^{t_0+(2\pi/\omega_0)} \frac{A^2}{2} dt + \frac{\omega_0}{2\pi} \int_{t_0}^{t_0+(2\pi/\omega_0)} \frac{A^2}{2} \cos [2(\omega_0 t + \theta)] dt \\ &= \frac{A^2}{2} \end{aligned} \quad (2.28)$$

where the identity $\cos^2(u) = \frac{1}{2} + \frac{1}{2} \cos(2u)$ has been used⁴ and the second integral is zero because the integration is over two complete periods of the integrand. ■

⁴See Appendix F.2 for trigonometric identities.

2.3 FOURIER SERIES

2.3.1 Complex Exponential Fourier Series

Given a signal $x(t)$ defined over the interval $(t_0, t_0 + T_0)$ with the definition $\omega_0 = 2\pi f_0 = \frac{2\pi}{T_0}$ we define the *complex exponential Fourier series* as

$$x(t) = \sum_{n=-\infty}^{\infty} X_n e^{jn\omega_0 t}, \quad t_0 \leq t < t_0 + T_0 \quad (2.29)$$

where

$$X_n = \frac{1}{T_0} \int_{t_0}^{t_0+T_0} x(t) e^{-jn\omega_0 t} dt \quad (2.30)$$

It can be shown to represent the signal $x(t)$ exactly in the interval $(t_0, t_0 + T_0)$, except at a point of jump discontinuity where it converges to the arithmetic mean of the left-hand and right-hand limits.⁵ Outside the interval $(t_0, t_0 + T_0)$, of course, nothing is guaranteed. However, we note that the right-hand side of (2.29) is periodic with period T_0 , since it is the sum of periodic rotating phasors with harmonic frequencies. Thus, if $x(t)$ is periodic with period T_0 , the Fourier series of (2.29) is an accurate representation for $x(t)$ for *all* t (except at points of discontinuity). The integration of (2.30) can then be taken over any period.

A useful observation about a Fourier series expansion of a signal is that the series is unique. For example, if we somehow find a Fourier expansion for a signal $x(t)$, we know that no other Fourier expansion for that $x(t)$ exists. The usefulness of this observation is illustrated with the following example.

EXAMPLE 2.6

Consider the signal

$$x(t) = \cos(\omega_0 t) + \sin^2(2\omega_0 t) \quad (2.31)$$

where $\omega_0 = 2\pi/T_0$. Find the complex exponential Fourier series.

Solution

We could compute the Fourier coefficients using (2.30), but by using appropriate trigonometric identities and Euler's theorem, we obtain

$$\begin{aligned} x(t) &= \cos(\omega_0 t) + \frac{1}{2} - \frac{1}{2} \cos(4\omega_0 t) \\ &= \frac{1}{2} e^{j\omega_0 t} + \frac{1}{2} e^{-j\omega_0 t} + \frac{1}{2} - \frac{1}{4} e^{j4\omega_0 t} - \frac{1}{4} e^{-j4\omega_0 t} \end{aligned} \quad (2.32)$$

Invoking uniqueness and equating the second line term by term with $\sum_{n=-\infty}^{\infty} X_n e^{jn\omega_0 t}$ we find that

$$X_0 = \frac{1}{2}$$

⁵Dirichlet's conditions state that sufficient conditions for convergence are that $x(t)$ be defined and bounded on the range $(t_0, t_0 + T_0)$ and have only a finite number of maxima and minima and a finite number of discontinuities on this interval.

$$X_1 = \frac{1}{2} = X_{-1} \quad (2.33)$$

$$X_4 = -\frac{1}{4} = X_{-4}$$

with all other X_n s equal to zero. Thus considerable labor is saved by noting that the Fourier series of a signal is unique. ■

2.3.2 Symmetry Properties of the Fourier Coefficients

Assuming $x(t)$ is real, it follows from (2.30) that

$$X_n^* = X_{-n} \quad (2.34)$$

by taking the complex conjugate inside the integral and noting that the same result is obtained by replacing n by $-n$. Writing X_n as

$$X_n = |X_n| e^{j\angle X_n} \quad (2.35)$$

we obtain

$$|X_n| = |X_{-n}| \quad \text{and} \quad \angle X_n = -\angle X_{-n} \quad (2.36)$$

Thus, for real signals, the magnitude of the Fourier coefficients is an even function of n , and the argument is odd.

Several symmetry properties can be derived for the Fourier coefficients, depending on the symmetry of $x(t)$. For example, suppose $x(t)$ is even; that is, $x(t) = x(-t)$. Then, using Euler's theorem to write the expression for the Fourier coefficients as (choose $t_0 = -T_0/2$)

$$X_n = \frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(t) \cos(n\omega_0 t) dt - \frac{j}{T_0} \int_{-T_0/2}^{T_0/2} x(t) \sin(n\omega_0 t) dt, \quad (2.37)$$

we see that the second term is zero, since $x(t) \sin(n\omega_0 t)$ is an odd function. Thus, X_n is purely real, and furthermore, X_n is an even function of n since $\cos n\omega_0 t$ is an even function of n . These consequences of $x(t)$ being even are illustrated by Example 2.6.

On the other hand, if $x(t) = -x(-t)$ [that is, $x(t)$ is odd], it readily follows that X_n is purely imaginary, since the first term in (2.37) is zero by virtue of $x(t) \cos(n\omega_0 t)$ being odd. In addition, X_n is an odd function of n , since $\sin(n\omega_0 t)$ is an odd function of n .

Another type of symmetry is (*odd*) *halfwave symmetry*, defined as

$$x\left(t \pm \frac{1}{2}T_0\right) = -x(t) \quad (2.38)$$

where T_0 is the period of $x(t)$. For signals with odd halfwave symmetry,

$$X_n = 0, \quad n = 0, \pm 2, \pm 4, \dots \quad (2.39)$$

which states that the Fourier series for such a signal consists only of odd-indexed terms. The proof of this is left to the problems.

2.3.3 Trigonometric Form of the Fourier Series

Using (2.36) and assuming $x(t)$ real, we can regroup the complex exponential Fourier series by pairs of terms of the form

$$\begin{aligned} X_n e^{jn\omega_0 t} + X_{-n} e^{-jn\omega_0 t} &= |X_n| e^{j(n\omega_0 t + \angle X_n)} + |X_n| e^{-j(n\omega_0 t + \angle X_n)} \\ &= 2 |X_n| \cos(n\omega_0 t + \angle X_n) \end{aligned} \quad (2.40)$$

where the facts that $|X_n| = |X_{-n}|$ and $\angle X_n = -\angle X_{-n}$ have been used. Hence, (2.29) can be written in the equivalent trigonometric form:

$$x(t) = X_0 + \sum_{n=1}^{\infty} 2 |X_n| \cos(n\omega_0 t + \angle X_n) \quad (2.41)$$

Expanding the cosine in (2.41), we obtain still another equivalent series of the form

$$x(t) = X_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega_0 t) + \sum_{n=1}^{\infty} B_n \sin(n\omega_0 t) \quad (2.42)$$

where

$$\begin{aligned} A_n &= 2 |X_n| \cos \angle X_n \\ &= \frac{2}{T_0} \int_{t_0}^{t_0+T_0} x(t) \cos(n\omega_0 t) dt \end{aligned} \quad (2.43)$$

and

$$\begin{aligned} B_n &= -2 |X_n| \sin \angle X_n \\ &= \frac{2}{T_0} \int_{t_0}^{t_0+T_0} x(t) \sin(n\omega_0 t) dt \end{aligned} \quad (2.44)$$

In either the trigonometric or the exponential forms of the Fourier series, X_0 represents the average or DC component of $x(t)$. The term for $n = 1$ is called the *fundamental* (along with the term for $n = -1$ if we are dealing with the complex exponential series), the term for $n = 2$ is called the *second harmonic*, and so on.

2.3.4 Parseval's Theorem

Using (2.26) for average power of a periodic signal,⁶ substituting (2.29) for $x(t)$, and interchanging the order of integration and summation, we obtain

$$P = \frac{1}{T_0} \int_{T_0} |x(t)|^2 dt = \frac{1}{T_0} \int_{T_0} \left(\sum_{m=-\infty}^{\infty} X_m e^{jm\omega_0 t} \right) \left(\sum_{n=-\infty}^{\infty} X_n e^{jn\omega_0 t} \right)^* dt = \sum_{n=-\infty}^{\infty} |X_n|^2 \quad (2.45)$$

⁶ $\int_{T_0} (\) dt$ represents integration over any period.

or

$$P = X_0^2 + \sum_{n=1}^{\infty} 2 |X_n|^2 \quad (2.46)$$

which is called Parseval's theorem. In words, (2.45) simply states that the average power of a periodic signal $x(t)$ is the sum of the powers in the phasor components of its Fourier series, or (2.46) states that its average power is the sum of the powers in its DC component plus that in its AC components [from (2.41) the power in each cosine component is its amplitude squared divided by 2, or $(2 |X_n|)^2 / 2 = 2 |X_n|^2$]. Note that powers of the Fourier components can be added because they are orthogonal (i.e., the integral of the product of two harmonics is zero).

2.3.5 Examples of Fourier Series

Table 2.1 gives Fourier series for several commonly occurring periodic waveforms. The left-hand column specifies the signal over one period. The definition of periodicity,

$$x(t) = x(t + T_0)$$

specifies it for all t . The derivation of the Fourier coefficients given in the right-hand column of Table 2.1 is left to the problems. Note that the full-rectified sinewave actually has the period $\frac{1}{2}T_0$.

Table 2.1 Fourier Series for Several Periodic Signals

Signal (one period)	Coefficients for exponential Fourier series
1. Asymmetrical pulse train; period = T_0 : $x(t) = A\Pi\left(\frac{t-t_0}{\tau}\right), \tau < T_0$ $x(t) = x(t + T_0), \text{ all } t$	$X_n = \frac{A\tau}{T_0} \text{sinc}(nf_0\tau) e^{-j2\pi n f_0 t_0}$ $n = 0, \pm 1, \pm 2, \dots$
2. Half-rectified sinewave; period = $T_0 = 2\pi/\omega_0$: $x(t) = \begin{cases} A \sin(\omega_0 t), & 0 \leq t \leq T_0/2 \\ 0, & -T_0/2 \leq t \leq 0 \end{cases}$ $x(t) = x(t + T_0), \text{ all } t$	$X_n = \begin{cases} \frac{A}{\pi(1-n^2)}, & n = 0, \pm 2, \pm 4, \dots \\ 0, & n = \pm 3, \pm 5, \dots \\ -\frac{1}{4}jnA, & n = \pm 1 \end{cases}$
3. Full-rectified sinewave; period = $T_0' = \pi/\omega_0$: $x(t) = A \sin(\omega_0 t) $	$X_n = \frac{2A}{\pi(1-4n^2)}, n = 0, \pm 1, \pm 2, \dots$
4. Triangular wave: $x(t) = \begin{cases} -\frac{4A}{T_0}t + A, & 0 \leq t \leq T_0/2 \\ \frac{4A}{T_0}t + A, & -T_0/2 \leq t \leq 0 \end{cases}$ $x(t) = x(t + T_0), \text{ all } t$	$X_n = \begin{cases} \frac{4A}{\pi^2 n^2}, & n \text{ odd} \\ 0, & n \text{ even} \end{cases}$

For the periodic pulse train, it is convenient to express the coefficients in terms of the *sinc function*, defined as

$$\text{sinc } z = \frac{\sin(\pi z)}{\pi z} \quad (2.47)$$

The sinc function is an even damped oscillatory function with zero crossings at integer values of its argument.

EXAMPLE 2.7

Specialize the results for the pulse train (no. 1) of Table 2.1 to the complex exponential and trigonometric Fourier series of a squarewave with even symmetry and amplitudes zero and A .

Solution

The solution proceeds by letting $t_0 = 0$ and $\tau = \frac{1}{2}T_0$ in item 1 of Table 2.1. Thus,

$$X_n = \frac{1}{2}A \text{sinc}\left(\frac{1}{2}n\right) \quad (2.48)$$

But

$$\begin{aligned} \text{sinc}(n/2) &= \frac{\sin(n\pi/2)}{n\pi/2} \\ &= \begin{cases} 1, & n = 0 \\ 0, & n = \text{even} \\ |2/n\pi|, & n = \pm 1, \pm 5, \pm 9, \dots \\ -|2/n\pi|, & n = \pm 3, \pm 7, \dots \end{cases} \end{aligned}$$

Thus,

$$\begin{aligned} x(t) &= \dots + \frac{A}{5\pi} e^{-j5\omega_0 t} - \frac{A}{3\pi} e^{-j3\omega_0 t} + \frac{A}{\pi} e^{-j\omega_0 t} \\ &\quad + \frac{A}{2} + \frac{A}{\pi} e^{j\omega_0 t} - \frac{A}{3\pi} e^{j3\omega_0 t} + \frac{A}{5\pi} e^{j5\omega_0 t} - \dots \\ &= \frac{A}{2} + \frac{2A}{\pi} \left[\cos(\omega_0 t) - \frac{1}{3} \cos(3\omega_0 t) + \frac{1}{5} \cos(5\omega_0 t) - \dots \right] \end{aligned} \quad (2.49)$$

The first equation is the complex exponential form of the Fourier series and the second equation is the trigonometric form. The DC component of this squarewave is $X_0 = \frac{1}{2}A$. Setting this term to zero in the preceding Fourier series, we have the Fourier series of a squarewave of amplitudes $\pm \frac{1}{2}A$. Such a squarewave has halfwave symmetry, and this is precisely the reason that no even harmonics are present in its Fourier series. ■

2.3.6 Line Spectra

The complex exponential Fourier series (2.29) of a signal is simply a summation of phasors. In Section 2.1 we showed how a phasor could be characterized in the frequency domain by two plots: one showing its amplitude versus frequency and one showing its phase. Similarly, a periodic signal can be characterized in the frequency domain by making two plots: one showing

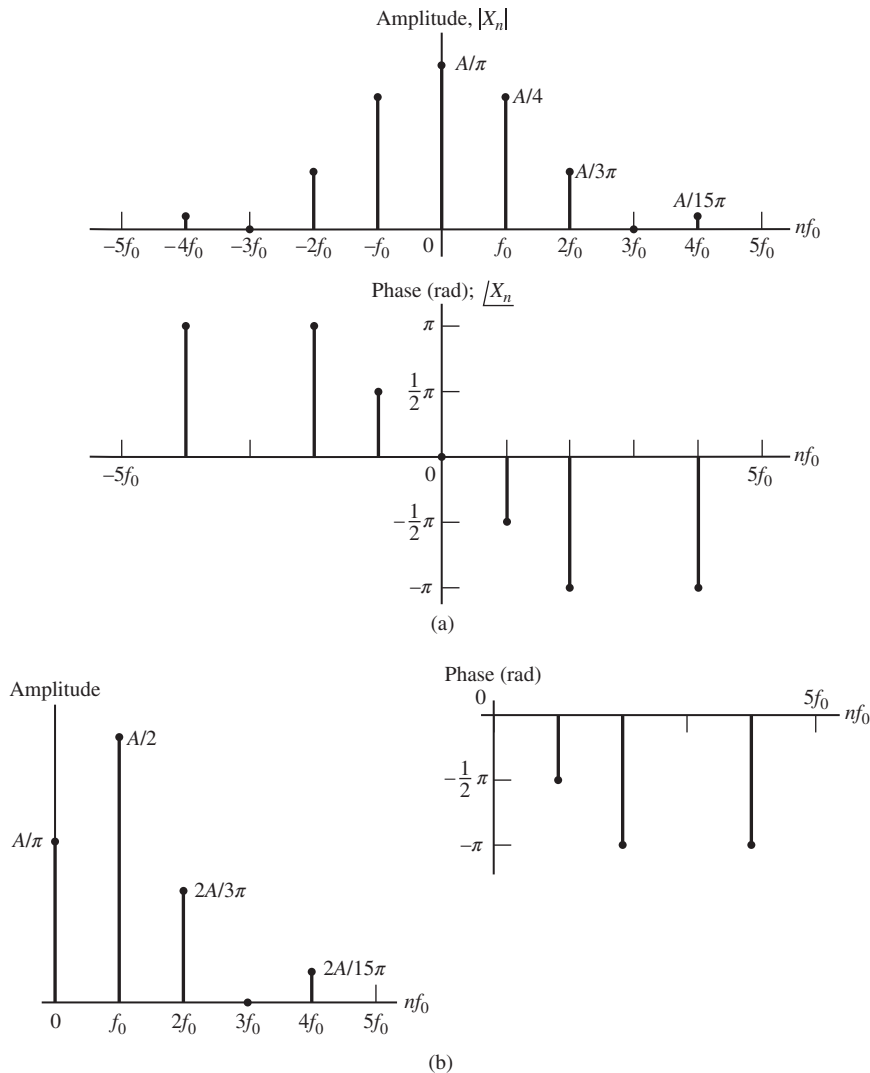


Figure 2.5 Line spectra for half-rectified sine wave. (a) Double-sided. (b) Single-sided.

amplitudes of the separate phasor components versus frequency and the other showing their phases versus frequency. The resulting plots are called the *two-sided amplitude*⁷ and *phase spectra*, respectively, of the signal. From (2.36) it follows that, for a real signal, the amplitude spectrum is even and the phase spectrum is odd, which is simply a result of the addition of complex conjugate phasors to get a real sinusoidal signal.

Figure 2.5(a) shows the double-sided spectrum for a half-rectified sine wave as plotted from the results given in Table 2.1. For $n = 2, 4, \dots$, X_n is represented as

⁷Magnitude spectrum would be a more accurate term, although *amplitude spectrum* is the customary term.

follows:

$$X_n = - \left| \frac{A}{\pi(1-n^2)} \right| = \frac{A}{\pi(n^2-1)} e^{-j\pi} \quad (2.50)$$

For $n = -2, -4, \dots$, it is represented as

$$X_n = - \left| \frac{A}{\pi(1-n^2)} \right| = \frac{A}{\pi(n^2-1)} e^{j\pi} \quad (2.51)$$

to ensure that the phase is odd, as it must be (note that $e^{\pm j\pi} = -1$). Thus, putting this together with $X_{\pm 1} = \mp jA/4$, we get

$$|X_n| = \begin{cases} \frac{1}{4}A, & n = \pm 1 \\ \left| \frac{A}{\pi(1-n^2)} \right|, & \text{all even } n \end{cases} \quad (2.52)$$

$$\angle X_n = \begin{cases} -\pi, & n = 2, 4, \dots \\ -\frac{1}{2}\pi, & n = 1 \\ 0, & n = 0 \\ \frac{1}{2}\pi, & n = -1 \\ \pi, & n = -2, -4, \dots \end{cases} \quad (2.53)$$

The single-sided line spectra are obtained by plotting the amplitudes and phase angles of the terms in the trigonometric Fourier series (2.41) versus nf_0 . Because the series (2.41) has only nonnegative frequency terms, the single-sided spectra exist only for $nf_0 \geq 0$. From (2.41) it is readily apparent that the single-sided phase spectrum of a periodic signal is identical to its double-sided phase spectrum for $nf_0 \geq 0$ and zero for $nf_0 < 0$. The single-sided amplitude spectrum is obtained from the double-sided amplitude spectrum by doubling the amplitudes of all lines for $nf_0 > 0$. The line at $nf_0 = 0$ stays the same. The single-sided spectra for the half-rectified sinewave are shown in Figure 2.5(b).

As a second example, consider the pulse train

$$x(t) = \sum_{n=-\infty}^{\infty} A\Pi\left(\frac{t - nT_0 - \frac{1}{2}\tau}{\tau}\right) \quad (2.54)$$

From Table 2.1, with $t_0 = \frac{1}{2}\tau$ substituted in item 1, the Fourier coefficients are

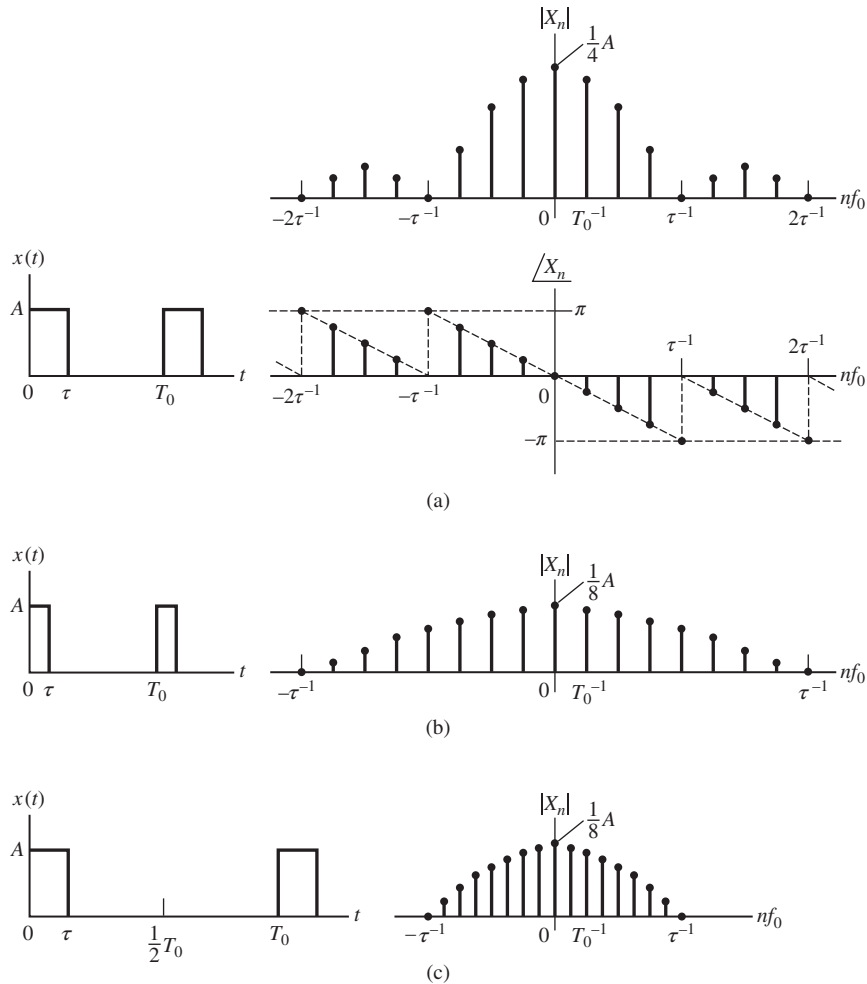
$$X_n = \frac{A\tau}{T_0} \text{sinc}(nf_0\tau) e^{-j\pi nf_0\tau} \quad (2.55)$$

The Fourier coefficients can be put in the form $|X_n| \exp(j\angle X_n)$, where

$$|X_n| = \frac{A\tau}{T_0} |\text{sinc}(nf_0\tau)| \quad (2.56)$$

and

$$\angle X_n = \begin{cases} -\pi nf_0\tau & \text{if } \text{sinc}(nf_0\tau) > 0 \\ -\pi nf_0\tau + \pi & \text{if } nf_0 > 0 \text{ and } \text{sinc}(nf_0\tau) < 0 \\ -\pi nf_0\tau - \pi & \text{if } nf_0 < 0 \text{ and } \text{sinc}(nf_0\tau) < 0 \end{cases} \quad (2.57)$$

**Figure 2.6**

Spectra for a periodic pulse train signal. (a) $\tau = \frac{1}{4}T_0$. (b) $\tau = \frac{1}{8}T_0$; T_0 same as in (a). (c) $\tau = \frac{1}{8}T_0$; τ same as in (a).

The $\pm\pi$ on the right-hand side of (2.57) on the second and third lines accounts for $|\text{sinc}(nf_0\tau)| = -\text{sinc}(nf_0\tau)$ whenever $\text{sinc}(nf_0\tau) < 0$. Since the phase spectrum must have odd symmetry if $x(t)$ is real, π is subtracted if $nf_0 < 0$ and added if $nf_0 > 0$. The reverse could have been done—the choice is arbitrary. With these considerations, the double-sided amplitude and phase spectra can now be plotted. They are shown in Figure 2.6 for several choices of τ and T_0 . Note that appropriate multiples of 2π are added or subtracted from the lines in the phase spectrum ($e^{\pm j2\pi} = 1$).

Comparing Figures 2.6(a) and 2.6(b), we note that the zeros of the envelope of the amplitude spectrum, which occur at multiples of $1/\tau$ Hz, move out along the frequency axis as the pulse width decreases. That is, *the time duration of a signal and its spectral width are inversely proportional*, a property that will be shown to be true in general later. Second,

comparing Figures 2.6(a) and 2.6(c), we note that the separation between lines in the spectra is $1/T_0$. Thus, the density of the spectral lines with frequency increases as the period of $x(t)$ increases.

COMPUTER EXAMPLE 2.1

The MATLAB™ program given below computes the amplitude and phase spectra for a half-rectified sinewave. The stem plots produced look exactly the same as those in Figure 2.5(a). Programs for plotting spectra of other waveforms are left to the computer exercises.

```
% file ch2ce1
% Plot of line spectra for half-rectified sinewave
%
clf
A = 1;
n.max = 11;           % maximum harmonic plotted
n = -n.max:1:n.max;
X = zeros(size(n));  % set all lines = 0; fill in nonzero ones
I = find(n == 1);
II = find(n == -1);
III = find(mod(n, 2) == 0);
X(I) = -j*A/4;
X(II) = j*A/4;
X(III) = A./(pi*(1. - n(III).^2));
[argX, magX] = cart2pol(real(X),imag(X)); % Convert to magnitude and
phase
IV = find(n >= 2 & mod(n, 2) == 0);
argX(IV) = argX(IV) - 2*pi;           % force phase to be odd
magXss(1:n.max) = 2*magX(n.max+1:2*n.max);
magXss(1) = magXss(1)/2;
argXss(1:n.max) = argX(n.max+1:2*n.max);
nn = 1:n.max;
subplot(2,2,1), stem(n, magX), ylabel('Amplitude'), xlabel('\itnf}0,
Hz'),...
    axis([-10.1 10.1 0 0.5])
subplot(2,2,2), stem(n, argX), xlabel('\itnf}0, Hz'), ylabel('Phase,
rad'),...
    axis([-10.1 10.1 -4 4])
subplot(2,2,3), stem(nn-1, magXss), ylabel('Amplitude'),
xlabel('\itnf}0, Hz')
subplot(2,2,4), stem(nn-1, argXss), xlabel('\itnf}0, Hz'),
ylabel('Phase, rad'),...
    xlabel('\itnf}0')
% End of script file
```

2.4 THE FOURIER TRANSFORM

To generalize the Fourier series representation (2.29) to a representation valid for aperiodic signals, we consider the two basic relationships (2.29) and (2.30). Suppose that $x(t)$ is aperiodic

but is an energy signal, so that it is integrable square in the interval $(-\infty, \infty)$.⁸ In the interval $|t| < \frac{1}{2}T_0$, we can represent $x(t)$ as the Fourier series

$$x(t) = \sum_{n=-\infty}^{\infty} \left[\frac{1}{T_0} \int_{-T_0/2}^{T_0/2} x(\lambda) e^{-j2\pi f_0 \lambda} d\lambda \right] e^{j2\pi n f_0 t}, \quad |t| < \frac{T_0}{2} \quad (2.58)$$

where $f_0 = 1/T_0$. To represent $x(t)$ for all time, we simply let $T_0 \rightarrow \infty$ such that $n f_0 = n/T_0$ becomes the continuous variable f , $1/T_0$ becomes the differential df , and the summation becomes an integral. Thus,

$$x(t) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} x(\lambda) e^{-j2\pi f \lambda} d\lambda \right] e^{j2\pi f t} df \quad (2.59)$$

Defining the inside integral as

$$X(f) = \int_{-\infty}^{\infty} x(\lambda) e^{-j2\pi f \lambda} d\lambda \quad (2.60)$$

we can write (2.59) as

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df \quad (2.61)$$

The existence of these integrals is assured, since $x(t)$ is an energy signal. We note that

$$X(f) = \lim_{T_0 \rightarrow \infty} T_0 X_n \quad (2.62)$$

which avoids the problem that $|X_n| \rightarrow 0$ as $T_0 \rightarrow \infty$.

The frequency-domain description of $x(t)$ provided by (2.60) is referred to as the *Fourier transform* of $x(t)$, written symbolically as $X(f) = \mathfrak{F}[x(t)]$. Conversion back to the time domain is achieved via the *inverse Fourier transform* (2.61), written symbolically as $x(t) = \mathfrak{F}^{-1}[X(f)]$.

Expressing (2.60) and (2.61) in terms of $f = \omega/2\pi$ results in easily remembered symmetrical expressions. Integrating (2.61) with respect to the variable ω requires a factor of $(2\pi)^{-1}$.

2.4.1 Amplitude and Phase Spectra

Writing $X(f)$ in terms of amplitude and phase as

$$X(f) = |X(f)| e^{j\theta(f)}, \quad \theta(f) = \angle X(f) \quad (2.63)$$

we can show, for real $x(t)$, that

$$|X(f)| = |X(-f)| \quad \text{and} \quad \theta(f) = -\theta(-f) \quad (2.64)$$

⁸Actually if $\int_{-\infty}^{\infty} |x(t)| dt < \infty$, the Fourier-transform integral converges. It more than suffices if $x(t)$ is an energy signal. Dirichlet's conditions give sufficient conditions for a signal to have a Fourier transform. In addition to being absolutely integrable, $x(t)$ should be single-valued with a finite number of maxima and minima and a finite number of discontinuities in any finite time interval.

just as for the Fourier series. This is done by using Euler's theorem to write (2.60) in terms of its real and imaginary parts:

$$R = \operatorname{Re} X(f) = \int_{-\infty}^{\infty} x(t) \cos(2\pi ft) dt \quad (2.65)$$

and

$$I = \operatorname{Im} X(f) = - \int_{-\infty}^{\infty} x(t) \sin(2\pi ft) dt \quad (2.66)$$

Thus, the real part of $X(f)$ is even and the imaginary part is odd if $x(t)$ is a real signal. Since $|X(f)|^2 = R^2 + I^2$ and $\tan \theta(f) = I/R$, the symmetry properties (2.64) follow. A plot of $|X(f)|$ versus f is referred to as the *amplitude spectrum*⁹ of $x(t)$, and a plot of $\angle X(f) = \theta(f)$ versus f is known as the *phase spectrum*.

2.4.2 Symmetry Properties

If $x(t) = x(-t)$, that is, if $x(t)$ is even, then $x(t) \sin(2\pi ft)$ is odd in (2.66) and $\operatorname{Im} X(f) = 0$. Furthermore, $\operatorname{Re} X(f)$ is an even function of f because cosine is an even function. Thus, the Fourier transform of a real, even function is real and even.

On the other hand, if $x(t)$ is odd, $x(t) \cos 2\pi ft$ is odd in (2.65) and $\operatorname{Re} X(f) = 0$. Thus, the Fourier transform of a real, odd function is imaginary. In addition, $\operatorname{Im} X(f)$ is an odd function of frequency because $\sin 2\pi ft$ is an odd function.

EXAMPLE 2.8

Consider the pulse

$$x(t) = A\Pi\left(\frac{t-t_0}{\tau}\right) \quad (2.67)$$

The Fourier transform is

$$\begin{aligned} X(f) &= \int_{-\infty}^{\infty} A\Pi\left(\frac{t-t_0}{\tau}\right) e^{j2\pi ft} dt \\ &= A \int_{t_0-\tau/2}^{t_0+\tau/2} e^{-j2\pi ft} dt = A\tau \operatorname{sinc}(f\tau) e^{-j2\pi ft_0} \end{aligned} \quad (2.68)$$

The amplitude spectrum of $x(t)$ is

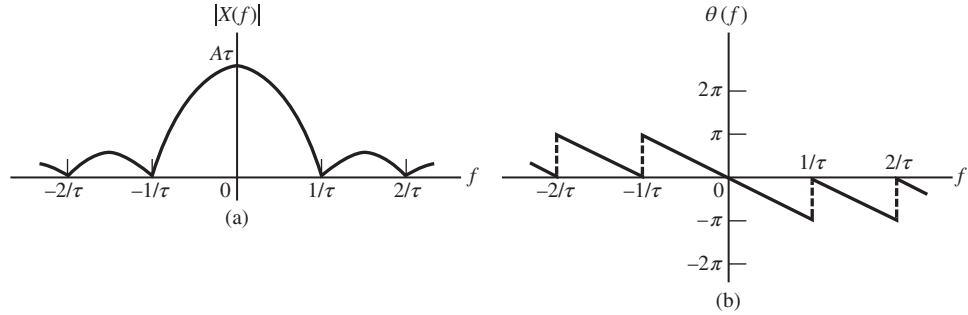
$$|X(f)| = A\tau |\operatorname{sinc}(f\tau)| \quad (2.69)$$

and the phase spectrum is

$$\theta(f) = \begin{cases} -2\pi t_0 f & \text{if } \operatorname{sinc}(f\tau) > 0 \\ -2\pi t_0 f \pm \pi & \text{if } \operatorname{sinc}(f\tau) < 0 \end{cases} \quad (2.70)$$

The term $\pm\pi$ is used to account for $\operatorname{sinc}(f\tau)$ being negative, and if $+\pi$ is used for $f > 0$, $-\pi$ is used for $f < 0$, or vice versa, to ensure that $\theta(f)$ is odd. When $|\theta(f)|$ exceeds 2π , an appropriate multiple

⁹Amplitude density spectrum would be more correct, since its dimensions are (amplitude units)(time) = (amplitude units)/(frequency), but we will use the term *amplitude spectrum* for simplicity.

**Figure 2.7**

Amplitude and phase spectra for a pulse signal. (a) Amplitude spectrum. (b) Phase spectrum ($t_0 = \frac{1}{2}\tau$ is assumed).

of 2π may be added or subtracted from $\theta(f)$. Figure 2.7 shows the amplitude and phase spectra for the signal (2.67). The similarity to Figure 2.6 is to be noted, especially the inverse relationship between spectral width and pulse duration.

2.4.3 Energy Spectral Density

The energy of a signal, defined by (2.22), can be expressed in the frequency domain as follows:

$$\begin{aligned} E &\triangleq \int_{-\infty}^{\infty} |x(t)|^2 dt \\ &= \int_{-\infty}^{\infty} x^*(t) \left[\int_{-\infty}^{\infty} X(f) e^{j2\pi ft} df \right] dt \end{aligned} \quad (2.71)$$

where $x(t)$ has been written in terms of its Fourier transform. Reversing the order of integration, we obtain

$$\begin{aligned} E &= \int_{-\infty}^{\infty} X(f) \left[\int_{-\infty}^{\infty} x^*(t) e^{j2\pi ft} dt \right] df \\ &= \int_{-\infty}^{\infty} X(f) \left[\int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt \right]^* df \\ &= \int_{-\infty}^{\infty} X(f) X^*(f) df \end{aligned}$$

or

$$E = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df \quad (2.72)$$

This is referred to as *Rayleigh's energy theorem* or Parseval's theorem for Fourier transforms.

Examining $|X(f)|^2$ and recalling the definition of $X(f)$ given by (2.60), we note that the former has the units of (volts-seconds) or, since we are considering power on a per-ohm basis, (watts-seconds)/hertz = joules/hertz. Thus, we see that $|X(f)|^2$ has the units of energy density, and we define the energy spectral density of a signal as

$$G(f) \triangleq |X(f)|^2 \quad (2.73)$$

By integrating $G(f)$ over all frequency, we obtain the signal's total energy.

EXAMPLE 2.9

Rayleigh's energy theorem (Parseval's theorem for Fourier transforms) is convenient for finding the energy in a signal whose square is not easily integrated in the time domain, or vice versa. For example, the signal

$$x(t) = 40 \operatorname{sinc}(20t) \longleftrightarrow X(f) = 2\Pi\left(\frac{f}{20}\right) \quad (2.74)$$

has energy density

$$G(f) = |X(f)|^2 = \left[2\Pi\left(\frac{f}{20}\right)\right]^2 = 4\Pi\left(\frac{f}{20}\right) \quad (2.75)$$

where $\Pi(f/20)$ need not be squared because it has amplitude 1 whenever it is nonzero. Using Rayleigh's energy theorem, we find that the energy in $x(t)$ is

$$E = \int_{-\infty}^{\infty} G(f) df = \int_{-10}^{10} 4 df = 80 \text{ J} \quad (2.76)$$

This checks with the result that is obtained by integrating $x^2(t)$ over all t using the definite integral $\int_{-\infty}^{\infty} \operatorname{sinc}^2(u) du = 1$.

The energy contained in the frequency interval $(0, W)$ can be found from the integral

$$\begin{aligned} E_W &= \int_{-W}^W G(f) df = 2 \int_0^W \left[2\Pi\left(\frac{f}{20}\right)\right]^2 df \\ &= \begin{cases} 8W, & W \leq 10 \\ 80, & W > 10 \end{cases} \end{aligned} \quad (2.77)$$

which follows because $\Pi\left(\frac{f}{20}\right) = 0, |f| > 10$. ■

2.4.4 Convolution

We digress somewhat from our consideration of the Fourier transform to define the convolution operation and illustrate it by example.

The convolution of two signals, $x_1(t)$ and $x_2(t)$, is a new function of time, $x(t)$, written symbolically in terms of x_1 and x_2 as

$$x(t) = x_1(t) * x_2(t) = \int_{-\infty}^{\infty} x_1(\lambda)x_2(t - \lambda) d\lambda \quad (2.78)$$

Note that t is a parameter as far as the integration is concerned. The integrand is formed from x_1 and x_2 by three operations: (1) time reversal to obtain $x_2(-\lambda)$, (2) time shifting to obtain $x_2(t - \lambda)$, and (3) multiplication of $x_1(\lambda)$ and $x_2(t - \lambda)$ to form the integrand. An example will illustrate the implementation of these operations to form $x_1 * x_2$. Note that the dependence on time is often suppressed.

EXAMPLE 2.10

Find the convolution of the two signals

$$x_1(t) = e^{-\alpha t}u(t) \text{ and } x_2(t) = e^{-\beta t}u(t), \quad \alpha > \beta > 0 \quad (2.79)$$

Solution

The steps involved in the convolution are illustrated in Figure 2.9 for $\alpha = 4$ and $\beta = 2$. Mathematically, we can form the integrand by direct substitution:

$$x(t) = x_1(t) * x_2(t) = \int_{-\infty}^{\infty} e^{-\alpha\lambda}u(\lambda)e^{-\beta(t-\lambda)}u(t-\lambda)d\lambda \quad (2.80)$$

But

$$u(\lambda)u(t-\lambda) = \begin{cases} 0, & \lambda < 0 \\ 1, & 0 < \lambda < t \\ 0, & \lambda > t \end{cases} \quad (2.81)$$

Thus,

$$x(t) = \begin{cases} 0, & t < 0 \\ \int_0^t e^{-\beta t} e^{-(\alpha-\beta)\lambda} d\lambda = \frac{1}{\alpha-\beta} (e^{-\beta t} - e^{-\alpha t}), & t \geq 0 \end{cases} \quad (2.82)$$

This result for $x(t)$ is also shown in Figure 2.8.

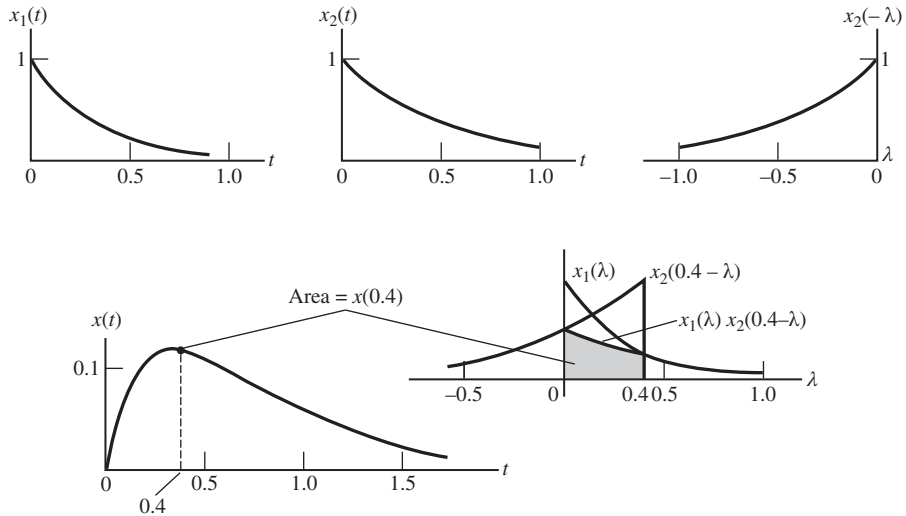


Figure 2.8

The operations involved in the convolution of two exponentially decaying signals.

2.4.5 Transform Theorems: Proofs and Applications

Several useful theorems¹⁰ involving Fourier transforms can be proved. These are useful for deriving Fourier-transform pairs as well as deducing general frequency-domain relationships. The notation $x(t) \longleftrightarrow X(f)$ will be used to denote a Fourier-transform pair.

Each theorem will be stated along with a proof in most cases. Several examples giving applications will be given after the statements of all the theorems. In the statements of the theorems $x(t)$, $x_1(t)$, and $x_2(t)$ denote signals with $X(f)$, $X_1(f)$, and $X_2(f)$ denoting their respective Fourier transforms. Constants are denoted by a , a_1 , a_2 , t_0 , and f_0 .

Superposition Theorem

$$a_1 x_1(t) + a_2 x_2(t) \longleftrightarrow a_1 X_1(f) + a_2 X_2(f) \quad (2.83)$$

Proof: By the defining integral for the Fourier transform,

$$\begin{aligned} \mathfrak{F}\{a_1 x_1(t) + a_2 x_2(t)\} &= \int_{-\infty}^{\infty} [a_1 x_1(t) + a_2 x_2(t)] e^{-j2\pi f t} dt \\ &= a_1 \int_{-\infty}^{\infty} x_1(t) e^{-j2\pi f t} dt + a_2 \int_{-\infty}^{\infty} x_2(t) e^{-j2\pi f t} dt \\ &= a_1 X_1(f) + a_2 X_2(f) \end{aligned} \quad (2.84)$$

Time-Delay Theorem

$$x(t - t_0) \longleftrightarrow X(f) e^{-j2\pi f t_0} \quad (2.85)$$

Proof: Using the defining integral for the Fourier transform, we have

$$\begin{aligned} \mathfrak{F}\{x(t - t_0)\} &= \int_{-\infty}^{\infty} x(t - t_0) e^{-j2\pi f t} dt \\ &= \int_{-\infty}^{\infty} x(\lambda) e^{-j2\pi f (\lambda + t_0)} d\lambda \\ &= e^{-j2\pi f t_0} \int_{-\infty}^{\infty} x(\lambda) e^{-j2\pi f \lambda} d\lambda \\ &= X(f) e^{-j2\pi f t_0} \end{aligned} \quad (2.86)$$

where the substitution $\lambda = t - t_0$ was used in the first integral.

Scale-Change Theorem

$$x(at) \longleftrightarrow \frac{1}{|a|} X\left(\frac{f}{a}\right) \quad (2.87)$$

¹⁰See Tables F.5 and F.6 in Appendix F for a listing of Fourier-transform pairs and theorems.

Proof: First, assume that $a > 0$. Then

$$\begin{aligned}\mathfrak{F}\{x(at)\} &= \int_{-\infty}^{\infty} x(at)e^{-j2\pi ft} dt \\ &= \int_{-\infty}^{\infty} x(\lambda)e^{-j2\pi f\lambda/a} \frac{d\lambda}{a} = \frac{1}{a} X\left(\frac{f}{a}\right)\end{aligned}\quad (2.88)$$

where the substitution $\lambda = at$ has been used. Next considering $a < 0$, we write

$$\begin{aligned}\mathfrak{F}\{x(at)\} &= \int_{-\infty}^{\infty} x(-|a|t)e^{-j2\pi ft} dt = \int_{-\infty}^{\infty} x(\lambda)e^{+j2\pi f\lambda/|a|} \frac{d\lambda}{|a|} \\ &= \frac{1}{|a|} X\left(-\frac{f}{|a|}\right) = \frac{1}{|a|} X\left(\frac{f}{a}\right)\end{aligned}\quad (2.89)$$

where use has been made of the relation $-|a| = a$ if $a < 0$.

Duality Theorem

$$X(t) \longleftrightarrow x(-f) \quad (2.90)$$

That is, if the Fourier transform of $x(t)$ is $X(f)$, then the Fourier transform of $X(f)$ with f replaced by t is the original time-domain signal with t replaced by $-f$.

Proof: The proof of this theorem follows by virtue of the fact that the only difference between the Fourier-transform integral and the inverse Fourier-transform integral is a minus sign in the exponent of the integrand.

Frequency-Translation Theorem

$$x(t)e^{j2\pi f_0 t} \longleftrightarrow X(f - f_0) \quad (2.91)$$

Proof: To prove the frequency-translation theorem, note that

$$\int_{-\infty}^{\infty} x(t)e^{j2\pi f_0 t} e^{-j2\pi ft} dt = \int_{-\infty}^{\infty} x(t)e^{-j2\pi(f-f_0)t} dt = X(f - f_0) \quad (2.92)$$

Modulation Theorem

$$x(t)\cos(2\pi f_0 t) \longleftrightarrow \frac{1}{2}X(f - f_0) + \frac{1}{2}X(f + f_0) \quad (2.93)$$

Proof: The proof of this theorem follows by writing $\cos(2\pi f_0 t)$ in exponential form as $\frac{1}{2}(e^{j2\pi f_0 t} + e^{-j2\pi f_0 t})$ and applying the superposition and frequency-translation theorems.

Differentiation Theorem

$$\frac{d^n x(t)}{dt^n} \longleftrightarrow (j2\pi f)^n X(f) \quad (2.94)$$

Proof: We prove the theorem for $n = 1$ by using integration by parts on the defining Fourier-transform integral as follows:

$$\begin{aligned}\mathfrak{F}\left\{\frac{dx}{dt}\right\} &= \int_{-\infty}^{\infty} \frac{dx(t)}{dt} e^{-j2\pi ft} dt \\ &= x(t)e^{-j2\pi ft} \Big|_{-\infty}^{\infty} + j2\pi f \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft} dt \\ &= j2\pi f X(f)\end{aligned}\quad (2.95)$$

where $u = e^{-j2\pi ft}$ and $dv = (dx/dt)dt$ have been used in the integration-by-parts formula, and the first term of the middle equation vanishes at each end point by virtue of $x(t)$ being an energy signal. The proof for values of $n > 1$ follows by induction.

Integration Theorem

$$\int_{-\infty}^t x(\lambda) d\lambda \leftrightarrow (j2\pi f)^{-1} X(f) + \frac{1}{2} X(0)\delta(f) \quad (2.96)$$

Proof: If $X(0) = 0$, the proof of the integration theorem can be carried out by using integration by parts as in the case of the differentiation theorem. We obtain

$$\begin{aligned}&\mathfrak{F}\left\{\int_{-\infty}^t x(\lambda) d(\lambda)\right\} \\ &= \left\{\int_{-\infty}^t x(\lambda) d(\lambda)\right\} \left(-\frac{1}{j2\pi f} e^{-j2\pi ft}\right) \Big|_{-\infty}^{\infty} + \frac{1}{j2\pi f} \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft} dt\end{aligned}\quad (2.97)$$

The first term vanishes if $X(0) = \int_{-\infty}^{\infty} x(t)dt = 0$, and the second term is just $X(f)/(j2\pi f)$. For $X(0) \neq 0$, a limiting argument must be used to account for the Fourier transform of the nonzero average value of $x(t)$.

Convolution Theorem

$$\begin{aligned}&\int_{-\infty}^{\infty} x_1(\lambda)x_2(t-\lambda) d\lambda \\ &\triangleq \int_{-\infty}^{\infty} x_1(t-\lambda)x_2(\lambda)d\lambda \leftrightarrow X_1(f)X_2(f)\end{aligned}\quad (2.98)$$

Proof: To prove the convolution theorem of Fourier transforms, we represent $x_2(t - \lambda)$ in terms of the inverse Fourier-transform integral as

$$x_2(t - \lambda) = \int_{-\infty}^{\infty} X_2(f)e^{j2\pi f(t-\lambda)} df \quad (2.99)$$

Denoting the convolution operation as $x_1(t) * x_2(t)$, we have

$$\begin{aligned}x_1(t) * x_2(t) &= \int_{-\infty}^{\infty} x_1(\lambda) \left[\int_{-\infty}^{\infty} X_2(f)e^{j2\pi f(t-\lambda)} df \right] d\lambda \\ &= \int_{-\infty}^{\infty} X_2(f) \left[\int_{-\infty}^{\infty} x_1(\lambda)e^{-j2\pi f\lambda} d\lambda \right] e^{j2\pi ft} df\end{aligned}\quad (2.100)$$

where the last step results from reversing the orders of integration. The bracketed term inside the integral is $X_1(f)$, the Fourier transform of $x_1(t)$. Thus,

$$x_1 * x_2 = \int_{-\infty}^{\infty} X_1(f)X_2(f)e^{j2\pi ft} df \quad (2.101)$$

which is the inverse Fourier transform of $X_1(f)X_2(f)$. Taking the Fourier transform of this result yields the desired transform pair.

Multiplication Theorem

$$x_1(t)x_2(t) \longleftrightarrow X_1(f) * X_2(f) = \int_{-\infty}^{\infty} X_1(\lambda)X_2(f - \lambda) d\lambda \quad (2.102)$$

Proof: The proof of the multiplication theorem proceeds in a manner analogous to the proof of the convolution theorem.

EXAMPLE 2.11

Use the duality theorem to show that

$$2AW \operatorname{sinc}(2Wt) \longleftrightarrow A\Pi\left(\frac{f}{2W}\right) \quad (2.103)$$

Solution

From Example 2.8, we know that

$$x(t) = A\Pi\left(\frac{t}{\tau}\right) \longleftrightarrow A\tau \operatorname{sinc} f\tau = X(f) \quad (2.104)$$

Considering $X(t)$, and using the duality theorem, we obtain

$$X(t) = A\tau \operatorname{sinc}(\tau t) \longleftrightarrow A\Pi\left(-\frac{f}{\tau}\right) = x(-f) \quad (2.105)$$

where τ is a parameter with dimension (s)⁻¹, which may be somewhat confusing at first sight! By letting $\tau = 2W$ and noting that $\Pi(u)$ is even, the given relationship follows. ■

EXAMPLE 2.12

Obtain the following Fourier-transform pairs:

1. $A\delta(t) \longleftrightarrow A$
2. $A\delta(t - t_0) \longleftrightarrow Ae^{-j2\pi ft_0}$
3. $A \longleftrightarrow A\delta(f)$
4. $Ae^{j2\pi f_0 t} \longleftrightarrow A\delta(f - f_0)$

Solution

Even though these signals are not energy signals, we can formally derive the Fourier transform of each by obtaining the Fourier transform of a “proper” energy signal that approaches the given signal in the limit as some parameter approaches zero or infinity. For example, formally,

$$\mathfrak{F}[A\delta(t)] = \mathfrak{F}\left[\lim_{\tau \rightarrow 0} \left(\frac{A}{\tau}\right) \Pi\left(\frac{t}{\tau}\right)\right] = \lim_{\tau \rightarrow 0} A \operatorname{sinc}(f\tau) = A \quad (2.106)$$

We can use a formal procedure such as this to define Fourier transforms for the other three signals as well. It is easier, however, to use the sifting property of the delta function and the appropriate Fourier-transform theorems. The same results are obtained. For example, we obtain the first transform pair directly by writing down the Fourier-transform integral with $x(t) = \delta(t)$ and invoking the sifting property:

$$\mathfrak{F}[A\delta(t)] = A \int_{-\infty}^{\infty} \delta(t) e^{-j2\pi f t} dt = A \quad (2.107)$$

Transform pair 2 follows by application of the time-delay theorem to pair 1.

Transform pair 3 can be obtained by using the inverse-transform relationship or the first transform pair and the duality theorem. Using the latter, we obtain

$$X(t) = A \longleftrightarrow A\delta(-f) = A\delta(f) = x(-f) \quad (2.108)$$

where the evenness property of the impulse function is used.

Transform pair 4 follows by applying the frequency-translation theorem to pair 3. The Fourier-transform pairs of Example 2.12 will be used often in the discussion of modulation. ■

EXAMPLE 2.13

Use the differentiation theorem to obtain the Fourier transform of the triangular signal, defined as

$$\Lambda\left(\frac{t}{\tau}\right) \triangleq \begin{cases} 1 - |t|/\tau, & |t| < \tau \\ 0, & \text{otherwise} \end{cases} \quad (2.109)$$

Solution

Differentiating $\Lambda(t/\tau)$ twice, we obtain, as shown in Figure 2.9

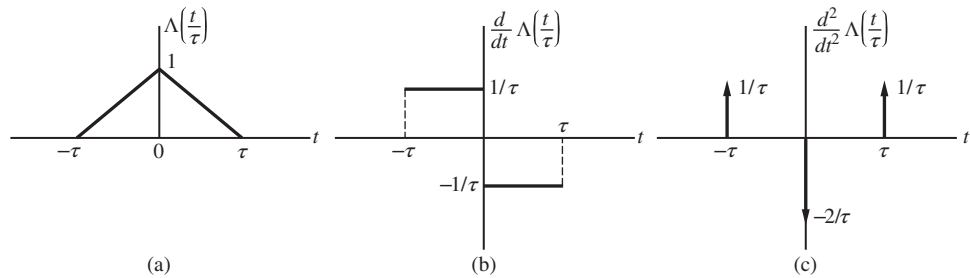
$$\frac{d^2\Lambda(t/\tau)}{dt^2} = \frac{1}{\tau}\delta(t + \tau) - \frac{2}{\tau}\delta(t) + \frac{1}{\tau}\delta(t - \tau) \quad (2.110)$$

Using the differentiation, superposition, and time-shift theorems and the result of Example 2.12, we obtain

$$\begin{aligned} \mathfrak{F}\left[\frac{d^2\Lambda(t/\tau)}{dt^2}\right] &= (j2\pi f)^2 \mathfrak{F}\left[\Lambda\left(\frac{t}{\tau}\right)\right] \\ &= \frac{1}{\tau}(e^{j2\pi f \tau} - 2 + e^{-j2\pi f \tau}) \end{aligned} \quad (2.111)$$

or, solving for $\mathfrak{F}\left[\Lambda\left(\frac{t}{\tau}\right)\right]$ and simplifying, we get

$$\mathfrak{F}\left[\Lambda\left(\frac{t}{\tau}\right)\right] = \frac{2 \cos 2\pi f \tau - 2}{\tau (j2\pi f)^2} = \tau \frac{\sin^2(\pi f \tau)}{(\pi f \tau)^2} \quad (2.112)$$

**Figure 2.9**

Triangular signal and its first two derivatives. (a) Triangular signal. (b) First derivative of the triangular signal. (c) Second derivative of the triangular signal.

where the identity $\frac{1}{2} [1 - \cos(2\pi ft)] = \sin^2(\pi ft)$ has been used. Summarizing, we have shown that

$$\Lambda\left(\frac{t}{\tau}\right) \longleftrightarrow \tau \operatorname{sinc}^2(f\tau) \quad (2.113)$$

where $[\sin(\pi f\tau)]/(\pi f\tau)$ has been replaced by $\operatorname{sinc}(f\tau)$.

EXAMPLE 2.14

As another example of obtaining Fourier transforms of signals involving impulses, let us consider the signal

$$y_s(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_s) \quad (2.114)$$

It is a periodic waveform referred to as the ideal sampling waveform and consists of a doubly infinite sequence of impulses spaced by T_s seconds.

Solution

To obtain the Fourier transform of $y_s(t)$, we note that it is periodic and, in a formal sense, therefore, can be represented by a Fourier series. Thus,

$$y_s(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_s) = \sum_{n=-\infty}^{\infty} Y_n e^{jn2\pi f_s t}, \quad f_s = \frac{1}{T_s} \quad (2.115)$$

where

$$Y_n = \frac{1}{T_s} \int_{T_s} \delta(t) e^{-jn2\pi f_s t} dt = f_s \quad (2.116)$$

by the sifting property of the impulse function. Therefore,

$$y_s(t) = f_s \sum_{n=-\infty}^{\infty} e^{jn2\pi f_s t} \quad (2.117)$$

Fourier-transforming term by term, we obtain

$$Y_s(f) = f_s \sum_{n=-\infty}^{\infty} \mathfrak{F}[1 \cdot e^{j2\pi n f_s t}] = f_s \sum_{n=-\infty}^{\infty} \delta(f - n f_s) \quad (2.118)$$

where we have used the results of Example 2.12. Summarizing, we have shown that

$$\sum_{m=-\infty}^{\infty} \delta(t - m T_s) \leftrightarrow f_s \sum_{n=-\infty}^{\infty} \delta(f - n f_s) \quad (2.119)$$

The transform pair (2.119) is useful in spectral representations of periodic signals by the Fourier transform, which will be considered shortly.

A useful expression can be derived from (2.119). Taking the Fourier transform of the left-hand side of (2.119) yields

$$\begin{aligned} \mathfrak{F} \left[\sum_{m=-\infty}^{\infty} \delta(t - m T_s) \right] &= \int_{-\infty}^{\infty} \left[\sum_{m=-\infty}^{\infty} \delta(t - m T_s) \right] e^{-j2\pi f t} dt \\ &= \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(t - m T_s) e^{-j2\pi f t} dt \\ &= \sum_{m=-\infty}^{\infty} e^{-j2\pi m T_s f} \end{aligned} \quad (2.120)$$

where we interchanged the orders of integration and summation and used the sifting property of the impulse function to perform the integration. Replacing m by $-m$ and equating the result to the right-hand side of (2.119) gives

$$\sum_{m=-\infty}^{\infty} e^{j2\pi m T_s f} = f_s \sum_{n=-\infty}^{\infty} \delta(f - n f_s) \quad (2.121)$$

This result will be used in Chapter 7. ■

EXAMPLE 2.15

The convolution theorem can be used to obtain the Fourier transform of the triangle $\Lambda(t/\tau)$ defined by (2.109).

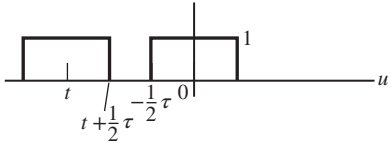
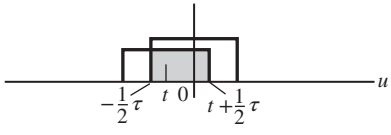
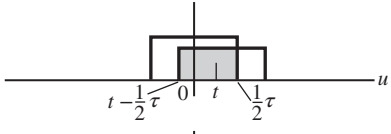
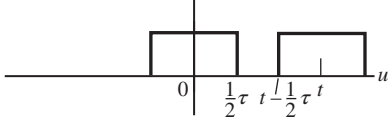
Solution

We proceed by first showing that the convolution of two rectangular pulses is a triangle. The steps in computing

$$y(t) = \int_{-\infty}^{\infty} \Pi\left(\frac{t-\lambda}{\tau}\right) \Pi\left(\frac{\lambda}{\tau}\right) d\lambda \quad (2.122)$$

are carried out in Table 2.2. Summarizing the results, we have

Table 2.2 Computation of $\Pi(t/\tau) * \Pi(t/\tau)$

Range	Integrand	Limits	Area
$-\infty < t < -\tau$			0
$-\tau < t < 0$		$-\frac{1}{2}\tau$ to $t + \frac{1}{2}\tau$	$\tau + t$
$0 < t < \tau$		$t - \frac{1}{2}\tau$ to $\frac{1}{2}\tau$	$\tau - t$
$\tau < t < \infty$			0

$$\tau \Lambda\left(\frac{t}{\tau}\right) = \Pi\left(\frac{t}{\tau}\right) * \Pi\left(\frac{t}{\tau}\right) = \begin{cases} 0, & t < -\tau \\ \tau - |t|, & |t| \leq \tau \\ 0, & t > \tau \end{cases} \quad (2.123)$$

$$\text{or } \Lambda\left(\frac{t}{\tau}\right) = \frac{1}{\tau} \Pi\left(\frac{t}{\tau}\right) * \Pi\left(\frac{t}{\tau}\right) \quad (2.124)$$

Using the transform pair

$$\Pi\left(\frac{t}{\tau}\right) \leftrightarrow \tau \operatorname{sinc} ft \quad (2.125)$$

and the convolution theorem of Fourier transforms (2.114), we obtain the transform pair

$$\Lambda\left(\frac{t}{\tau}\right) \leftrightarrow \tau \operatorname{sinc}^2 f\tau \quad (2.126)$$

as in Example 2.13 by applying the differentiation theorem. ■

A useful result is the convolution of an impulse $\delta(t - t_0)$ with a signal $x(t)$, where $x(t)$ is assume continuous at $t = t_0$. Carrying out the operation, we obtain

$$\delta(t - t_0) * x(t) = \int_{-\infty}^{\infty} \delta(\lambda - t_0)x(t - \lambda) d\lambda = x(t - t_0) \quad (2.127)$$

by the sifting property of the delta function. That is, convolution of $x(t)$ with an impulse occurring at time t_0 simply shifts $x(t)$ to t_0 .

EXAMPLE 2.16

Consider the Fourier transform of the cosinusoidal pulse

$$x(t) = A\Pi\left(\frac{t}{\tau}\right)\cos(\omega_0 t), \quad \omega_0 = 2\pi f_0 \quad (2.128)$$

Using the transform pair (see Example 2.12, Item 4)

$$e^{\pm j2\pi f_0 t} \leftrightarrow \delta(f \mp f_0) \quad (2.129)$$

obtained earlier and Euler's theorem, we find that

$$\cos(2\pi f_0 t) \leftrightarrow \frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0) \quad (2.130)$$

We have also shown that

$$A\Pi\left(\frac{t}{\tau}\right) \leftrightarrow A\tau \operatorname{sinc}(f\tau)$$

Therefore, using the multiplication theorem of Fourier transforms (2.118), we obtain

$$\begin{aligned} X(f) &= \mathfrak{F}\left[A\Pi\left(\frac{t}{\tau}\right)\cos(\omega_0 t)\right] = [A\tau \operatorname{sinc}(f\tau)] * \left\{\frac{1}{2}[\delta(f - f_0) + \delta(f + f_0)]\right\} \\ &= \frac{1}{2}A\tau \left\{\operatorname{sinc}[(f - f_0)\tau] + \operatorname{sinc}[(f + f_0)\tau]\right\} \end{aligned} \quad (2.131)$$

where $\delta(f - f_0) * Z(f) = Z(f - f_0)$ for $Z(f)$ continuous at $f = f_0$ has been used. Figure 2.10(c) shows $X(f)$. The same result can be obtained via the modulation theorem. ■

2.4.6 Fourier Transforms of Periodic Signals

The Fourier transform of a periodic signal, in a strict mathematical sense, does not exist, since periodic signals are not energy signals. However, using the transform pairs derived in Example 2.12 for a constant and a phasor signal, we could, in a formal sense, write down the Fourier transform of a periodic signal by Fourier-transforming its complex Fourier series term by term.

A somewhat more useful form for the Fourier transform of a periodic signal is obtained by applying the convolution theorem and the transform pair (2.119) for the ideal sampling waveform. To obtain it, consider the result of convolving the ideal sampling waveform with a pulse-type signal $p(t)$ to obtain a new signal $x(t)$, where $x(t)$ is a periodic power signal. This is apparent when one carries out the convolution with the aid of (2.127):

$$x(t) = \left[\sum_{m=-\infty}^{\infty} \delta(t - mT_s) \right] * p(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_s) * p(t) = \sum_{m=-\infty}^{\infty} p(t - mT_s) \quad (2.132)$$

Applying the convolution theorem and the Fourier-transform pair of (2.119), we find that the Fourier transform of $x(t)$ is

$$X(f) = \mathfrak{F}\left\{ \sum_{m=-\infty}^{\infty} \delta(t - mT_s) \right\} P(f)$$

$$\begin{aligned}
 &= \left[f_s \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] P(f) = f_s \sum_{n=-\infty}^{\infty} \delta(f - nf_s) P(f) \\
 &= \sum_{n=-\infty}^{\infty} f_s P(nf_s) \delta(f - nf_s)
 \end{aligned} \tag{2.133}$$

where $P(f) = \mathfrak{F}[p(t)]$ and the fact that $P(f) \delta(f - nf_s) = P(nf_s) \delta(f - nf_s)$ has been used. Summarizing, we have obtained the Fourier-transform pair

$$\sum_{m=-\infty}^{\infty} p(t - mT_s) \longleftrightarrow \sum_{n=-\infty}^{\infty} f_s P(nf_s) \delta(f - nf_s) \tag{2.134}$$

The usefulness of (2.134) is illustrated with an example.

EXAMPLE 2.17

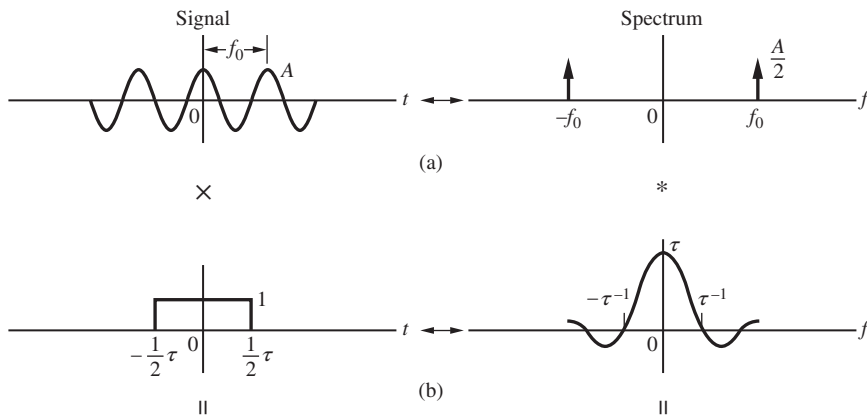
The Fourier transform of a single cosinusoidal pulse was found in Example 2.16 and is shown in Figure 2.10(c). The Fourier transform of a periodic cosinusoidal pulse train, which could represent the output of a radar transmitter, for example, is obtained by writing it as

$$\begin{aligned}
 y(t) &= \left[\sum_{n=-\infty}^{\infty} \delta(t - nT_s) \right] * \Pi\left(\frac{t}{\tau}\right) \cos(2\pi f_0 t), \quad f_0 \gg 1/\tau \\
 &= \sum_{m=-\infty}^{\infty} \Pi\left(\frac{t - mT_s}{\tau}\right) \cos[2\pi f_0(t - mT_s)], \quad f_s \leq \tau^{-1}
 \end{aligned} \tag{2.135}$$

This signal is illustrated in Figure 2.10(e). Identifying $p(t) = \Pi\left(\frac{t}{\tau}\right) \cos(2\pi f_0 t)$ we get, by the modulation theorem, that $P(f) = \frac{A\tau}{2} [\text{sinc}(f - f_0)\tau + \text{sinc}(f + f_0)\tau]$. Applying (2.134), the Fourier transform of $y(t)$ is

$$Y(f) = \sum_{n=-\infty}^{\infty} \frac{A f_s \tau}{2} [\text{sinc}(nf_s - f_0)\tau + \text{sinc}(nf_s + f_0)\tau] \delta(f - nf_s) \tag{2.136}$$

The spectrum is illustrated on the right-hand side of Figure 2.10(e).



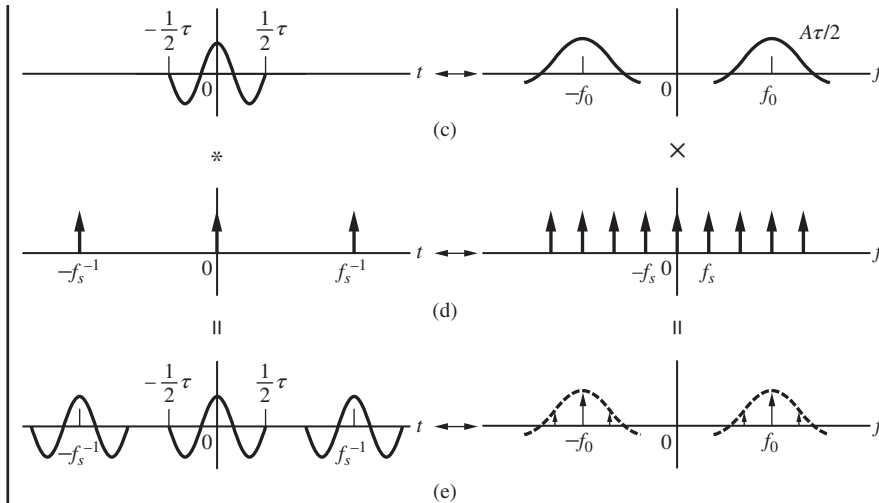


Figure 2.10

(a)–(c) Application of the multiplication theorem. (c)–(e) Application of the convolution theorem.

Note: \times denotes multiplication; \ast denotes convolution; \longleftrightarrow denotes transform pairs.

2.4.7 Poisson Sum Formula

We can develop the *Poisson sum formula* by taking the inverse Fourier transform of the right-hand side of (2.134). When we use the transform pair $\exp(-j2\pi n f_s t) \longleftrightarrow \delta(f - n f_s)$ (see Example 2.12), it follows that

$$\mathfrak{F}^{-1} \left\{ \sum_{n=-\infty}^{\infty} f_s P(n f_s) \delta(f - n f_s) \right\} = f_s \sum_{n=-\infty}^{\infty} P(n f_s) e^{j2\pi n f_s t} \quad (2.137)$$

Equating this to the left-hand side of (2.134), we obtain the Poisson sum formula:

$$\sum_{m=-\infty}^{\infty} p(t - m T_s) = f_s \sum_{n=-\infty}^{\infty} P(n f_s) e^{j2\pi n f_s t} \quad (2.138)$$

The Poisson sum formula is useful when one goes from the Fourier transform to sampled approximations of it. For example, Equation (2.138) says that the sample values $P(n f_s)$ of $P(f) = \mathfrak{F}\{p(t)\}$ are the Fourier series coefficients of the periodic function $T_s \sum_{n=-\infty}^{\infty} p(t - m T_s)$.

2.5 POWER SPECTRAL DENSITY AND CORRELATION

Recalling the definition of energy spectral density, Equation (2.73), we see that it is of use only for energy signals for which the integral of $G(f)$ over all frequencies gives total energy, a finite quantity. For power signals, it is meaningful to speak in terms of *power spectral density*. Analogous to $G(f)$, we define the power spectral density $S(f)$ of a signal $x(t)$ as a real, even,

nonnegative function of frequency, which gives total average power per ohm when integrated; that is,

$$P = \int_{-\infty}^{\infty} S(f) df = \langle x^2(t) \rangle \quad (2.139)$$

where $\langle x^2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x^2(t) dt$ denotes the time average of $x^2(t)$. Since $S(f)$ is a function that gives the variation of density of power with frequency, we conclude that it must consist of a series of impulses for the periodic power signals that we have so far considered. Later, in Chapter 7, we will consider power spectra of random signals.

EXAMPLE 2.18

Considering the cosinusoidal signal

$$x(t) = A \cos(2\pi f_0 t + \theta) \quad (2.140)$$

we note that its average power per ohm, $\frac{1}{2}A^2$, is concentrated at the single frequency f_0 hertz. However, since the power spectral density must be an even function of frequency, we split this power equally between $+f_0$ and $-f_0$ hertz. Thus, the power spectral density of $x(t)$ is, from intuition, given by

$$S(f) = \frac{1}{4}A^2\delta(f - f_0) + \frac{1}{4}A^2\delta(f + f_0) \quad (2.141)$$

Checking this by using (2.139), we see that integration over all frequencies results in the average power per ohm of $\frac{1}{2}A^2$. ■

2.5.1 The Time-Average Autocorrelation Function

To introduce the time-average autocorrelation function, we return to the energy spectral density of an energy signal, (2.73). Without any apparent reason, suppose we take the inverse Fourier transform of $G(f)$, letting the independent variable be τ :

$$\begin{aligned} \phi(\tau) &\triangleq \mathfrak{F}^{-1}[G(f)] = \mathfrak{F}^{-1}[X(f)X^*(f)] \\ &= \mathfrak{F}^{-1}[X(f)] * \mathfrak{F}^{-1}[X^*(f)] \end{aligned} \quad (2.142)$$

The last step follows by application of the convolution theorem. Applying the time-reversal theorem (Item 3b in Table F.6 in Appendix F) to write $\mathfrak{F}^{-1}[X^*(f)] = x(-\tau)$ and then the convolution theorem, we obtain

$$\begin{aligned} \phi(\tau) &= x(\tau) * x(-\tau) = \int_{-\infty}^{\infty} x(\lambda)x(\lambda + \tau) d\lambda \\ &= \lim_{T \rightarrow \infty} \int_{-T}^T x(\lambda)x(\lambda + \tau) d\lambda \quad (\text{energy signal}) \end{aligned} \quad (2.143)$$

Equation (2.143) will be referred to as the *time-average autocorrelation function* for energy signals. We see that it gives a measure of the similarity, or coherence, between a signal and a delayed version of the signal. Note that $\phi(0) = E$, the signal energy. Also note the similarity of the correlation operation to convolution. The major point of (2.142) is that the autocorrelation function and energy spectral density are Fourier-transform pairs. We forgo further discussion

of the time-average autocorrelation function for energy signals in favor of analogous results for power signals.

The time-average autocorrelation function $R(\tau)$ of a power signal $x(t)$ is defined as the time average

$$R(\tau) = \langle x(t)x(t+\tau) \rangle \\ \triangleq \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)x(t+\tau) dt \quad (\text{power signal}) \quad (2.144)$$

If $x(t)$ is periodic with period T_0 , the integrand of (2.144) is periodic, and the time average can be taken over a single period:

$$R(\tau) = \frac{1}{T_0} \int_{T_0} x(t)x(t+\tau) dt \quad [x(t) \text{ periodic}]$$

Just like $\phi(\tau)$, $R(\tau)$ gives a measure of the similarity between a power signal at time t and at time $t + \tau$; it is a function of the delay variable τ , since time, t , is the variable of integration. In addition to being a measure of the similarity between a signal and its time displacement, we note that the total average power of the signal is

$$R(0) = \langle x^2(t) \rangle = \int_{-\infty}^{\infty} S(f) df \quad (2.145)$$

Thus, we suspect that the time-average autocorrelation function and power spectral density of a power signal are closely related, just as they are for energy signals. This relationship is stated formally by the *Wiener–Khinchine theorem*, which says that the time-average autocorrelation function of a signal and its power spectral density are Fourier-transform pairs:

$$S(f) = \mathfrak{F}[R(\tau)] = \int_{-\infty}^{\infty} R(\tau) e^{-j2\pi f \tau} d\tau \quad (2.146)$$

and

$$R(\tau) = \mathfrak{F}^{-1}[S(f)] = \int_{-\infty}^{\infty} S(f) e^{j2\pi f \tau} df \quad (2.147)$$

A formal proof of the Wiener–Khinchine theorem will be given in Chapter 7. We simply take (2.146) as the definition of power spectral density at this point. We note that (2.145) follows immediately from (2.147) by setting $\tau = 0$.

2.5.2 Properties of $R(\tau)$

The time-average autocorrelation function has several useful properties, which are listed below:

1. $R(0) = \langle x^2(t) \rangle \geq |R(\tau)|$, for all τ ; that is, an absolute maximum of $R(\tau)$ exists at $\tau = 0$.
2. $R(-\tau) = \langle x(t)x(t-\tau) \rangle = R(\tau)$; that is, $R(\tau)$ is even.
3. $\lim_{|\tau| \rightarrow \infty} R(\tau) = \langle x(t) \rangle^2$ if $x(t)$ does not contain periodic components.
4. If $x(t)$ is periodic in t with period T_0 , then $R(\tau)$ is periodic in τ with period T_0 .
5. The time-average autocorrelation function of any power signal has a Fourier transform that is nonnegative.

Property 5 results by virtue of the fact that normalized power is a nonnegative quantity. These properties will be proved in Chapter 7.

The autocorrelation function and power spectral density are important tools for systems analysis involving random signals.

EXAMPLE 2.19

We desire the autocorrelation function and power spectral density of the signal $x(t) = \text{Re}[2 + 3 \exp(j10\pi t) + 4j \exp(j10\pi t)]$ or $x(t) = 2 + 3 \cos(10\pi t) - 4 \sin(10\pi t)$. The first step is to write the signal as a constant plus a single sinusoid. To do so, we note that

$$x(t) = \text{Re} \left[2 + \sqrt{3^2 + 4^2} \exp [j \tan^{-1} (4/3)] \exp (j10\pi t) \right] = 2 + 5 \cos [10\pi t + \tan^{-1} (4/3)]$$

We may proceed in one of two ways. The first is to find the autocorrelation function of $x(t)$ and Fourier-transform it to get the power spectral density. The second is to write down the power spectral density and inverse Fourier-transform it to get the autocorrelation function.

Following the first method, we find the autocorrelation function:

$$\begin{aligned} R(\tau) &= \frac{1}{T_0} \int_{T_0} x(t)x(t + \tau) dt \\ &= \frac{1}{0.2} \int_0^{0.2} \{2 + 5 \cos [10\pi t + \tan^{-1} (4/3)]\} \{2 + 5 \cos [10\pi (t + \tau) + \tan^{-1} (4/3)]\} dt \\ &= 5 \int_0^{0.2} \left\{ 4 + 10 \cos [10\pi t + \tan^{-1} (4/3)] + 10 \cos [10\pi (t + \tau) + \tan^{-1} (4/3)] \right. \\ &\quad \left. + 25 \cos [10\pi t + \tan^{-1} (4/3)] \cos [10\pi (t + \tau) + \tan^{-1} (4/3)] \right\} dt \\ &= 5 \int_0^{0.2} 4 dt + 50 \int_0^{0.2} \cos [10\pi t + \tan^{-1} (4/3)] dt \\ &\quad + 50 \int_0^{0.2} \cos [10\pi (t + \tau) + \tan^{-1} (4/3)] dt \\ &\quad + \frac{125}{2} \int_0^{0.2} \cos (10\pi \tau) dt + \frac{125}{2} \int_0^{0.2} \cos [20\pi t + 10\pi \tau + 2 \tan^{-1} (4/3)] dt \\ &= 5 \int_0^{0.2} 4 dt + 0 + 0 + \frac{125}{2} \int_0^{0.2} \cos (10\pi \tau) dt \\ &\quad + \frac{125}{2} \int_0^{0.2} \cos [20\pi t + 10\pi \tau + 2 \tan^{-1} (4/3)] dt \\ &= 4 + \frac{25}{2} \cos (10\pi \tau) \end{aligned} \tag{2.148}$$

where integrals involving cosines of t are zero by virtue of integrating a cosine over an integer number of periods, and the trigonometric relationship $\cos x \cos y = \frac{1}{2} \cos (x + y) + \frac{1}{2} \cos (x - y)$ has been used. The power spectral density is the Fourier transform of the autocorrelation function, or

$$S(f) = \mathfrak{F} \left[4 + \frac{25}{2} \cos (10\pi \tau) \right]$$

$$\begin{aligned}
 &= 4\mathfrak{F}[1] + \frac{25}{2}\mathfrak{F}[\cos(10\pi\tau)] \\
 &= 4\delta(f) + \frac{25}{4}\delta(f-5) + \frac{25}{4}\delta(f+5)
 \end{aligned} \tag{2.149}$$

Note that integration of this over all f gives $P = 4 + \frac{25}{2} = 16.5$ watts/ohm, which is the DC power plus the AC power (the latter is split between 5 and -5 hertz). We could have proceeded by writing down the power spectral density first, using power arguments, and inverse Fourier-transforming it to get the autocorrelation function.

Note that all properties of the autocorrelation function are satisfied except the third, which does not apply. ■

EXAMPLE 2.20

The sequence 1110010 is an example of a *pseudonoise* or *m*-sequence; they are important in the implementation of digital communication systems and will be discussed further in Chapter 9. For now, we use this *m*-sequence as another illustration for computing autocorrelation functions and power spectra. Consider Figure 2.11(a), which shows the waveform equivalent of this *m*-sequence obtained by replacing each 0 by -1 , multiplying each sequence member by a square pulse function $\Pi\left(\frac{t-t_0}{\Delta}\right)$, summing, and assuming the resulting waveform is repeated forever thereby making it periodic. To compute the autocorrelation function, we apply (2.145), which is

$$R(\tau) = \frac{1}{T_0} \int_{T_0} x(t)x(t+\tau) dt$$

since a periodic repetition of the waveform is assumed. Consider the waveform $x(t)$ multiplied by $x(t+n\Delta)$ [shown in Figure 2.11(b) for $n=2$]. The product is shown in Figure 2.11(c), where it is seen that the net area under the product $x(t)x(t+n\Delta)$ is $-\Delta$, which gives $R(2\Delta) = -\frac{\Delta}{7\Delta} = -\frac{1}{7}$ for this case. In fact, this answer results for any τ equal to a nonzero integer multiple of Δ . For $\tau=0$, the net area under the product $x(t)x(t+0)$ is 7Δ , which gives $R(0) = \frac{7\Delta}{7\Delta} = 1$. These correlation results are shown in Figure 2.11(d) by the open circles where it is noted that they repeat each $\tau=7\Delta$. For a given noninteger delay value, the autocorrelation function is obtained as the linear interpolation of the autocorrelation function values for the integer delays bracketing the desired delay value. One can see that this is the case by considering the integral $\int_{T_0} x(t)x(t+\tau) dt$ and noting that the area under the product $x(t)x(t+\tau)$ must be a linear function of τ due to $x(t)$ being composed of square pulses. Thus, the autocorrelation function is as shown in Figure 2.11(d) by the solid line. For one period, it can be expressed as

$$R(\tau) = \frac{8}{7}\Lambda\left(\frac{\tau}{\Delta}\right) - \frac{1}{7}, \quad |\tau| \leq \frac{T_0}{2}$$

The power spectral density is the Fourier transform of the autocorrelation function, which can be obtained by applying (2.146). The detailed derivation of it is left to the problems. The result is

$$S(f) = \frac{8}{49} \sum_{n=-\infty}^{\infty} \text{sinc}^2\left(\frac{n}{7\Delta}\right) \delta\left(f - \frac{n}{7\Delta}\right) - \frac{1}{7}\delta(f)$$

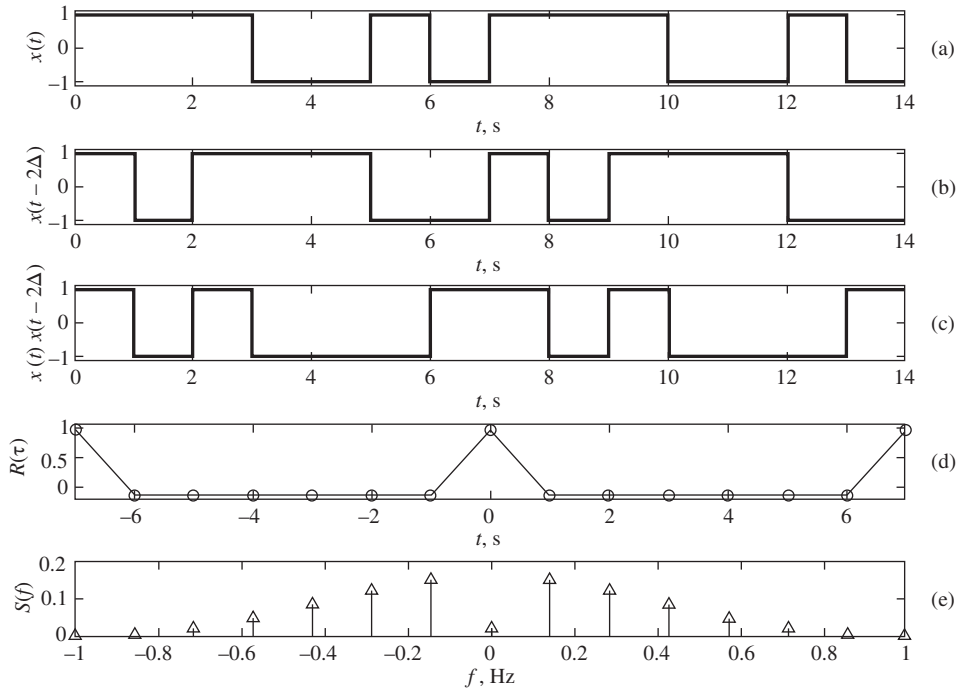


Figure 2.11

Waveforms pertinent to computing the autocorrelation function and power spectrum of an m -sequence of length 7.

and is shown in Figure 2.11(e). Note that near $f = 0$, $S(f) = \left(\frac{8}{49} - \frac{1}{7}\right) \delta(f) = \frac{1}{49} \delta(f)$, which says that the DC power is $\frac{1}{49} = \frac{1}{7^2}$ watts. The student should think about why this is the correct result. (*Hint*: What is the DC value of $x(t)$ and to what power does this correspond?)

The autocorrelation function and power spectral density are important tools for systems analysis involving random signals.

2.6 SIGNALS AND LINEAR SYSTEMS

In this section we are concerned with the characterization of systems and their effects on signals. In system modeling, the actual elements, such as resistors, capacitors, inductors, springs, and masses, that compose a particular system are usually not of concern. Rather, we view a system in terms of the operation it performs on an input to produce an output. Symbolically, this is accomplished, for a single-input, single-output system, by writing

$$y(t) = \mathcal{H}[x(t)] \quad (2.150)$$

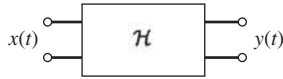


Figure 2.12
Operator representation of a linear system.

where $\mathcal{H}[\cdot]$ is the operator that produces the output $y(t)$ from the input $x(t)$, as illustrated in Figure 2.12. We now consider certain classes of systems, the first of which is linear time-invariant systems.

2.6.1 Definition of a Linear Time-Invariant System

If a system is linear, superposition holds. That is, if $x_1(t)$ results in the output $y_1(t)$ and $x_2(t)$ results in the output $y_2(t)$, then the output due to $\alpha_1 x_1(t) + \alpha_2 x_2(t)$, where α_1 and α_2 are constants, is given by

$$\begin{aligned} y(t) &= \mathcal{H}[\alpha_1 x_1(t) + \alpha_2 x_2(t)] = \alpha_1 \mathcal{H}[x_1(t)] + \alpha_2 \mathcal{H}[x_2(t)] \\ &= \alpha_1 y_1(t) + \alpha_2 y_2(t) \end{aligned} \quad (2.151)$$

If the system is *time-invariant*, or *fixed*, the delayed input $x(t - t_0)$ gives the delayed output $y(t - t_0)$; that is,

$$y(t - t_0) = \mathcal{H}[x(t - t_0)] \quad (2.152)$$

With these properties explicitly stated, we are now ready to obtain more concrete descriptions of linear time-invariant (LTI) systems.

2.6.2 Impulse Response and the Superposition Integral

The *impulse response* $h(t)$ of an LTI system is defined to be the response of the system to an impulse applied at $t = 0$, that is

$$h(t) \triangleq \mathcal{H}[\delta(t)] \quad (2.153)$$

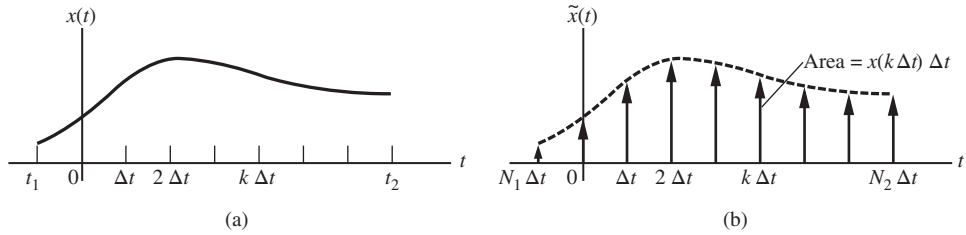
By the time-invariant property of the system, the response to an impulse applied at any time t_0 is $h(t - t_0)$, and the response to the linear combination of impulses $\alpha_1 \delta(t - t_1) + \alpha_2 \delta(t - t_2)$ is $\alpha_1 h(t - t_1) + \alpha_2 h(t - t_2)$ by the superposition property and time-invariance. Through induction, we may therefore show that the response to the input

$$x(t) = \sum_{n=1}^N \alpha_n \delta(t - t_n) \quad (2.154)$$

is

$$y(t) = \sum_{n=1}^N \alpha_n h(t - t_n) \quad (2.155)$$

We will use (2.155) to obtain the superposition integral, which expresses the response of an LTI system to an arbitrary input (with suitable restrictions) in terms of the impulse response

**Figure 2.13**

A signal and an approximate representation. (a) Signal. (b) Approximation with a sequence of impulses.

of the system. Considering the arbitrary input signal $x(t)$ of Figure 2.13(a), we can represent it as

$$x(t) = \int_{-\infty}^{\infty} x(\lambda) \delta(t - \lambda) d\lambda \quad (2.156)$$

by the sifting property of the unit impulse. Approximating the integral of (2.156) as a sum, we obtain

$$x(t) \cong \sum_{n=N_1}^{N_2} x(n \Delta t) \delta(t - n \Delta t) \Delta t, \quad \Delta t \ll 1 \quad (2.157)$$

where $t_1 = N_1 \Delta t$ is the starting time of the signal and $t_2 = N_2 \Delta t$ is the ending time. The output, using (2.155) with $\alpha_n = x(n \Delta t) \Delta t$ and $t_n = n \Delta t$, is

$$\tilde{y}(t) = \sum_{n=N_1}^{N_2} x(n \Delta t) h(t - n \Delta t) \Delta t \quad (2.158)$$

where the tilde denotes the output resulting from the approximation to the input given by (2.157). In the limit as Δt approaches zero and $n \Delta t$ approaches the continuous variable λ , the sum becomes an integral, and we obtain

$$y(t) = \int_{-\infty}^{\infty} x(\lambda) h(t - \lambda) d\lambda \quad (2.159)$$

where the limits have been changed to $\pm\infty$ to allow arbitrary starting and ending times for $x(t)$. Making the substitution $\sigma = t - \lambda$, we obtain the equivalent result

$$y(t) = \int_{-\infty}^{\infty} x(t - \sigma) h(\sigma) d\sigma \quad (2.160)$$

Because these equations were obtained by superposition of a number of elementary responses due to each individual impulse, they are referred to as superposition integrals. A simplification results if the system under consideration is causal, that is, a system that does not respond before an input is applied. For a causal system, $h(t - \lambda) = 0$ for $t < \lambda$, and the upper limit on (2.159) can be set equal to t . Furthermore, if $x(t) = 0$ for $t < 0$, the lower limit becomes zero.

2.6.3 Stability

A fixed, linear system is bounded-input, bounded-output (BIBO) stable if every bounded input results in a bounded output. It can be shown¹¹ that a system is BIBO stable if and only if

$$\int_{-\infty}^{\infty} |h(t)| dt < \infty \quad (2.161)$$

2.6.4 Transfer (Frequency Response) Function

Applying the convolution theorem of Fourier transforms, item 8 of Table F.6 in Appendix F, to either (2.159) or (2.160), we obtain

$$Y(f) = H(f)X(f) \quad (2.162)$$

where $X(f) = \mathfrak{F}\{x(t)\}$, $Y(f) = \mathfrak{F}\{y(t)\}$, and

$$H(f) = \mathfrak{F}\{h(t)\} = \int_{-\infty}^{\infty} h(t)e^{-j2\pi ft} dt \quad (2.163)$$

or

$$h(t) = \mathfrak{F}^{-1}\{H(f)\} = \int_{-\infty}^{\infty} H(f)e^{j2\pi ft} df \quad (2.164)$$

$H(f)$ is referred to as the *transfer (frequency response) function* of the system. We see that either $h(t)$ or $H(f)$ is an equally good characterization of the system. By an inverse Fourier transform on (2.162), the output becomes

$$y(t) = \int_{-\infty}^{\infty} X(f)H(f)e^{j2\pi ft} df \quad (2.165)$$

2.6.5 Causality

A system is causal if it does not anticipate its input. In terms of the impulse response, it follows that for a time-invariant causal system,

$$h(t) = 0, \quad t < 0 \quad (2.166)$$

When causality is viewed from the standpoint of the frequency response function of the system, a celebrated theorem by Wiener and Paley¹² states that if

$$\int_{-\infty}^{\infty} |h(t)|^2 dt = \int_{-\infty}^{\infty} |H(f)|^2 df < \infty \quad (2.167)$$

with $h(t) \equiv 0$ for $t < 0$, it is then necessary that

$$\int_{-\infty}^{\infty} \frac{|\ln |H(f)||}{1 + f^2} df < \infty \quad (2.168)$$

Conversely, if $|H(f)|$ is square-integrable, and if the integral in (2.168) is unbounded, then we cannot make $h(t) \equiv 0, t < 0$, no matter what we choose for $\underline{H(f)}$. Consequences of

¹¹See Ziemer, Tranter, and Fannin (1998), Chapter 2.

¹²See William Siebert, *Circuits, Signals, and Systems*, New York: McGraw Hill, 1986, p. 476.

(2.168) are that no causal filter can have $|H(f)| \equiv 0$ over a finite band of frequencies (i.e., a causal filter cannot perfectly reject any finite band of frequencies). In fact, the Paley–Wiener criterion restricts the rate at which $|H(f)|$ for a causal LTI system can vanish. For example,

$$|H_1(f)| = e^{-k_1|f|} \Rightarrow |\ln |H_1(f)|| = k_1|f| \quad (2.169)$$

and

$$|H_2(f)| = e^{-k_2f^2} \Rightarrow |\ln |H_2(f)|| = k_2f^2 \quad (2.170)$$

where k_1 and k_2 are positive constants, are not allowable amplitude responses for causal LTI filters because (2.168) does not give a finite result in either case.

The sufficiency statement of the Paley–Wiener criterion is stated as follows: Given any square-integrable function $|H(f)|$ for which (2.168) is satisfied, there exists an $\angle H(f)$ such that $H(f) = |H(f)| \exp \left[j \angle H(f) \right]$ is the Fourier transform of $h(t)$ for a causal filter.

EXAMPLE 2.21

(a) Show that the system with impulse response

$$h(t) = e^{-2t} \cos(10\pi t) u(t)$$

is stable.

(b) Is it causal?

Solution

(a) We consider the integral

$$\begin{aligned} \int_{-\infty}^{\infty} |h(t)| dt &= \int_{-\infty}^{\infty} \left| e^{-2t} \cos(10\pi t) u(t) \right| dt \\ &= \int_0^{\infty} e^{-2t} |\cos(10\pi t)| dt \\ &\leq \int_0^{\infty} e^{-2t} dt = -\frac{1}{2} e^{-2t} \Big|_0^{\infty} = \frac{1}{2} < \infty \end{aligned}$$

Therefore, it is BIBO stable. Note that the third line follows from the second line by virtue of $|\cos(10\pi t)| \leq 1$.

(b) The system is causal because $h(t) = 0$ for $t < 0$. ■

2.6.6 Symmetry Properties of $H(f)$

The frequency response function, $H(f)$, of an LTI system is, in general, a complex quantity. We therefore write it in terms of magnitude and argument as

$$H(f) = |H(f)| \exp \left[j \angle H(f) \right] \quad (2.171)$$

where $|H(f)|$ is called the *amplitude- (magnitude) response function* and $\angle H(f)$ is called the *phase response function* of the LTI system. Also, $H(f)$ is the Fourier transform of a real-time

function $h(t)$. Therefore, it follows that

$$|H(f)| = |H(-f)| \quad (2.172)$$

and

$$\angle H(f) = -\angle H(-f) \quad (2.173)$$

That is, the amplitude response of a system with real-valued impulse response is an even function of frequency and its phase response is an odd function of frequency.

EXAMPLE 2.22

Consider the lowpass RC filter shown in Figure 2.14. We may find its frequency response function by a number of methods. First, we may write down the governing differential equation (integral-differential equation, in general) as

$$RC \frac{dy(t)}{dt} + y(t) = x(t) \quad (2.174)$$

and Fourier-transform it, obtaining

$$(j2\pi f RC + 1)Y(f) = X(f)$$

or

$$\begin{aligned} H(f) &= \frac{Y(f)}{X(f)} = \frac{1}{1 + j(f/f_3)} \\ &= \frac{1}{\sqrt{1 + (f/f_3)^2}} e^{-j \tan^{-1}(f/f_3)} \end{aligned} \quad (2.175)$$

where $f_3 = 1/(2\pi RC)$ is the 3-dB frequency, or half-power frequency. Second, we can use Laplace transform theory with s replaced by $j2\pi f$. Third, we can use AC sinusoidal steady-state analysis. The amplitude and phase responses of this system are illustrated in Figures 2.15(a) and (b), respectively.

Using the Fourier-transform pair

$$\alpha e^{-\alpha t} u(t) \longleftrightarrow \frac{\alpha}{\alpha + j2\pi f} \quad (2.176)$$

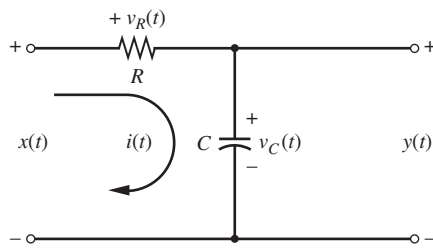
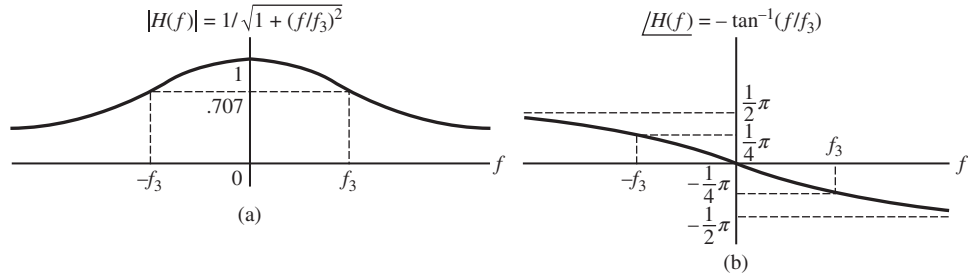


Figure 2.14
An RC lowpass filter.

**Figure 2.15**

Amplitude and phase responses of the RC lowpass filter. (a) Amplitude response. (b) Phase response.

we find the impulse response of the filter to be

$$h(t) = \frac{1}{RC} e^{-t/RC} u(t) \quad (2.177)$$

Finally, we consider the response of the filter to the pulse

$$x(t) = A \Pi \left(\frac{t - \frac{1}{2}T}{T} \right) \quad (2.178)$$

Using appropriate Fourier-transform pairs, we can readily find $Y(f)$, but its inverse Fourier transformation requires some effort. Thus, it appears that the superposition integral is the best approach in this case. Choosing the form

$$y(t) = \int_{-\infty}^{\infty} h(t - \sigma) x(\sigma) d\sigma \quad (2.179)$$

we find, by direct substitution in $h(t)$, that

$$h(t - \sigma) = \frac{1}{RC} e^{-(t-\sigma)/RC} u(t - \sigma) = \begin{cases} \frac{1}{RC} e^{-(t-\sigma)/RC}, & \sigma < t \\ 0, & \sigma > t \end{cases} \quad (2.180)$$

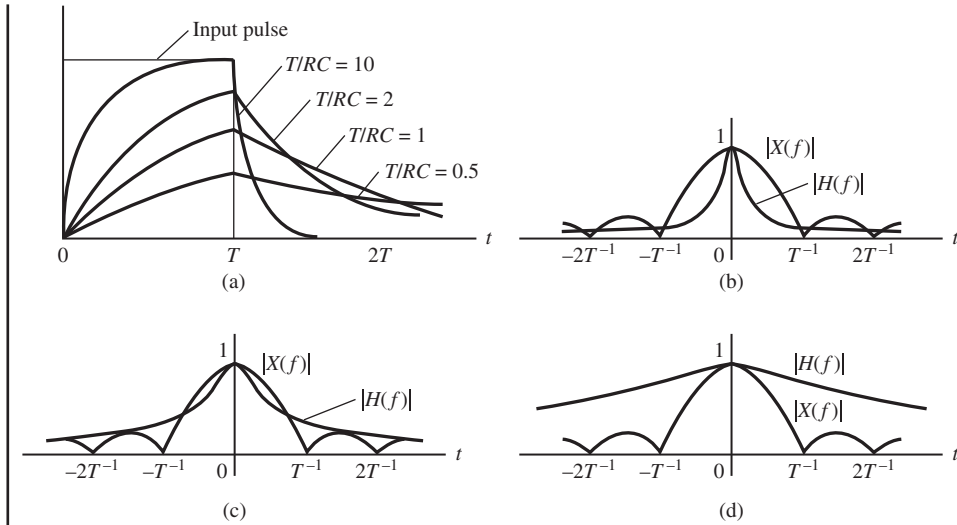
Since $x(\sigma)$ is zero for $\sigma < 0$ and $\sigma > T$, we find that

$$y(t) = \begin{cases} 0, & t < 0 \\ \int_0^t \frac{A}{RC} e^{-(t-\sigma)/RC} d\sigma, & 0 \leq t \leq T \\ \int_0^T \frac{A}{RC} e^{-(t-\sigma)/RC} d\sigma, & t > T \end{cases} \quad (2.181)$$

Carrying out the integrations, we obtain

$$y(t) = \begin{cases} 0, & t < 0 \\ A (1 - e^{-t/RC}), & 0 < t < T \\ A (e^{-(t-T)/RC} - e^{-t/RC}), & t > T \end{cases} \quad (2.182)$$

This result is plotted in Figure 2.16 for several values of T/RC . Also shown are $|X(f)|$ and $|H(f)|$. Note that $T/RC = 2\pi f_3/T^{-1}$ is proportional to the ratio of the 3-dB frequency of the filter to the spectral width (T^{-1}) of the pulse. When this ratio is large, the spectrum of the input pulse is essentially passed undistorted by the system, and the output looks like the input. On the other hand, for $2\pi f_3/T^{-1} \ll 1$, the system distorts the input signal spectrum, and $y(t)$ looks nothing like the input. These ideas will be put on a firmer basis when signal distortion is discussed.

**Figure 2.16**

(a) Waveforms and (b)–(d) spectra for a lowpass RC filter with pulse input. (a) Input and output signals. (b) $T/RC = 0.5$. (c) $T/RC = 2$. (d) $T/RC = 10$.

Note that the output could have been found by writing the input as $x(t) = A[u(t) - u(t - T)]$ and using superposition to write the output in terms of the step response as $y(t) = y_s(t) - y_s(t - T)$. The student may show that the step response is $y_s(t) = A(1 - e^{-t/RC})u(t)$. Thus, the output is $y(t) = A(1 - e^{-t/RC})u(t) - A(1 - e^{-(t-T)/RC})u(t - T)$, which can be shown to be equivalent to the result obtained above in (2.182). ■

2.6.7 Input-Output Relationships for Spectral Densities

Consider a fixed linear two-port system with frequency response function $H(f)$, input $x(t)$, and output $y(t)$. If $x(t)$ and $y(t)$ are energy signals, their energy spectral densities are $G_x(f) = |X(f)|^2$ and $G_y(f) = |Y(f)|^2$, respectively. Since $Y(f) = H(f)X(f)$, it follows that

$$G_y(f) = |H(f)|^2 G_x(f) \quad (2.183)$$

A similar relationship holds for power signals and spectra:

$$S_y(f) = |H(f)|^2 S_x(f) \quad (2.184)$$

This will be proved in Chapter 7.

2.6.8 Response to Periodic Inputs

Consider the steady-state response of a fixed linear system to the complex exponential input signal $Ae^{j2\pi f_0 t}$. Using the superposition integral, we obtain

$$y_{ss}(t) = \int_{-\infty}^{\infty} h(\lambda) A e^{j2\pi f_0(t-\lambda)} d\lambda$$

$$\begin{aligned}
&= Ae^{j2\pi f_0 t} \int_{-\infty}^{\infty} h(\lambda) e^{-j2\pi f_0 \lambda} d\lambda \\
&= H(f_0) Ae^{j2\pi f_0 t} \quad (2.185)
\end{aligned}$$

That is, the output is a complex exponential signal of the same frequency but with amplitude scaled by $|H(f_0)|$ and phase shifted by $\angle H(f_0)$ relative to the amplitude and phase of the input. Using superposition, we conclude that the steady-state output due to an arbitrary periodic input is represented by the complex exponential Fourier series

$$y(t) = \sum_{n=-\infty}^{\infty} X_n H(nf_0) e^{jn2\pi f_0 t} \quad (2.186)$$

or

$$y(t) = \sum_{n=-\infty}^{\infty} |X_n| |H(nf_0)| \exp \left\{ j \left[2\pi n f_0 t + \angle X_n + \angle H(nf_0) \right] \right\} \quad (2.187)$$

$$= X_0 H(0) + 2 \sum_{n=1}^{\infty} |X_n| |H(nf_0)| \cos \left[2\pi n f_0 t + \angle X_n + \angle H(nf_0) \right] \quad (2.188)$$

where (2.172) and (2.173) have been used to get the second equation. Thus, for a periodic input, the magnitude of each spectral component of the input is attenuated (or amplified) by the amplitude-response function *at the frequency of the particular spectral component*, and the phase of each spectral component is shifted by the value of the phase-shift function of the system at the *frequency of the particular spectral component*.

EXAMPLE 2.23

Consider the response of a filter having the frequency response function

$$H(f) = 2\Pi \left(\frac{f}{42} \right) e^{-j\pi f/10} \quad (2.189)$$

to a unit-amplitude triangular signal with period 0.1 s. From Table 2.1 and Equation (2.29), the exponential Fourier series of the input signal is

$$\begin{aligned}
x(t) &= \dots \frac{4}{25\pi^2} e^{-j100\pi t} + \frac{4}{9\pi^2} e^{-j60\pi t} + \frac{4}{\pi^2} e^{-j20\pi t} \\
&\quad + \frac{4}{\pi^2} e^{j20\pi t} + \frac{4}{9\pi^2} e^{j60\pi t} + \frac{4}{25\pi^2} e^{j100\pi t} + \dots \\
&= \frac{8}{\pi^2} \left[\cos(20\pi t) + \frac{1}{9} \cos(60\pi t) + \frac{1}{25} \cos(100\pi t) + \dots \right] \quad (2.190)
\end{aligned}$$

The filter eliminates all harmonics above 21 Hz and passes all those below 21 Hz, imposing an amplitude scale factor of 2 and a phase shift of $-\pi f/10$ rad. The only harmonic of the triangular wave to be passed by the filter is the fundamental, which has a frequency of 10 Hz, giving a phase shift of $-\pi(10)/10 = -\pi$ rad.

The output is therefore

$$y(t) = \frac{16}{\pi^2} \cos 20\pi \left(t - \frac{1}{20} \right) \quad (2.191)$$

where the phase shift is seen to be equivalent to a delay of $\frac{1}{20}$ s. ■

2.6.9 Distortionless Transmission

Equation (2.188) shows that both the amplitudes and phases of the spectral components of a periodic input signal will, in general, be altered as the signal is sent through a two-port LTI system. This modification may be desirable in signal *processing* applications, but it amounts to distortion in signal *transmission* applications. While it may appear at first that ideal signal transmission results only if there is *no* attenuation and phase shift of the spectral components of the input, this requirement is too stringent. A system will be classified as distortionless if it introduces the same attenuation and time delay to all spectral components of the input, for then the output looks like the input. In particular, if the output of a system is given in terms of the input as

$$y(t) = H_0 x(t - t_0) \quad (2.192)$$

where H_0 and t_0 are constants, the output is a scaled, delayed replica of the input ($t_0 > 0$ for causality). Employing the time-delay theorem to Fourier transform (2.192) and using the definition $H(f) = Y(f)/X(f)$, we obtain

$$H(f) = H_0 e^{-j2\pi f t_0} \quad (2.193)$$

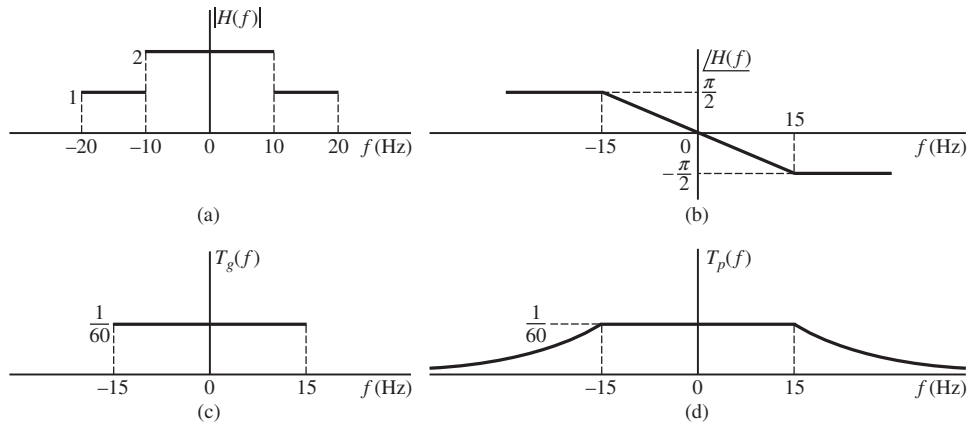
as the frequency response function of a distortionless system; that is, the amplitude response of a distortionless system is constant and the phase shift is linear with frequency. Of course, these restrictions are necessary only within the frequency ranges where the input has significant spectral content. Figure 2.17 and Example 2.24, considered shortly, will illustrate these comments.

In general, we can isolate three major types of distortion. First, if the system is linear but the amplitude response is not constant with frequency, the system is said to introduce *amplitude distortion*. Second, if the system is linear but the phase shift is not a linear function of frequency, the system introduces *phase, or delay, distortion*. Third, if the system is not linear, we have *nonlinear distortion*. Of course, these three types of distortion may occur in combination with one another.

2.6.10 Group and Phase Delay

One can often identify phase distortion in a linear system by considering the derivative of phase with respect to frequency. A distortionless system exhibits a phase response in which phase is directly proportional to frequency. Thus, the derivative of the phase response function with respect to frequency of a distortionless system is a constant. The negative of this constant is called the group delay of the LTI system. In other words, the group delay is defined by the equation

$$T_g(f) = -\frac{1}{2\pi} \frac{d\theta(f)}{df} \quad (2.194)$$

**Figure 2.17**

Amplitude and phase response and group and phase delays of the filter for Example 2.24.

(a) Amplitude response. (b) Phase response. (c) Group delay. (d) Phase delay.

in which $\theta(f)$ is the phase response of the system. For a distortionless system, the phase response function is given by (2.193) as

$$\theta(f) = -2\pi f t_0 \quad (2.195)$$

This yields a group delay of

$$T_g(f) = -\frac{1}{2\pi} \frac{d}{df} (-2\pi f t_0)$$

or

$$T_g(f) = t_0 \quad (2.196)$$

This confirms the preceding observation that the group delay of a distortionless LTI system is a constant.

Group delay is the delay that a group of two or more frequency components undergo in passing through a linear system. If a linear system has a single-frequency component as the input, the system is always distortionless, since the output can be written as an amplitude-scaled and phase-shifted (time-delayed) version of the input. As an example, assume that the input to a linear system is given by

$$x(t) = A \cos(2\pi f_1 t) \quad (2.197)$$

It follows from (2.188) that the output can be written as

$$y(t) = A |H(f_1)| \cos[2\pi f_1 t + \theta(f_1)] \quad (2.198)$$

where $\theta(f_1)$ is the phase response of the system evaluated at $f = f_1$. Equation (2.198) can be written as

$$y(t) = A |H(f_1)| \cos \left\{ 2\pi f_1 \left[t + \frac{\theta(f_1)}{2\pi f_1} \right] \right\} \quad (2.199)$$

The delay of the single component is defined as the phase delay:

$$T_p(f) = -\frac{\theta(f)}{2\pi f} \quad (2.200)$$

Thus, (2.199) can be written as

$$y(t) = A|H(f_1)| \cos \{2\pi f_1 [t - T_p(f_1)]\} \quad (2.201)$$

Use of (2.195) shows that for a distortionless system, the phase delay is given by

$$T_p(f) = -\frac{1}{2\pi f} (-2\pi f t_0) = t_0 \quad (2.202)$$

Thus, we see that distortionless systems have equal group and phase delays. The following example should clarify the preceding definitions.

EXAMPLE 2.24

Consider a system with amplitude response and phase shift as shown in Figure 2.17 and the following four inputs:

1. $x_1(t) = \cos(10\pi t) + \cos(12\pi t)$
2. $x_2(t) = \cos(10\pi t) + \cos(26\pi t)$
3. $x_3(t) = \cos(26\pi t) + \cos(34\pi t)$
4. $x_4(t) = \cos(32\pi t) + \cos(34\pi t)$

Although this system is somewhat unrealistic from a practical standpoint, we can use it to illustrate various combinations of amplitude and phase distortion. Using (2.188), we obtain the following corresponding outputs:

1.

$$\begin{aligned} y_1(t) &= 2 \cos \left(10\pi t - \frac{1}{6}\pi \right) + 2 \cos \left(12\pi t - \frac{1}{5}\pi \right) \\ &= 2 \cos \left[10\pi \left(t - \frac{1}{60} \right) \right] + 2 \cos \left[12\pi \left(t - \frac{1}{60} \right) \right] \end{aligned}$$

2.

$$\begin{aligned} y_2(t) &= 2 \cos \left(10\pi t - \frac{1}{6}\pi \right) + \cos \left(26\pi t - \frac{13}{30}\pi \right) \\ &= 2 \cos \left[10\pi \left(t - \frac{1}{60} \right) \right] + \cos \left[26\pi \left(t - \frac{1}{60} \right) \right] \end{aligned}$$

3.

$$\begin{aligned} y_3(t) &= \cos\left(26\pi t - \frac{13}{30}\pi\right) + \cos\left(34\pi t - \frac{1}{2}\pi\right) \\ &= \cos\left[26\pi\left(t - \frac{1}{60}\right)\right] + \cos\left[34\pi\left(t - \frac{1}{68}\right)\right] \end{aligned}$$

4.

$$\begin{aligned} y_4(t) &= \cos\left(32\pi t - \frac{1}{2}\pi\right) + \cos\left(34\pi t - \frac{1}{2}\pi\right) \\ &= \cos\left[32\pi\left(t - \frac{1}{64}\right)\right] + \cos\left[34\pi\left(t - \frac{1}{68}\right)\right] \end{aligned}$$

Checking these results with (2.192), we see that only the input $x_1(t)$ is passed without distortion by the system. For $x_2(t)$, amplitude distortion results, and for $x_3(t)$ and $x_4(t)$, phase (delay) distortion is introduced.

The group delay and phase delay are also illustrated in Figure 2.17. It can be seen that for $|f| \leq 15$ Hz, the group and phase delays are both equal to $\frac{1}{60}$ s. For $|f| > 15$ Hz, the group delay is zero, and the phase delay is

$$T_p(f) = \frac{1}{4|f|}, \quad |f| > 15 \text{ Hz} \quad (2.203)$$

2.6.11 Nonlinear Distortion

To illustrate the idea of nonlinear distortion, let us consider a nonlinear system with the input-output characteristic

$$y(t) = a_1 x(t) + a_2 x^2(t) \quad (2.204)$$

where a_1 and a_2 are constants, and with the input

$$x(t) = A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t) \quad (2.205)$$

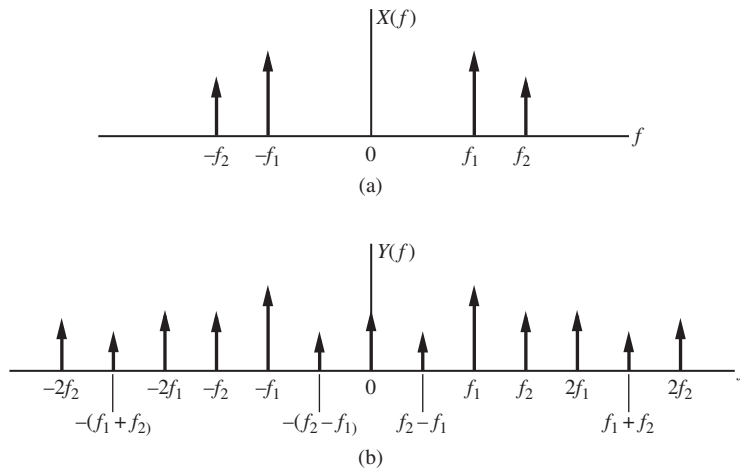
The output is therefore

$$y(t) = a_1 [A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)] + a_2 [A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)]^2 \quad (2.206)$$

Using trigonometric identities, we can write the output as

$$\begin{aligned} y(t) &= a_1 [A_1 \cos(\omega_1 t) + A_2 \cos(\omega_2 t)] \\ &\quad + \frac{1}{2} a_2 (A_1^2 + A_2^2) + \frac{1}{2} a_2 [A_1^2 \cos(2\omega_1 t) + A_2^2 \cos(2\omega_2 t)] \\ &\quad + a_2 A_1 A_2 \{ \cos[(\omega_1 + \omega_2)t] + \cos[(\omega_1 - \omega_2)t] \} \end{aligned} \quad (2.207)$$

As can be seen from (2.207) and as illustrated in Figure 2.18, the system has produced frequencies in the output other than the frequencies of the input. In addition to the first term in (2.207), which may be considered the desired output, there are distortion terms at harmonics of the input frequencies (in this case, second) as well as distortion terms involving sums and differences of the harmonics (in this case, first) of the input frequencies. The former

**Figure 2.18**

Input and output spectra for a nonlinear system with discrete frequency input. (a) Input spectrum. (b) Output spectrum.

are referred to as *harmonic distortion terms*, and the latter are referred to as *intermodulation distortion terms*. Note that a second-order nonlinearity could be used as a device to produce a component at *double* the frequency of an input sinusoid. Third-order nonlinearities can be used as *triplers*, and so forth.

A general input signal can be handled by applying the multiplication theorem given in Table F.6 in Appendix F. Thus, for the nonlinear system with the transfer characteristic given by (2.204), the output spectrum is

$$Y(f) = a_1 X(f) + a_2 X(f) * X(f) \quad (2.208)$$

The second term is considered distortion, and is seen to give interference at all frequencies occupied by the desired output (the first term). It is impossible to isolate harmonic and intermodulation distortion components as before. For example, if

$$X(f) = A\Pi\left(\frac{f}{2W}\right) \quad (2.209)$$

Then the distortion term is

$$a_2 X(f) * X(f) = 2a_2 W A^2 \Lambda\left(\frac{f}{2W}\right) \quad (2.210)$$

The input and output spectra are shown in Figure 2.19. Note that the spectral width of the distortion term is *double* that of the input.

2.6.12 Ideal Filters

It is often convenient to work with filters having idealized frequency response functions with rectangular amplitude-response functions that are constant within the passband and zero elsewhere. We will consider three general types of ideal filters: lowpass, highpass, and

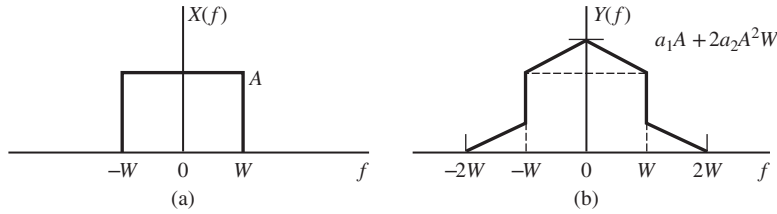


Figure 2.19

Input and output spectra for a nonlinear system with an input whose spectrum is nonzero over a continuous band of frequencies. (a) Input spectrum. (b) Output spectrum.

bandpass. Within the passband, a linear phase response is assumed. Thus, if B is the single-sided bandwidth (*width of the stopband*¹³ for the highpass filter) of the filter in question, the transfer functions of ideal lowpass, highpass, and bandpass filters are easily expressed.

1. For the ideal lowpass filter

$$H_{\text{LP}}(f) = H_0 \Pi(f/2B) e^{-j2\pi f t_0} \quad (2.211)$$

2. For the ideal highpass filter

$$H_{\text{HP}}(f) = H_0 [1 - \Pi(f/2B)] e^{-j2\pi f t_0} \quad (2.212)$$

3. Finally, for the ideal bandpass filter

$$H_{\text{BP}}(f) = [H_1(f - f_0) + H_1(f + f_0)] e^{-j2\pi f t_0} \quad (2.213)$$

where $H_1(f) = H_0 \Pi(f/B)$.

The amplitude-response and phase response functions for these filters are shown in Figure 2.20.

The corresponding impulse responses are obtained by inverse Fourier transformation of the respective frequency response function. For example, the impulse response of an ideal lowpass filter is, from Example 2.12 and the time-delay theorem, given by

$$h_{\text{LP}}(t) = 2BH_0 \text{sinc} [2B(t - t_0)] \quad (2.214)$$

Since $h_{\text{LP}}(t)$ is not zero for $t < 0$, we see that an ideal lowpass filter is noncausal. Nevertheless, ideal filters are useful concepts because they simplify calculations and can give satisfactory results for spectral considerations.

Turning to the ideal bandpass filter, we may use the modulation theorem to write its impulse response as

$$h_{\text{BP}}(t) = 2h_1(t - t_0) \cos [2\pi f_0(t - t_0)] \quad (2.215)$$

where

$$h_1(t) = \mathfrak{F}^{-1}[H_1(f)] = H_0 B \text{sinc} (Bt) \quad (2.216)$$

¹³The *stopband* of a filter will be defined here as the frequency range(s) for which $|H(f)|$ is below 3 dB of its maximum value.

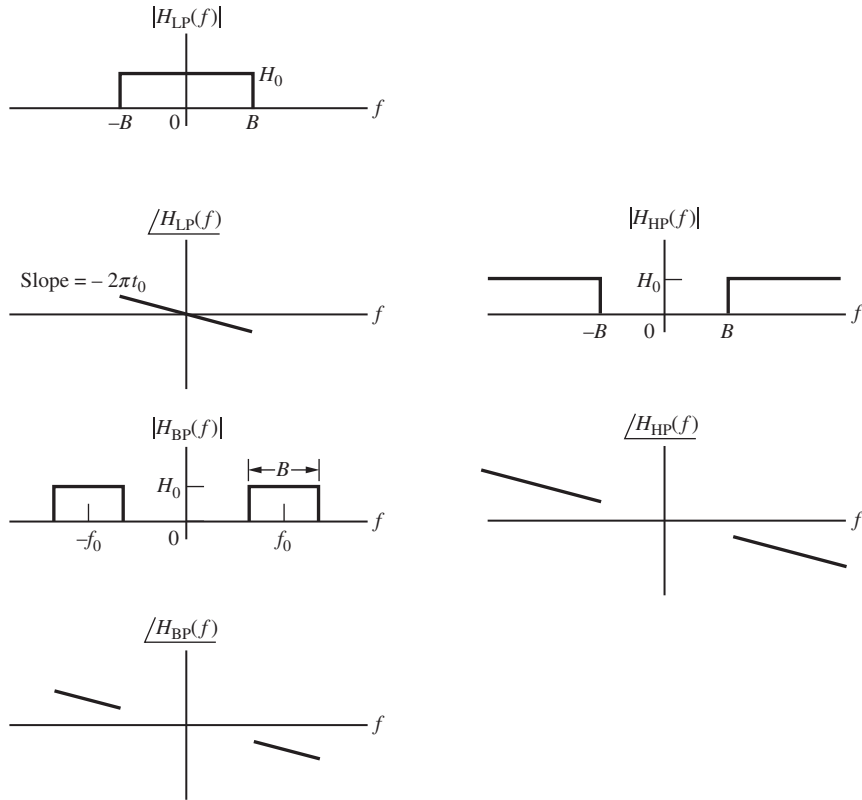


Figure 2.20
Amplitude-response and phase response functions for ideal filters.

Thus the impulse response of an ideal bandpass filter is the oscillatory signal

$$h_{\text{BP}}(t) = 2H_0B \operatorname{sinc} [B(t - t_0)] \cos [2\pi f_0(t - t_0)] \quad (2.217)$$

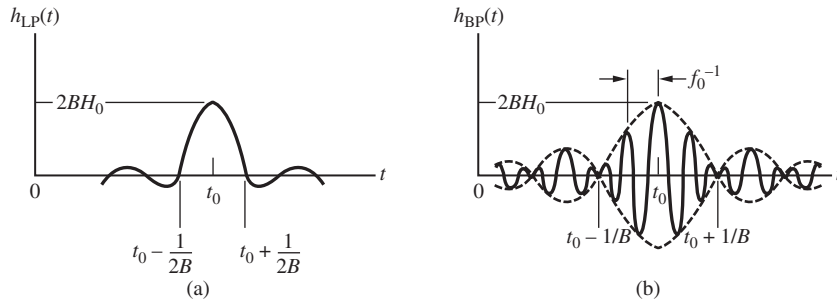
Figure 2.21 illustrates $h_{\text{LP}}(t)$ and $h_{\text{BP}}(t)$. If $f_0 \gg B$, it is convenient to view $h_{\text{BP}}(t)$ as the slowly varying envelope $2H_0 \operatorname{sinc} (Bt)$ modulating the high-frequency oscillatory signal $\cos (2\pi f_0 t)$ and shifted to the right by t_0 seconds.

Derivation of the impulse response of an ideal highpass filter is left to the problems (Problem 2.63).

2.6.13 Approximation of Ideal Lowpass Filters by Realizable Filters

Although ideal filters are noncausal and therefore unrealizable devices,¹⁴ there are several practical filter types that may be designed to approximate ideal filter characteristics as closely as desired. In this section we consider three such approximations for the lowpass case. Bandpass and highpass approximations may be obtained through suitable frequency

¹⁴See Williams and Taylor (1988), Chapter 2, for a detailed discussion of classical filter designs.

**Figure 2.21**

Impulse responses for ideal lowpass and bandpass filters. (a) $h_{LP}(t)$. (b) $h_{BP}(t)$.

transformation. The three filter types to be considered are (1) Butterworth, (2) Chebyshev, and (3) Bessel.

The Butterworth filter is a filter design chosen to maintain a constant amplitude response in the passband at the cost of less stopband attenuation. An n th-order Butterworth filter is characterized by a transfer function, in terms of the complex frequency s , of the form

$$H_{BW}(s) = \frac{\omega_3^n}{(s - s_1)(s - s_2) \cdots (s - s_n)} \quad (2.218)$$

where the poles s_1, s_2, \dots, s_n are symmetrical with respect to the real axis and equally spaced about a semicircle of radius ω_3 in the left half s -plane and $f_3 = \omega_3/2\pi$ is the 3-dB cutoff frequency.¹⁵ Typical pole locations are shown in Figure 2.22(a). For example, the system function of a second-order Butterworth filter is

$$H_{2\text{nd-order BW}}(s) = \frac{\omega_3^2}{\left(s + \frac{1+j}{\sqrt{2}}\omega_3\right)\left(s + \frac{1-j}{\sqrt{2}}\omega_3\right)} = \frac{\omega_3^2}{s^2 + \sqrt{2}\omega_3 s + \omega_3^2} \quad (2.219)$$

where $f_3 = \frac{\omega_3}{2\pi}$ is the 3-dB cutoff frequency in hertz. The amplitude response for an n th-order Butterworth filter is of the form

$$|H_{BU}(f)| = \frac{1}{\sqrt{1 + (f/f_3)^{2n}}} \quad (2.220)$$

Note that as n approaches infinity, $|H_{BU}(f)|$ approaches an ideal lowpass filter characteristic. However, the filter delay also approaches infinity.

The Chebyshev (type 1) lowpass filter has an amplitude response chosen to maintain a minimum allowable attenuation in the passband while maximizing the attenuation in the stopband. A typical pole-zero diagram is shown in Figure 2.22(b). The amplitude response of a Chebyshev filter is of the form

$$|H_C(f)| = \frac{1}{\sqrt{1 + \epsilon^2 C_n^2(f)}} \quad (2.221)$$

¹⁵From basic circuit theory courses you will recall that the poles and zeros of a rational function of s , $H(s) = N(s)/D(s)$, are those values of complex frequency $s = \sigma + j\omega$ for which $D(s) = 0$ and $N(s) = 0$, respectively.

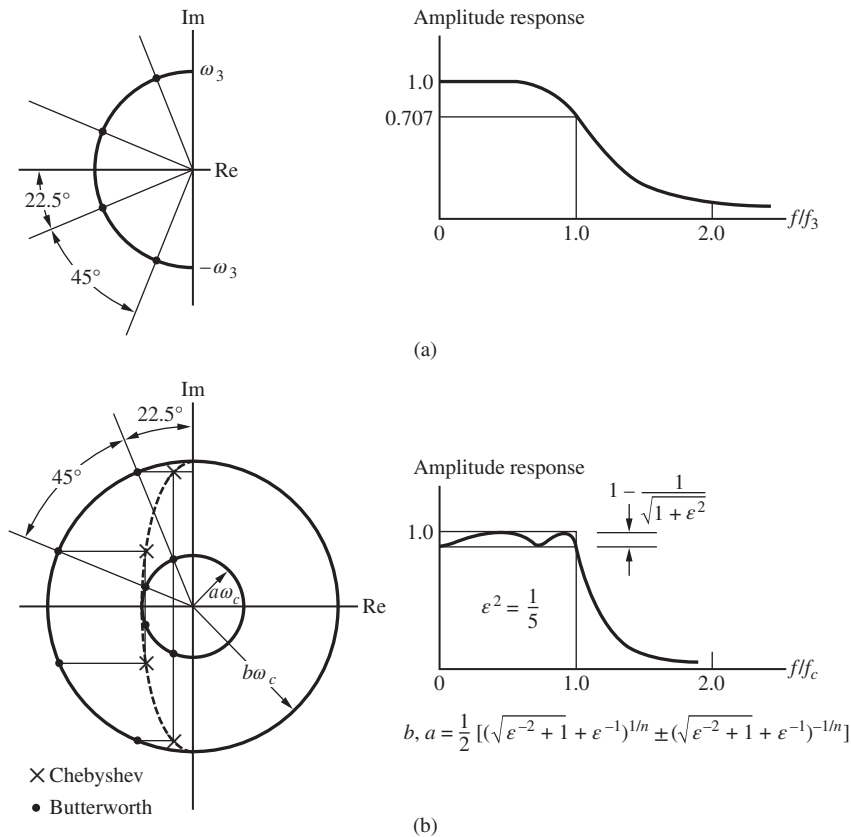


Figure 2.22 Pole locations and amplitude responses for fourth-order Butterworth and Chebyshev filters. (a) Butterworth filter. (b) Chebyshev filter.

The parameter ϵ is specified by the minimum allowable attenuation in the passband, and $C_n(f)$, known as a Chebyshev polynomial, is given by the recursion relation

$$C_n(f) = 2 \left(\frac{f}{f_c} \right) C_{n-1}(f) - C_{n-2}(f), \quad n = 2, 3, \dots \quad (2.222)$$

where

$$C_1(f) = \frac{f}{f_c} \text{ and } C_0(f) = 1 \quad (2.223)$$

Regardless of the value of n , it turns out that $C_n(f_c) = 1$, so that $H_C(f_c) = (1 + \epsilon^2)^{-1/2}$. (Note that f_c is not necessarily the 3-dB frequency here.)

The Bessel lowpass filter is a design that attempts to maintain a linear phase response in the passband at the expense of the amplitude response. The cutoff frequency of a Bessel filter is defined by

$$f_c = (2\pi t_0)^{-1} = \frac{\omega_c}{2\pi} \quad (2.224)$$

where t_0 is the nominal delay of the filter. The frequency response function of an n th-order Bessel filter is given by

$$H_{BE}(f) = \frac{K_n}{B_n(f)} \quad (2.225)$$

where K_n is a constant chosen to yield $H(0) = 1$, and $B_n(f)$ is a Bessel polynomial of order n defined by

$$B_n(f) = (2n - 1)B_{n-1}(f) - \left(\frac{f}{f_c}\right)^2 B_{n-2}(f) \quad (2.226)$$

where

$$B_0(f) = 1 \text{ and } B_1(f) = 1 + j\left(\frac{f}{f_c}\right) \quad (2.227)$$

Figure 2.23 illustrates the amplitude-response and group-delay characteristics of third-order Butterworth, Bessel, and Chebyshev filters. All three filters are normalized to have

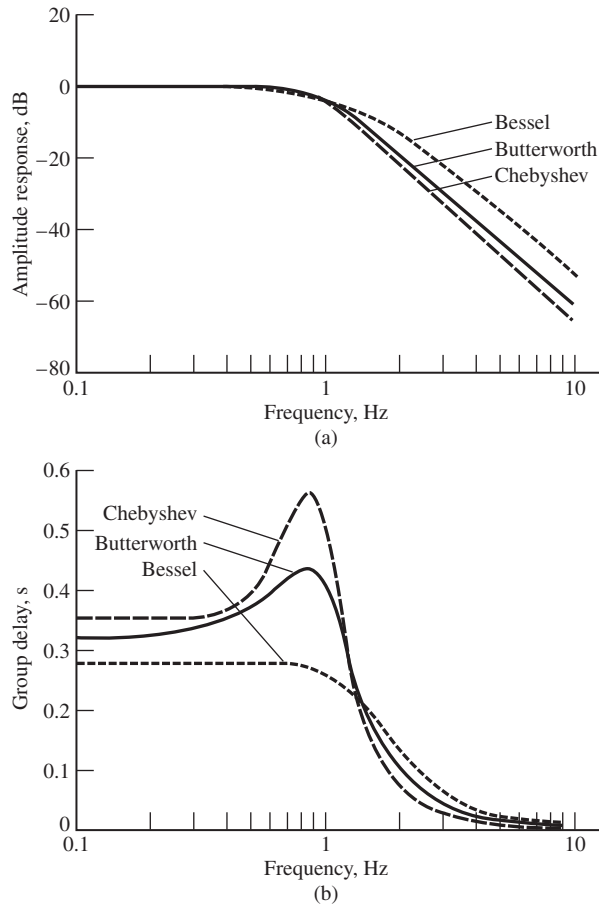


Figure 2.23

Comparison of third-order Butterworth, Chebyshev (0.1-dB ripple), and Bessel filters. (a) Amplitude response. (b) Group delay. All filters are designed to have a 1-Hz, 3-dB bandwidth.

3-dB amplitude attenuation at a frequency of f_c Hz. The amplitude responses show that the Chebyshev filters have more attenuation than the Butterworth and Bessel filters do for frequencies exceeding the 3-dB frequency. Increasing the passband ($f < f_c$) ripple of a Chebyshev filter increases the stopband ($f > f_c$) attenuation.

The group delay characteristics shown in Figure 2.23(b) illustrate, as expected, that the Bessel filter has the most constant group delay. Comparison of the Butterworth and the 0.1-dB ripple Chebyshev group delays shows that although the group delay of the Chebyshev filter has a higher peak, it has a more constant group delay for frequencies less than about $0.4f_c$.

COMPUTER EXAMPLE 2.2

The MATLAB™ program given below can be used to plot the amplitude and phase responses of Butterworth and Chebyshev filters of any order and any cutoff frequency (3-dB frequency for Butterworth). The ripple is also an input for the Chebyshev filter. Several MATLAB™ subprograms are used, such as `logspace`, `butter`, `cheby1`, `freqs`, and `cart2pol`. It is suggested that the student use the help feature of MATLAB™ to find out how these are used. For example, a line `freqs (num, den, W)` in the command window automatically plots amplitude and phase responses. However, we have used `semilogx` here to plot the amplitude response in dB versus frequency in hertz on a logarithmic scale.

```
% file: c2ce2
% Frequency response for Butterworth and Chebyshev 1 filters
%
clf
filt.type = input('Enter filter type; 1 = Butterworth; 2 = Chebyshev 1');
n.max = input('Enter maximum order of filter ');
fc = input('Enter cutoff frequency (3-dB for Butterworth) in Hz ');
if filt.type == 2
    R = input('Enter Chebyshev filter ripple in dB ');
end
W = logspace(0, 3, 1000); % Set up frequency axis; hertz assumed
for n = 1:n.max
    if filt.type == 1 % Generate num. and den. polynomials
        [num,den]=butter(n, 2*pi*fc, 's');
    elseif filt.type == 2
        [num,den]=cheby1(n, R, 2*pi*fc, 's');
    end
    H = freqs(num, den, W); % Generate complex frequency response
    [phase, mag] = cart2pol(real(H),imag(H)); % Convert H to polar
coordinates
    subplot(2,1,1),semilogx(W/(2*pi),20*log10(mag)),...
    axis([min(W/(2*pi)) max(W/(2*pi)) -20 0]),...
    if n == 1 % Put on labels and title; hold for future plots
        ylabel('|H| in dB')
        hold on
        if filt.type == 1
            title(['Butterworth filter responses: order 1 -
                ',num2str(n.max),'; ...
                cutoff freq = ',num2str(fc),' Hz'])
        elseif filt.type == 2
            title(['Chebyshev filter responses: order 1 -
                ',num2str(n.max),'; ...
                ripple = ',num2str(R),' dB; cutoff freq = ',num2str(fc),'
                Hz'])
        end
    end
```

```

end
subplot(2,1,2),semilogx(W/(2*pi),180*phase/pi),...
axis([min(W/(2*pi)) max(W/(2*pi)) -200 200]),...
if n == 1
grid on
hold on
xlabel('f, Hz'),ylabel('phase in degrees')
end
end
% End of script file

```

2.6.14 Relationship of Pulse Resolution and Risettime to Bandwidth

In our consideration of signal distortion, we assumed bandlimited signal spectra. We found that the input signal to a filter is merely delayed and attenuated if the filter has constant amplitude response and linear phase response throughout the passband of the signal. But suppose the input signal is not bandlimited. What rule of thumb can we use to estimate the required bandwidth? This is a particularly important problem in pulse transmission, where the detection and resolution of pulses at a filter output are of interest.

A satisfactory definition for pulse duration and bandwidth, and the relationship between them, is obtained by consulting Figure 2.24. In Figure 2.24(a), a pulse with a single maximum, taken at $t = 0$ for convenience, is shown with a rectangular approximation of height $x(0)$ and duration T . It is required that the approximating pulse and $|x(t)|$ have equal areas. Thus,

$$Tx(0) = \int_{-\infty}^{\infty} |x(t)| dt \geq \int_{-\infty}^{\infty} x(t) dt = X(0) \quad (2.228)$$

where we have used the relationship

$$X(0) = \mathfrak{F}[x(t)]|_{f=0} = \int_{-\infty}^{\infty} x(t)e^{-j2\pi t \cdot 0} dt \quad (2.229)$$

Turning to Figure 2.24(b), we obtain a similar inequality for the rectangular approximation to the pulse spectrum. Specifically, we may write

$$2WX(0) = \int_{-\infty}^{\infty} |X(f)| df \geq \int_{-\infty}^{\infty} X(f) df = x(0) \quad (2.230)$$

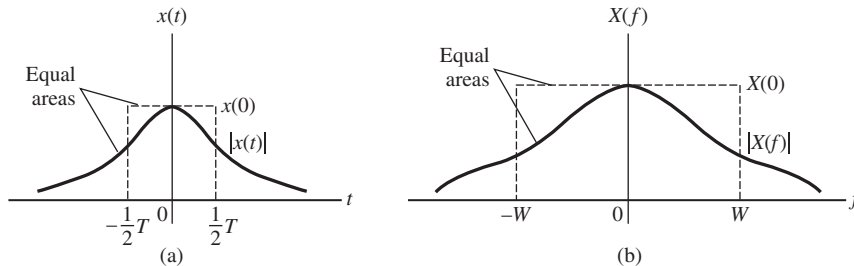


Figure 2.24

Arbitrary pulse signal and spectrum. (a) Pulse and rectangular approximation. (b) Amplitude spectrum and rectangular approximation.

where we have used the relationship

$$x(0) = \mathfrak{F}^{-1} [X(f)] \Big|_{t=0} = \int_{-\infty}^{\infty} X(f) e^{j2\pi f \cdot 0} df \quad (2.231)$$

Thus, we have the pair of inequalities

$$\frac{x(0)}{X(0)} \geq \frac{1}{T} \text{ and } 2W \geq \frac{x(0)}{X(0)} \quad (2.232)$$

which, when combined, result in the relationship of pulse duration and bandwidth

$$2W \geq \frac{1}{T} \quad (2.233)$$

or

$$W \geq \frac{1}{2T} \text{ Hz} \quad (2.234)$$

Other definitions of pulse duration and bandwidth could have been used, but a relationship similar to (2.233) and (2.234) would have resulted.

This inverse relationship between pulse duration and bandwidth has been illustrated by all the examples involving pulse spectra that we have considered so far (for example, Examples 2.8, 2.11, 2.13).

If pulses with bandpass spectra are considered, the relationship is

$$W \geq \frac{1}{T} \text{ Hz} \quad (2.235)$$

This is illustrated by Example 2.16.

A result similar to (2.233) and (2.234) also holds between the *risetime* T_R and bandwidth of a pulse. A suitable definition of *risetime* is the time required for a pulse's leading edge to go from 10% to 90% of its final value. For the bandpass case, (2.235) holds with T replaced by T_R , where T_R is the risetime of the *envelope* of the pulse.

Risetime can be used as a measure of a system's distortion. To see how this is accomplished, we will express the step response of a filter in terms of its impulse response. From the superposition integral of (2.160), with $x(t - \sigma) = u(t - \sigma)$, the step response of a filter with impulse response $h(t)$ is

$$\begin{aligned} y_s(t) &= \int_{-\infty}^{\infty} h(\sigma) u(t - \sigma) d\sigma \\ &= \int_{-\infty}^t h(\sigma) d\sigma \end{aligned} \quad (2.236)$$

This follows because $u(t - \sigma) = 0$ for $\sigma > t$. Therefore, the step response of an LTI system is the integral of its impulse response. This is not too surprising, since the unit step function is the integral of a unit impulse function.¹⁶

Examples 2.25 and 2.26 demonstrate how the risetime of a system's output due to a step input is a measure of the fidelity of the system.

¹⁶This result is a special case of a more general result for an LTI system: if the response of a system to a given input is known and that input is modified through a linear operation, such as integration, then the output to the modified input is obtained by performing the same linear operation on the output due to the original input.

EXAMPLE 2.25

The impulse response of a lowpass RC filter is given by

$$h(t) = \frac{1}{RC} e^{-t/RC} u(t) \quad (2.237)$$

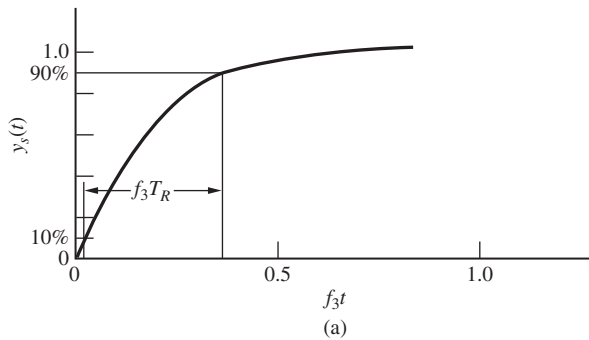
for which the step response is found to be

$$y_s(t) = (1 - e^{-2\pi f_3 t}) u(t) \quad (2.238)$$

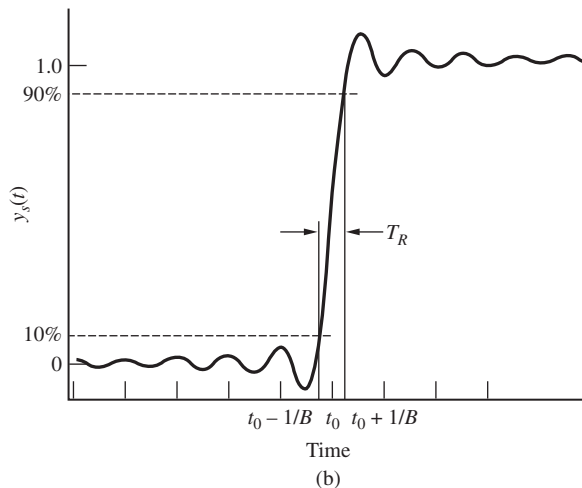
where the 3-dB bandwidth of the filter, defined following (2.175), has been used. The step response is plotted in Figure 2.25(a), where it is seen that the 10% to 90% risetime is approximately

$$T_R = \frac{0.35}{f_3} = 2.2RC \quad (2.239)$$

which demonstrates the inverse relationship between bandwidth and risetime.

**Figure 2.25**

Step response of (a) a lowpass RC filter and (b) an ideal lowpass filter, illustrating 10% to 90% risetime of each.



EXAMPLE 2.26

Using (2.214) with $H_0 = 1$, the step response of an ideal lowpass filter is

$$\begin{aligned} y_s(t) &= \int_{-\infty}^t 2B \operatorname{sinc} [2B(\sigma - t_0)] d\sigma \\ &= \int_{-\infty}^t 2B \frac{\sin [2\pi B(\sigma - t_0)]}{2\pi B(\sigma - t_0)} d\sigma \end{aligned} \quad (2.240)$$

By changing variables in the integrand to $u = 2\pi B(\sigma - t_0)$, the step response becomes

$$y_s(t) = \frac{1}{2\pi} \int_{-\infty}^{2\pi B(t-t_0)} \frac{\sin(u)}{u} du = \frac{1}{2} + \frac{1}{\pi} \operatorname{Si}[2\pi B(t - t_0)] \quad (2.241)$$

where $\operatorname{Si}(x) = \int_0^x (\sin u/u) du = -\operatorname{Si}(-x)$ is the sine-integral function.¹⁷ A plot of $y_s(t)$ for an ideal lowpass filter, such as is shown in Figure 2.25(b), reveals that the 10% to 90% risetime is approximately

$$T_R \cong \frac{0.44}{B} \quad (2.242)$$

Again, the inverse relationship between bandwidth and risetime is demonstrated. ■

2.7 SAMPLING THEORY

In many applications it is useful to represent a signal in terms of sample values taken at appropriately spaced intervals. Such sample-data systems find application in control systems and pulse-modulation communication systems.

In this section we consider the representation of a signal $x(t)$ by a so-called ideal *instantaneous sampled waveform* of the form

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} x(nT_s) \delta(t - nT_s) \quad (2.243)$$

where T_s is the sampling interval. Two questions to be answered in connection with such sampling are, “What are the restrictions on $x(t)$ and T_s to allow perfect recovery of $x(t)$ from $x_\delta(t)$?” and “How is $x(t)$ recovered from $x_\delta(t)$?” Both questions are answered by the uniform sampling theorem for lowpass signals, which may be stated as follows:

Theorem

If a signal $x(t)$ contains no frequency components for frequencies above $f = W$ hertz, then it is completely described by instantaneous sample values uniformly spaced in time with period $T_s < \frac{1}{2W}$. The signal can be exactly reconstructed from the sampled waveform given by (2.243) by passing it through an ideal lowpass filter with bandwidth B , where $W < B < f_s - W$ with $f_s = T_s^{-1}$. The frequency $2W$ is referred to as the *Nyquist frequency*.

¹⁷See M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, New York: Dover Publications, 1972, pp. 238ff (Copy of the 10th National Bureau of Standards Printing).

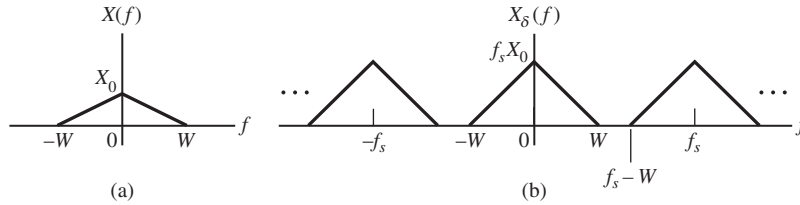


Figure 2.26

Signal spectra for lowpass sampling. (a) Assumed spectrum for $x(t)$. (b) Spectrum of the sampled signal.

To prove the sampling theorem, we find the spectrum of (2.243). Since $\delta(t - nT_s)$ is zero everywhere except at $t = nT_s$, (2.243) can be written as

$$x_\delta(t) = \sum_{n=-\infty}^{\infty} x(t)\delta(t - nT_s) = x(t) \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \quad (2.244)$$

Applying the multiplication theorem of Fourier transforms, (2.102), the Fourier transform of (2.244) is

$$X_\delta(f) = X(f) * \left[f_s \sum_{n=-\infty}^{\infty} \delta(f - nf_s) \right] \quad (2.245)$$

where the transform pair (2.119) has been used. Interchanging the orders of summation and convolution, and noting that

$$X(f) * \delta(f - nf_s) = \int_{-\infty}^{\infty} X(u) \delta(f - u - nf_s) du = X(f - nf_s) \quad (2.246)$$

by the sifting property of the delta function, we obtain

$$X_\delta(f) = f_s \sum_{n=-\infty}^{\infty} X(f - nf_s) \quad (2.247)$$

Thus, assuming that the spectrum of $x(t)$ is bandlimited to W Hz and that $f_s > 2W$ as stated in the sampling theorem, we may readily sketch $X_\delta(f)$. Figure 2.26 shows a typical choice for $X(f)$ and the corresponding $X_\delta(f)$. We note that sampling simply results in a periodic repetition of $X(f)$ in the frequency domain with a spacing f_s . If $f_s < 2W$, the separate terms in (2.247) overlap, and there is no apparent way to recover $x(t)$ from $x_\delta(t)$ without distortion. On the other hand, if $f_s > 2W$, the term in (2.247) for $n = 0$ is easily separated from the rest by ideal lowpass filtering. Assuming an ideal lowpass filter with the frequency response function

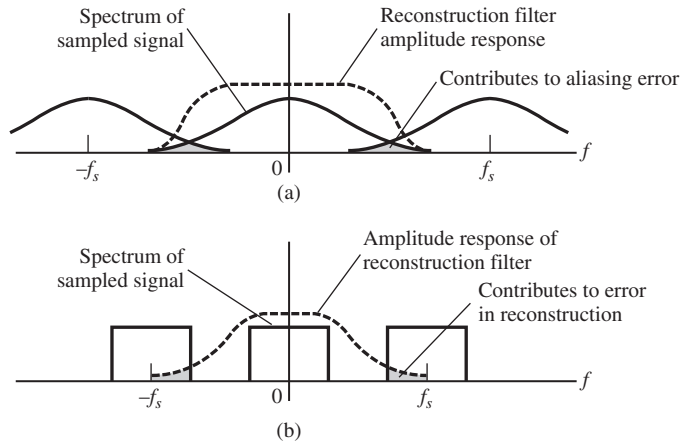
$$H(f) = H_0 \Pi \left(\frac{f}{2B} \right) e^{-j2\pi f t_0}, \quad W \leq B \leq f_s - W \quad (2.248)$$

the output spectrum, with $x_\delta(t)$ at the input, is

$$Y(f) = f_s H_0 X(f) e^{-j2\pi f t_0} \quad (2.249)$$

and by the time-delay theorem, the output waveform is

$$y(t) = f_s H_0 x(t - t_0) \quad (2.250)$$

**Figure 2.27**

Spectra illustrating two types of errors encountered in reconstruction of sampled signals. (a) Illustration of aliasing error in the reconstruction of sampled signals. (b) Illustration of error due to nonideal reconstruction filter.

Thus, if the conditions of the sampling theorem are satisfied, we see that distortionless recovery of $x(t)$ from $x_s(t)$ is possible. Conversely, if the conditions of the sampling theorem are not satisfied, either because $x(t)$ is not bandlimited or because $f_s < 2W$, we see that distortion at the output of the reconstruction filter is inevitable. Such distortion, referred to as *aliasing*, is illustrated in Figure 2.27(a). It can be combated by filtering the signal before sampling or by increasing the sampling rate. A second type of error, illustrated in Figure 2.27(b), occurs in the reconstruction process and is due to the nonideal frequency response characteristics of practical filters. This type of error can be minimized by choosing reconstruction filters with sharper rolloff characteristics or by increasing the sampling rate. Note that the error due to aliasing and the error due to imperfect reconstruction filters are both *proportional to signal level*. Thus, increasing the signal amplitude does not improve the signal-to-error ratio.

An alternative expression for the reconstructed output from the ideal lowpass filter can be obtained by noting that when (2.243) is passed through a filter with impulse response $h(t)$, the output is

$$y(t) = \sum_{n=-\infty}^{\infty} x(nT_s)h(t - nT_s) \quad (2.251)$$

But $h(t)$ corresponding to (2.248) is given by (2.214). Thus,

$$y(t) = 2BH_0 \sum_{n=-\infty}^{\infty} x(nT_s) \operatorname{sinc} [2B(t - t_0 - nT_s)] \quad (2.252)$$

and we see that just as a periodic signal can be completely represented by its Fourier coefficients, a *bandlimited signal can be completely represented by its sample values*.

By setting $B = \frac{1}{2}f_s$, $H_0 = T_s$, and $t_0 = 0$ for simplicity, (2.252) becomes

$$y(t) = \sum_n x(nT_s) \operatorname{sinc} (f_s t - n) \quad (2.253)$$

This expansion is equivalent to a generalized Fourier series, for we may show that

$$\int_{-\infty}^{\infty} \text{sinc}(f_s t - n) \text{sinc}(f_s t - m) dt = \delta_{nm} \quad (2.254)$$

where $\delta_{nm} = 1$, $n = m$, and is 0 otherwise.

Turning next to bandpass spectra, for which the upper limit on frequency f_u is much larger than the single-sided bandwidth W , one may naturally inquire as to the feasibility of sampling at rates less than $f_s > 2f_u$. The *uniform sampling theorem for bandpass spectra* gives the conditions for which this is possible.

Theorem

If a signal has a spectrum of bandwidth W Hz and upper frequency limit f_u , then a rate f_s at which the signal can be sampled is $2f_u/m$, where m is the largest integer not exceeding f_u/W . All higher sampling rates are not necessarily usable unless they exceed $2f_u$.

EXAMPLE 2.27

Consider the bandpass signal $x(t)$ with the spectrum shown in Figure 2.28. According to the bandpass sampling theorem, it is possible to reconstruct $x(t)$ from sample values taken at a rate of

$$f_s = \frac{2f_u}{m} = \frac{2(3)}{2} = 3 \text{ samples per second} \quad (2.255)$$

whereas the lowpass sampling theorem requires 6 samples per second.

To show that this is possible, we sketch the spectrum of the sampled signal. According to (2.247), which holds in general,

$$X_\delta(f) = 3 \sum_{-\infty}^{\infty} X(f - 3n) \quad (2.256)$$

The resulting spectrum is shown in Figure 2.28(b), and we see that it is theoretically possible to recover $x(t)$ from $x_\delta(t)$ by bandpass filtering.

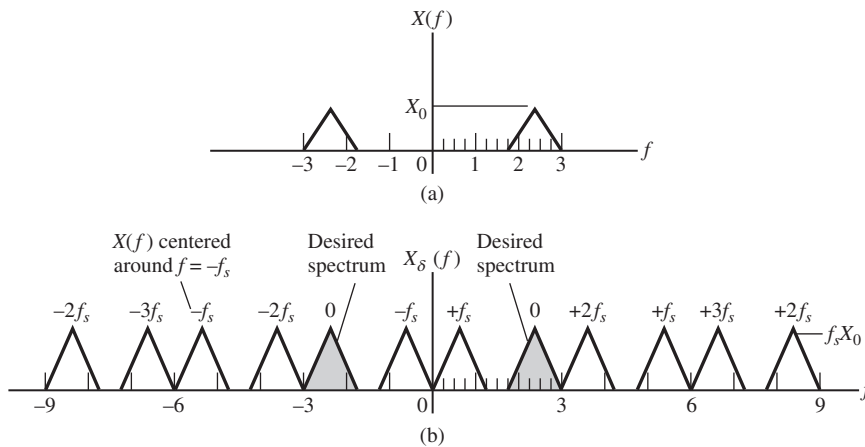


Figure 2.28

Signal spectra for bandpass sampling. (a) Assumed bandpass signal spectrum. (b) Spectrum of the sampled signal.

Another way of sampling a bandpass signal of bandwidth W is to resolve it into two lowpass quadrature signals of bandwidth $\frac{1}{2}W$. Both of these may then be sampled at a minimum rate of $2\left(\frac{1}{2}W\right) = W$ samples per second, thus resulting in an overall minimum sampling rate of $2W$ samples per second.

2.8 THE HILBERT TRANSFORM

(It may be advantageous to postpone this section until consideration of single-sideband systems in Chapter 3.)

2.8.1 Definition

Consider a filter that simply phase-shifts all frequency components of its input by $-\frac{1}{2}\pi$ radians; that is, its frequency response function is

$$H(f) = -j \operatorname{sgn} f \quad (2.257)$$

where the sgn function (read “signum f ”) is defined as

$$\operatorname{sgn}(f) = \begin{cases} 1, & f > 0 \\ 0, & f = 0 \\ -1, & f < 0 \end{cases} \quad (2.258)$$

We note that $|H(f)| = 1$ and $\angle H(f)$ is odd, as it must be. If $X(f)$ is the input spectrum to the filter, the output spectrum is $-j \operatorname{sgn}(f)X(f)$, and the corresponding time function is

$$\begin{aligned} \hat{x}(t) &= \mathfrak{F}^{-1}[-j \operatorname{sgn}(f)X(f)] \\ &= h(t) * x(t) \end{aligned} \quad (2.259)$$

where $h(t) = -j\mathfrak{F}^{-1}[\operatorname{sgn} f]$ is the impulse response of the filter. To obtain $\mathfrak{F}^{-1}[\operatorname{sgn} f]$ without resorting to contour integration, we consider the inverse transform of the function

$$G(f; \alpha) = \begin{cases} e^{-\alpha f}, & f > 0 \\ -e^{\alpha f}, & f < 0 \end{cases} \quad (2.260)$$

We note that $\lim_{\alpha \rightarrow 0} G(f; \alpha) = \operatorname{sgn} f$. Thus, our procedure will be to inverse Fourier-transform $G(f; \alpha)$ and take the limit of the result as α approaches zero. Performing the inverse transformation, we obtain

$$\begin{aligned} g(t; \alpha) &= \mathfrak{F}^{-1}[G(f; \alpha)] \\ &= \int_0^{\infty} e^{-\alpha f} e^{j2\pi ft} df - \int_{-\infty}^0 e^{\alpha f} e^{j2\pi ft} df = \frac{j4\pi t}{\alpha^2 + (2\pi t)^2} \end{aligned} \quad (2.261)$$

Taking the limit as α approaches zero, we get the transform pair

$$\frac{j}{\pi t} \longleftrightarrow \operatorname{sgn}(f) \quad (2.262)$$

Using this result in (2.259), we obtain the output of the filter:

$$\hat{x}(t) = \int_{-\infty}^{\infty} \frac{x(\lambda)}{\pi(t-\lambda)} d\lambda = \int_{-\infty}^{\infty} \frac{x(t-\eta)}{\pi\eta} d\eta \quad (2.263)$$

The signal $\hat{x}(t)$ is defined as the *Hilbert transform* of $x(t)$. Since the Hilbert transform corresponds to a phase shift of $-\frac{1}{2}\pi$, we note that the Hilbert transform of $\hat{x}(t)$ corresponds to the frequency response function $(-j \operatorname{sgn} f)^2 = -1$, or a phase shift of π radians. Thus,

$$\widehat{\hat{x}}(t) = -x(t) \quad (2.264)$$

EXAMPLE 2.28

For an input to a Hilbert transform filter of

$$x(t) = \cos(2\pi f_0 t) \quad (2.265)$$

which has a spectrum given by

$$X(f) = \frac{1}{2}\delta(f-f_0) + \frac{1}{2}\delta(f+f_0) \quad (2.266)$$

we obtain an output spectrum from the Hilbert transformer of

$$\hat{X}(f) = \frac{1}{2}\delta(f-f_0)e^{-j\pi/2} + \frac{1}{2}\delta(f+f_0)e^{j\pi/2} \quad (2.267)$$

Taking the inverse Fourier transform of (2.267), we find the output signal to be

$$\begin{aligned} \hat{x}(t) &= \frac{1}{2}e^{j2\pi f_0 t}e^{-j\pi/2} + \frac{1}{2}e^{-j2\pi f_0 t}e^{j\pi/2} \\ &= \cos(2\pi f_0 t - \pi/2) \end{aligned}$$

$$\text{or } \widehat{\cos(2\pi f_0 t)} = \sin(2\pi f_0 t) \quad (2.268)$$

Of course, the Hilbert transform could have been found by inspection in this case by subtracting $\frac{1}{2}\pi$ from the argument of the cosine. Doing this for the signal $\sin \omega_0 t$, we find that

$$\widehat{\sin(2\pi f_0 t)} = \sin\left(2\pi f_0 t - \frac{1}{2}\pi\right) = -\cos(2\pi f_0 t) \quad (2.269)$$

We may use the two results obtained to show that

$$e^{j2\pi f_0 t} = -j \operatorname{sgn}(f_0) e^{j2\pi f_0 t} \quad (2.270)$$

This is done by considering the two cases $f_0 > 0$ and $f_0 < 0$, and using Euler's theorem in conjunction with the results of (2.268) and (2.269). The result (2.270) also follows directly by considering the response of a Hilbert transform filter with frequency response $H_{\text{HT}}(f) = -j \operatorname{sgn}(2\pi f)$ to the input $x(t) = e^{j2\pi f_0 t}$. ■

2.8.2 Properties

The Hilbert transform has several useful properties that will be illustrated later. Three of these properties will be proved here.

1. The energy (or power) in a signal $x(t)$ and its Hilbert transform $\hat{x}(t)$ are equal. To show this, we consider the energy spectral densities at the input and output of a Hilbert transform filter. Since $H(f) = -j \operatorname{sgn} f$, these densities are related by

$$\left| \hat{X}(f) \right|^2 \triangleq \left| \mathfrak{F} [\hat{x}(t)] \right|^2 = |-j \operatorname{sgn} (f)|^2 |X(f)|^2 = |X(f)|^2 \quad (2.271)$$

where $\hat{X}(f) = \mathfrak{F} [\hat{x}(t)] = -j \operatorname{sgn} (f) X(f)$. Thus, since the energy spectral densities at input and output are equal, so are the total energies. A similar proof holds for power signals.

2. A signal and its Hilbert transform are orthogonal; that is,

$$\int_{-\infty}^{\infty} x(t)\hat{x}(t) dt = 0 \text{ (energy signals)} \quad (2.272)$$

or

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T x(t)\hat{x}(t) dt = 0 \text{ (power signals)} \quad (2.273)$$

Considering (2.272), we note that the left-hand side can be written as

$$\int_{-\infty}^{\infty} x(t)\hat{x}(t) dt = \int_{-\infty}^{\infty} X(f)\hat{X}^*(f) df \quad (2.274)$$

by Parseval's theorem generalized, where $\hat{X}(f) = \mathfrak{F}[\hat{x}(t)] = -j \operatorname{sgn} (f) X(f)$. It therefore follows that

$$\int_{-\infty}^{\infty} x(t)\hat{x}(t) dt = \int_{-\infty}^{\infty} (+j \operatorname{sgn} f) |X(f)|^2 df \quad (2.275)$$

But the integrand of the right-hand side of (2.275) is odd, being the product of the even function $|X(f)|^2$ and the odd function $j \operatorname{sgn} f$. Therefore, the integral is zero, and (2.272) is proved. A similar proof holds for (2.273).

3. If $c(t)$ and $m(t)$ are signals with nonoverlapping spectra, where $m(t)$ is lowpass and $c(t)$ is highpass, then

$$\widehat{m(t)c(t)} = m(t)\hat{c}(t) \quad (2.276)$$

To prove this relationship, we use the Fourier integral to represent $m(t)$ and $c(t)$ in terms of their spectra, $M(f)$ and $C(f)$, respectively. Thus,

$$m(t)c(t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} M(f)C(f') \exp[j2\pi(f + f')t] df df' \quad (2.277)$$

where we assume $M(f) = 0$ for $|f| > W$ and $C(f') = 0$ for $|f'| < W$. The Hilbert transform of (2.277) is

$$\begin{aligned} \widehat{m(t)c(t)} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} M(f)C(f') \exp[j2\pi(\widehat{f + f'})t] df df' \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} M(f)C(f') [-j \operatorname{sgn} (f + f')] \exp [j2\pi (f + f') t] df df' \end{aligned} \quad (2.278)$$

where (2.270) has been used. However, the product $M(f)C(f')$ is nonvanishing only for $|f| < W$ and $|f'| > W$, and we may replace $\text{sgn}(f + f')$ by $\text{sgn}(f')$ in this case. Thus,

$$\widehat{m(t)c(t)} = \int_{-\infty}^{\infty} M(f) \exp(j2\pi ft) df \int_{-\infty}^{\infty} C(f') [-j \text{sgn}(f') \exp(j2\pi f't)] df' \quad (2.279)$$

However, the first integral on the right-hand side is just $m(t)$, and the second integral is $\widehat{c(t)}$, since

$$c(t) = \int_{-\infty}^{\infty} C(f') \exp(j2\pi f't) df'$$

and

$$\begin{aligned} \widehat{c(t)} &= \int_{-\infty}^{\infty} C(f') \exp(j2\pi f't) df' \\ &= \int_{-\infty}^{\infty} C(f') [-j \text{sgn}(f') \exp(j2\pi f't)] df' \end{aligned} \quad (2.280)$$

Hence, (2.279) is equivalent to (2.276), which was the relationship to be proved.

EXAMPLE 2.29

Given that $m(t)$ is a lowpass signal with $M(f) = 0$ for $|f| > W$, we may directly apply (2.276) in conjunction with (2.275) and (2.269) to show that

$$m(t) \widehat{\cos \omega_0 t} = m(t) \sin \omega_0 t \quad (2.281)$$

and

$$m(t) \widehat{\sin \omega_0 t} = -m(t) \cos \omega_0 t \quad (2.282)$$

if $f_0 = \omega_0/2\pi > W$. ■

2.8.3 Analytic Signals

An analytic signal $x_p(t)$, corresponding to the real signal $x(t)$, is defined as

$$x_p(t) = x(t) + j\widehat{x}(t) \quad (2.283)$$

where $\widehat{x}(t)$ is the Hilbert transform of $x(t)$. We now consider several properties of an analytic signal.

We used the term *envelope* in connection with the ideal bandpass filter. The *envelope* of a signal is defined mathematically as the magnitude of the analytic signal $x_p(t)$. The concept of an envelope will acquire more importance when we discuss modulation in Chapter 3.

EXAMPLE 2.30

In Section 2.6.12, (2.217), we showed that the impulse response of an ideal bandpass filter with bandwidth B , delay t_0 , and center frequency f_0 is given by

$$h_{\text{BP}}(t) = 2H_0 B \operatorname{sinc} [B(t - t_0)] \cos [\omega_0(t - t_0)] \quad (2.284)$$

Assuming that $B < f_0$, we can use the result of Example 2.29 to determine the Hilbert transform of $h_{\text{BP}}(t)$. The result is

$$\hat{h}_{\text{BP}}(t) = 2H_0 B \operatorname{sinc} [B(t - t_0)] \sin [\omega_0(t - t_0)] \quad (2.285)$$

Thus, the envelope is

$$\begin{aligned} |h_{\text{BP}}(t)| &= |x(t) + j\hat{x}(t)| & (2.286) \\ &= \sqrt{[x(t)]^2 + [\hat{x}(t)]^2} \\ &= \sqrt{\{2H_0 B \operatorname{sinc} [B(t - t_0)]\}^2 \{\cos^2 [\omega_0(t - t_0)] + \sin^2 [\omega_0(t - t_0)]\}} \end{aligned}$$

or

$$|h_{\text{BP}}(t)| = 2H_0 B \left| \operatorname{sinc} [B(t - t_0)] \right| \quad (2.287)$$

as shown in Figure 2.22(b) by the dashed lines. The envelope is obviously easy to identify if the signal is composed of a lowpass signal multiplied by a high-frequency sinusoid. Note, however, that the envelope is mathematically defined for any signal. ■

The spectrum of the analytic signal is also of interest. We will use it to advantage in Chapter 3 when we investigate single-sideband modulation. Since the analytic signal, from (2.283), is defined as

$$x_p(t) = x(t) + j\hat{x}(t)$$

it follows that the Fourier transform of $x_p(t)$ is

$$X_p(f) = X(f) + j \{-j \operatorname{sgn} (f) X(f)\} \quad (2.288)$$

where the term in braces is the Fourier transform of $\hat{x}(t)$. Thus,

$$X_p(f) = X(f) [1 + \operatorname{sgn} f] \quad (2.289)$$

or

$$X_p(f) = \begin{cases} 2X(f), & f > 0 \\ 0, & f < 0 \end{cases} \quad (2.290)$$

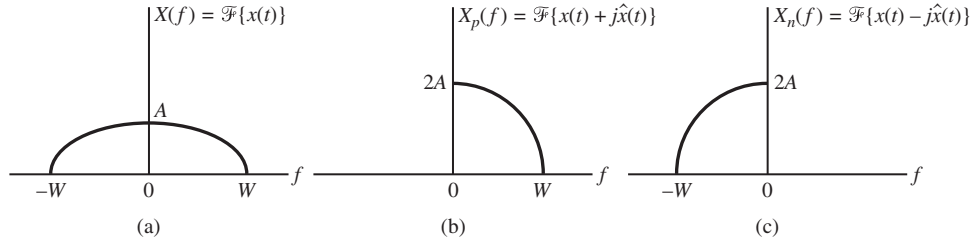
The subscript p is used to denote that the spectrum is nonzero only for positive frequencies.

Similarly, we can show that the signal

$$x_n(t) = x(t) - j\hat{x}(t) \quad (2.291)$$

is nonzero only for negative frequencies. Replacing $\hat{x}(t)$ by $-\hat{x}(t)$ in the preceding discussion results in

$$X_n(f) = X(f)[1 - \operatorname{sgn} f] \quad (2.292)$$

**Figure 2.29**

Spectra of analytic signals. (a) Spectrum of $x(t)$. (b) Spectrum of $x(t) + j\hat{x}(t)$. (c) Spectrum of $x(t) - j\hat{x}(t)$.

or

$$X_n(f) = \begin{cases} 0, & f > 0 \\ 2X(f), & f < 0 \end{cases} \quad (2.293)$$

These spectra are illustrated in Figure 2.30.

Two observations may be made at this point. First, if $X(f)$ is nonzero at $f = 0$, then $X_p(f)$ and $X_n(f)$ will be discontinuous at $f = 0$. Also, we should not be confused that $|X_n(f)|$ and $|X_p(f)|$ are not even, since the corresponding time-domain signals are not real.

2.8.4 Complex Envelope Representation of Bandpass Signals

If $X(f)$ in (2.288) corresponds to a signal with a bandpass spectrum, as shown in Figure 2.29(a), it then follows by (2.290) that $X_p(f)$ is just twice the positive frequency portion of $X(f) = \mathfrak{F}\{x(t)\}$, as shown in Figure 2.29(b). By the frequency-translation theorem, it follows that $x_p(t)$ can be written as

$$x_p(t) = \tilde{x}(t)e^{j2\pi f_0 t} \quad (2.294)$$

where $\tilde{x}(t)$ is a complex-valued lowpass signal (hereafter referred to as the *complex envelope*) and f_0 is a reference frequency chosen for convenience.¹⁸ The spectrum (assumed to be real for ease of plotting) of $\tilde{x}(t)$ is shown in Figure 2.29(c).

To find $\tilde{x}(t)$, we may proceed along one of two paths [note that simply taking the magnitude of (2.294) gives only $|\tilde{x}(t)|$ but not its argument]. First, using (2.283), we can find the analytic signal $x_p(t)$ and then solve (2.294) for $\tilde{x}(t)$. That is,

$$\tilde{x}(t) = x_p(t) e^{-j2\pi f_0 t} \quad (2.295)$$

Second, we can find $\tilde{x}(t)$ by using a frequency-domain approach to obtain $X(f)$, then scale its positive frequency components by a factor of 2 to give $X_p(f)$, and translate the resultant spectrum by f_0 Hz to the left. The inverse Fourier transform of this translated spectrum is then $\tilde{x}(t)$. For example, for the spectra shown in Figure 2.30, the complex envelope, using

¹⁸If the spectrum of $x_p(t)$ has a center of symmetry, a natural choice for f_0 would be this point of symmetry, but it need not be.

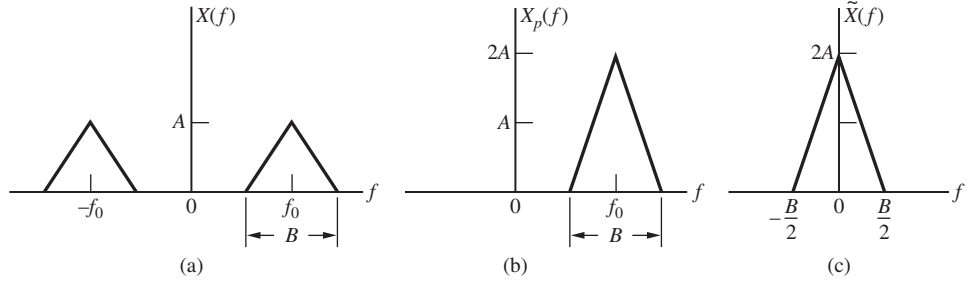


Figure 2.30 Spectra pertaining to the formation of a complex envelope of a signal $x(t)$. (a) A bandpass signal spectrum. (b) Twice the positive-frequency portion of $X(f)$ corresponding to $\mathfrak{F}[x(t) + j\hat{x}(t)]$. (c) Spectrum of $\tilde{x}(t)$.

Fig. 2.30(c), is

$$\tilde{x}(t) = \mathfrak{F}^{-1} \left[2A\Lambda \left(\frac{2f}{B} \right) \right] = AB \operatorname{sinc}^2(Bt) \quad (2.296)$$

The complex envelope is real in this case because the spectrum $X(f)$ is symmetrical around $f = f_0$.

Since $x_p(t) = x(t) + j\hat{x}(t)$, where $x(t)$ and $\hat{x}(t)$ are the real and imaginary parts, respectively, of $x_p(t)$, it follows from (2.294) that

$$x_p(t) = \tilde{x}(t)e^{j2\pi f_0 t} \triangleq x(t) + j\hat{x}(t) \quad (2.297)$$

or

$$x(t) = \operatorname{Re} [\tilde{x}(t)e^{j2\pi f_0 t}] \quad (2.298)$$

and

$$\hat{x}(t) = \operatorname{Im} [\tilde{x}(t)e^{j2\pi f_0 t}] \quad (2.299)$$

Thus, from (2.298), the real signal $x(t)$ can be expressed in terms of its complex envelope as

$$\begin{aligned} x(t) &= \operatorname{Re} [\tilde{x}(t)e^{j2\pi f_0 t}] \\ &= \operatorname{Re} [\tilde{x}(t)] \cos(2\pi f_0 t) - \operatorname{Im} [\tilde{x}(t)] \sin(2\pi f_0 t) \\ &= x_R(t) \cos(2\pi f_0 t) - x_I(t) \sin(2\pi f_0 t) \end{aligned} \quad (2.300)$$

where

$$\tilde{x}(t) \triangleq x_R(t) + jx_I(t) \quad (2.301)$$

The signals $x_R(t)$ and $x_I(t)$ are known as the *inphase* and *quadrature components* of $x(t)$.

EXAMPLE 2.31

Consider the real bandpass signal

$$x(t) = \cos(22\pi t) \quad (2.302)$$

Its Hilbert transform is

$$\hat{x}(t) = \sin(22\pi t) \quad (2.303)$$

so the corresponding analytic signal is

$$\begin{aligned} x_p(t) &= x(t) + j\hat{x}(t) \\ &= \cos(22\pi t) + j \sin(22\pi t) \\ &= e^{j22\pi t} \end{aligned} \quad (2.304)$$

In order to find the corresponding complex envelope, we need to specify f_0 , which, for the purposes of this example, we take as $f_0 = 10$ Hz. Thus, from (2.295), we have

$$\begin{aligned} \tilde{x}(t) &= x_p(t) e^{-j2\pi f_0 t} \\ &= e^{j22\pi t} e^{-j20\pi t} \\ &= e^{j2\pi t} \\ &= \cos(2\pi t) + j \sin(2\pi t) \end{aligned} \quad (2.305)$$

so that, from (2.301), we obtain

$$x_R(t) = \cos(2\pi t) \text{ and } x_I(t) = \sin(2\pi t) \quad (2.306)$$

Putting these into (2.300), we get

$$\begin{aligned} x(t) &= x_R(t) \cos(2\pi f_0 t) - x_I(t) \sin(2\pi f_0 t) \\ &= \cos(2\pi t) \cos(20\pi t) - \sin(2\pi t) \sin(20\pi t) \\ &= \cos(22\pi t) \end{aligned} \quad (2.307)$$

which is, not surprisingly, what we began with in (2.302). ■

2.8.5 Complex Envelope Representation of Bandpass Systems

Consider a bandpass system with impulse response $h(t)$, which is represented in terms of a complex envelope $\tilde{h}(t)$ as

$$h(t) = \operatorname{Re} \left[\tilde{h}(t) e^{j2\pi f_0 t} \right] \quad (2.308)$$

where $\tilde{h}(t) = h_R(t) + jh_I(t)$. Assume that the input is also bandpass with representation (2.298). The output, by the superposition integral, is

$$y(t) = x(t) * h(t) = \int_{-\infty}^{\infty} h(\lambda) x(t - \lambda) d\lambda \quad (2.309)$$

By Euler's theorem, we can represent $h(t)$ and $x(t)$ as

$$h(t) = \frac{1}{2} \tilde{h}(t) e^{j2\pi f_0 t} + \text{c.c.} \quad (2.310)$$

and

$$x(t) = \frac{1}{2} \tilde{x}(t) e^{j2\pi f_0 t} + \text{c.c.} \quad (2.311)$$

respectively, where c.c. stands for the complex conjugate of the immediately preceding term. Using these in (2.309), the output can be expressed as

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} \left[\frac{1}{2} \tilde{h}(\lambda) e^{j2\pi f_0 \lambda} + \text{c.c.} \right] \left[\frac{1}{2} \tilde{x}(t - \lambda) e^{j2\pi f_0 (t - \lambda)} + \text{c.c.} \right] d\lambda \\ &= \frac{1}{4} \int_{-\infty}^{\infty} \tilde{h}(\lambda) \tilde{x}(t - \lambda) d\lambda e^{j2\pi f_0 t} + \text{c.c.} \\ &\quad + \frac{1}{4} \int_{-\infty}^{\infty} \tilde{h}(\lambda) \tilde{x}^*(t - \lambda) e^{j4\pi f_0 \lambda} d\lambda e^{-j2\pi f_0 t} + \text{c.c.} \end{aligned} \quad (2.312)$$

The second pair of terms, $\frac{1}{4} \int_{-\infty}^{\infty} \tilde{h}(\lambda) \tilde{x}^*(t - \lambda) e^{j4\pi f_0 \lambda} d\lambda e^{-j2\pi f_0 t} + \text{c.c.}$, is approximately zero by virtue of the factor $e^{j4\pi f_0 \lambda} = \cos(4\pi f_0 \lambda) + j \sin(4\pi f_0 \lambda)$ in the integrand (\tilde{h} and \tilde{x} are slowly varying with respect to this complex exponential, and therefore, the integrand cancels to zero, half-cycle by half-cycle). Thus,

$$\begin{aligned} y(t) &\cong \frac{1}{4} \int_{-\infty}^{\infty} \tilde{h}(\lambda) \tilde{x}(t - \lambda) d\lambda e^{j2\pi f_0 t} + \text{c.c.} \\ &= \frac{1}{2} \text{Re} \left\{ \left[\tilde{h}(t) * \tilde{x}(t) \right] e^{j2\pi f_0 t} \right\} \triangleq \frac{1}{2} \text{Re} \left\{ \tilde{y}(t) e^{j2\pi f_0 t} \right\} \end{aligned} \quad (2.313)$$

where

$$\tilde{y}(t) = \tilde{h}(t) * \tilde{x}(t) = \mathfrak{F}^{-1} \left[\tilde{H}(f) \tilde{X}(f) \right] \quad (2.314)$$

in which $\tilde{H}(f)$ and $\tilde{X}(f)$ are the respective Fourier transforms of $\tilde{h}(t)$ and $\tilde{x}(t)$.

EXAMPLE 2.32

As an example of the application of (2.313), consider the input

$$x(t) = \Pi(t/\tau) \cos(2\pi f_0 t) \quad (2.315)$$

to a filter with impulse response

$$h(t) = \alpha e^{-\alpha t} u(t) \cos(2\pi f_0 t) \quad (2.316)$$

Using the complex envelope analysis just developed with $\tilde{x}(t) = \Pi(t/\tau)$ and $\tilde{h}(t) = \alpha e^{-\alpha t} u(t)$, we have as the complex envelope of the filter output

$$\begin{aligned} \tilde{y}(t) &= \Pi(t/\tau) * \alpha e^{-\alpha t} u(t) \\ &= \left[1 - e^{-\alpha(t+\tau/2)} \right] u(t + \tau/2) - \left[1 - e^{-\alpha(t-\tau/2)} \right] u(t - \tau/2) \end{aligned} \quad (2.317)$$

Multiplying this by $\frac{1}{2}e^{j2\pi f_0 t}$ and taking the real part results in the output of the filter in accordance with (2.313). The result is

$$y(t) = \frac{1}{2} \left\{ [1 - e^{-\alpha(t+\tau/2)}] u(t + \tau/2) - [1 - e^{-(t-\tau/2)}] u(t - \tau/2) \right\} \cos(2\pi f_0 t) \quad (2.318)$$

To check this result, we convolve (2.315) and (2.316) directly. The superposition integral becomes

$$\begin{aligned} y(t) &= x(t) * h(t) \\ &= \int_{-\infty}^{\infty} \Pi(\lambda/\tau) \cos(2\pi f_0 \lambda) \alpha e^{-\alpha(t-\lambda)} u(t-\lambda) \cos[2\pi f_0(t-\lambda)] d\lambda \end{aligned} \quad (2.319)$$

But

$$\cos(2\pi f_0 \lambda) \cos[2\pi f_0(t-\lambda)] = \frac{1}{2} \cos(2\pi f_0 t) + \frac{1}{2} \cos[2\pi f_0(t-2\lambda)] \quad (2.320)$$

so that the superposition integral becomes

$$\begin{aligned} y(t) &= \frac{1}{2} \int_{-\infty}^{\infty} \Pi(\lambda/\tau) \alpha e^{-\alpha(t-\lambda)} u(t-\lambda) d\lambda \cos(2\pi f_0 t) \\ &\quad + \frac{1}{2} \int_{-\infty}^{\infty} \Pi(\lambda/\tau) \alpha e^{-\alpha(t-\lambda)} u(t-\lambda) \cos[2\pi f_0(t-2\lambda)] d\lambda \end{aligned} \quad (2.321)$$

If $f_0^{-1} \ll \tau$ and $f_0^{-1} \ll \alpha^{-1}$, the second integral is approximately zero, so that we have only the first integral, which is $\Pi(\lambda/\tau)$ convolved with $\alpha e^{-\alpha t} u(t)$ and the result multiplied by $\frac{1}{2} \cos(2\pi f_0 t)$, which is the same as (2.318). ■

2.9 THE DISCRETE FOURIER TRANSFORM AND FAST FOURIER TRANSFORM

In order to compute the Fourier spectrum of a signal by means of a digital computer, the time-domain signal must be represented by sample values and the spectrum must be computed at a discrete number of frequencies. It can be shown that the following sum gives an approximation to the Fourier spectrum of a signal at frequencies $k/(NT_s)$, $k = 0, 1, \dots, N-1$:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi nk/N}, \quad k = 0, 1, \dots, N-1 \quad (2.322)$$

where $x_0, x_1, x_2, \dots, x_{N-1}$ are N sample values of the signal taken at T_s -second intervals for which the Fourier spectrum is desired. The sum (2.322) is called the *discrete Fourier transform (DFT)* of the sequence $\{x_n\}$. According to the sampling theorem, if the samples are spaced by T_s seconds, the spectrum repeats every $f_s = T_s^{-1}$ Hz. Since there are N frequency samples in this interval, it follows that the frequency resolution of (2.322) is $f_s/N = 1/(NT_s) \triangleq 1/T$. To obtain the sample sequence $\{x_n\}$ from the DFT sequence $\{X_k\}$, the sum

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, \quad k = 0, 1, 2, \dots, N-1 \quad (2.323)$$

is used. That (2.322) and (2.323) form a transform pair can be shown by substituting (2.322) into (2.323) and using the sum formula for a geometric series:

$$S_N \equiv \sum_{k=0}^{N-1} x^k = \begin{cases} \frac{1-x^N}{1-x}, & x \neq 1 \\ N, & x = 1 \end{cases} \quad (2.324)$$

As indicated above, the DFT and inverse DFT are approximations to the true Fourier spectrum of a signal $x(t)$ at the discrete set of frequencies $\{0, 1/T, 2/T, \dots, (N-1)/T\}$. The error can be small if the DFT and its inverse are applied properly to a signal. To indicate the approximations involved, we must visualize the spectrum of a sampled signal that is truncated to a finite number of sample values and whose spectrum is then sampled at a discrete number N of points. To see the approximations involved, we use the following Fourier-transform theorems:

1. The Fourier transform of an ideal sampling waveform (Example 2.14):

$$y_s(t) = \sum_{m=-\infty}^{\infty} \delta(t - mT_s) \longleftrightarrow f_s^{-1} \sum_{n=-\infty}^{\infty} \delta(f - nf_s), \quad f_s = T_s^{-1}$$

2. The Fourier transform of a rectangular window function:

$$\Pi(t/T) \longleftrightarrow T \operatorname{sinc}(fT)$$

3. The convolution theorem of Fourier transforms:

$$x_1(t) * x_2(t) \longleftrightarrow X_1(f)X_2(f)$$

4. The multiplication theorem of Fourier transforms:

$$x_1(t)x_2(t) \longleftrightarrow X_1(f) * X_2(f)$$

The approximations involved are illustrated by the following example.

EXAMPLE 2.33

An exponential signal is to be sampled, the samples truncated to a finite number, and the result represented by a finite number of samples of the Fourier spectrum of the sampled truncated signal. The continuous-time signal and its Fourier transform are

$$x(t) = e^{-|t|/\tau} \longleftrightarrow X(f) = \frac{2\tau}{1 + 2(\pi f \tau)^2} \quad (2.325)$$

This signal and its spectrum are shown in Figure 2.31(a). However, we are representing the signal by sample values spaced by T_s seconds, which entails multiplying the original signal by the ideal sampling waveform $y_s(t)$, given by (2.114). The resulting spectrum of this sampled signal is the convolution of $X(f)$ with the Fourier transform of $y_s(t)$, given by (2.119), which is $Y_s(f) = f_s \sum_{n=-\infty}^{\infty} \delta(f - nf_s)$. The result of this convolution in the frequency domain is

$$X_s(f) = f_s \sum_{n=-\infty}^{\infty} \frac{2\tau}{1 + [2\pi\tau(f - f_s)]^2} \quad (2.326)$$

The resulting sampled signal and its spectrum are shown in Figure 2.31(b).

In calculating the DFT, only a T -second chunk of $x(t)$ can be used (N samples spaced by $T_s = T/N$). This means that the sampled time-domain signal is effectively multiplied by a window function $\Pi(t/T)$.

In the frequency domain, this corresponds to convolution with the Fourier transform of the rectangular window function, which is $T \text{sinc}(fT)$. The resulting windowed, sampled signal and its spectrum are sketched in Figure 2.31(c). Finally, the spectrum is available only at N discrete frequencies separated by the reciprocal of the window duration $1/T$. This corresponds to convolution in the time domain with

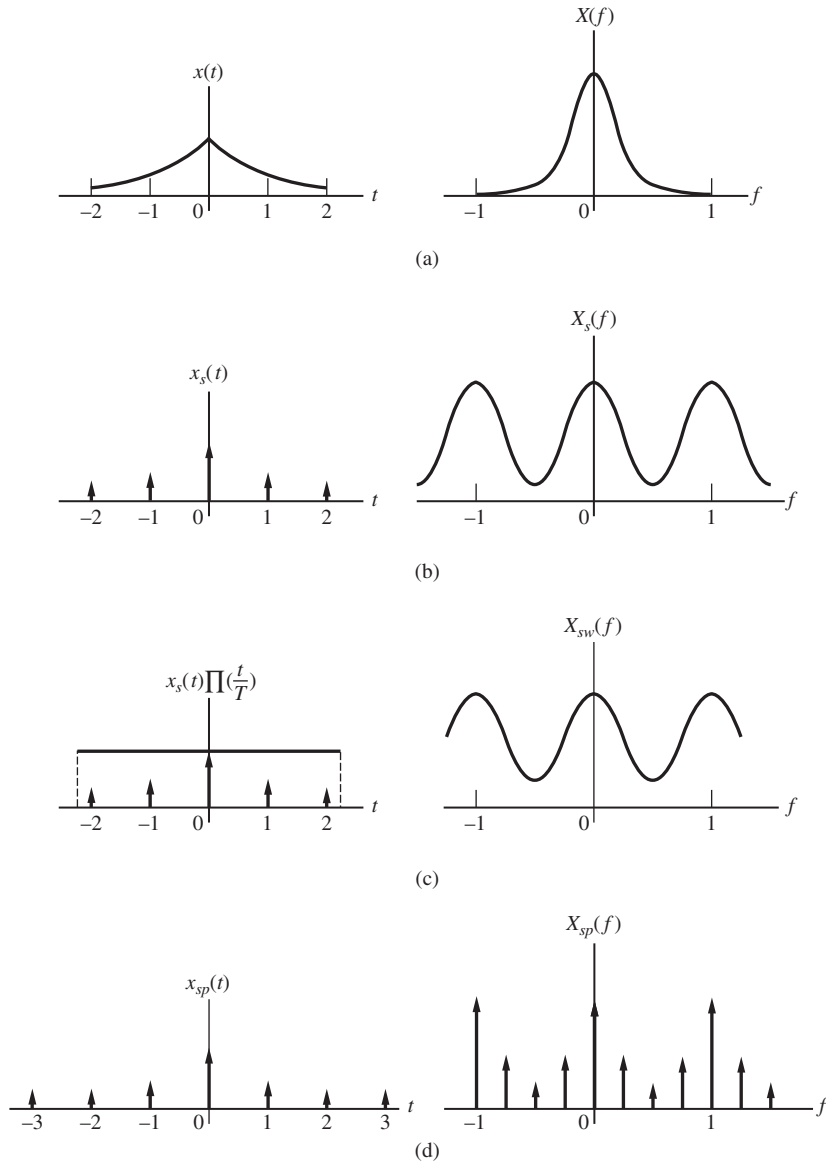


Figure 2.31

Signals and spectra illustrating the computation of the DFT. (a) Signal to be sampled and its spectrum ($\tau = 1$ s). (b) Sampled signal and its spectrum ($f_s = 1$ Hz). (c) Windowed, sampled signal and its spectrum ($T = 4^+$ s). (d) Sampled signal spectrum and corresponding periodic repetition of the sampled, windowed signal.

a sequence of delta functions. The resulting signal and spectrum are shown in Figure 2.31(d). It can be seen that unless one is careful, there is indeed a considerable likelihood that the DFT spectrum will look nothing like the spectrum of the original continuous-time signal. Means for minimizing these errors are discussed in several references on the subject.¹⁹

A little thought will indicate that to compute the complete DFT spectrum of a signal, approximately N^2 complex multiplications are required in addition to a number of complex additions. It is possible to find algorithms that allow the computation of the DFT spectrum of a signal using only approximately $N \log_2 N$ complex multiplications, which gives significant computational savings for N large. Such algorithms are referred to as fast *Fourier-transform (FFT) algorithms*. Two main types of FFT algorithms are those based on *decimation in time (DIT)* and those based on *decimation in frequency (DIF)*.

Fortunately, FFT algorithms are included in most computer mathematics packages such as MATLABTM, so we do not have to go to the trouble of writing our own FFT programs although it is an instructive exercise to do so. The following computer example computes the FFT of a sampled double-sided exponential pulse and compares spectra of the continuous-time and sampled pulses.

COMPUTER EXAMPLE 2.3

The MATLAB program given below computes the fast Fourier transform (FFT) of a double-sided exponentially decaying signal truncated to $-15.5 \leq t \leq 15.5$ sampled each $T_s = 1$ s. The periodicity property of the FFT means that the resulting FFT coefficients correspond to a waveform that is the periodic extension of this exponential waveform. The frequency extent of the FFT is $[0, f_s(1 - 1/N)]$ with the frequencies above $f_s/2$ corresponding to negative frequencies. Results are shown in Fig. 2.32.

```
% file: c2ce3
%
clf
tau = 2;
Ts = 1;
fs = 1/Ts;
ts = -15.5:Ts:15.5;
N = length(ts);
fss = 0:fs/N:fs-fs/N;
xss = exp(-abs(ts)/tau);
Xss = fft(xss);
t = -15.5:0.01:15.5;
f = 0:.01:fs-fs/N;
X = 2*fs*tau./(1+(2*pi*f*tau).^2);
subplot(2,1,1), stem(ts, xss)
hold on
subplot(2,1,1), plot(t, exp(-abs(t)/tau), '--'), xlabel('t, s'), yla-
bel('Signal & samples'), ...
legend('x(nTs)', 'x(t)')
subplot(2,1,2), stem(fss, abs(Xss))
hold on
subplot(2,1,2), plot(f, X, '--'), xlabel('f, Hz'), ylabel('FFT and
Fourier transform')
legend('|X_k|', '|X(f)|')
% End of script file
```

¹⁹Ziemer, Tranter, and Fannin (1998), Chapter 10.

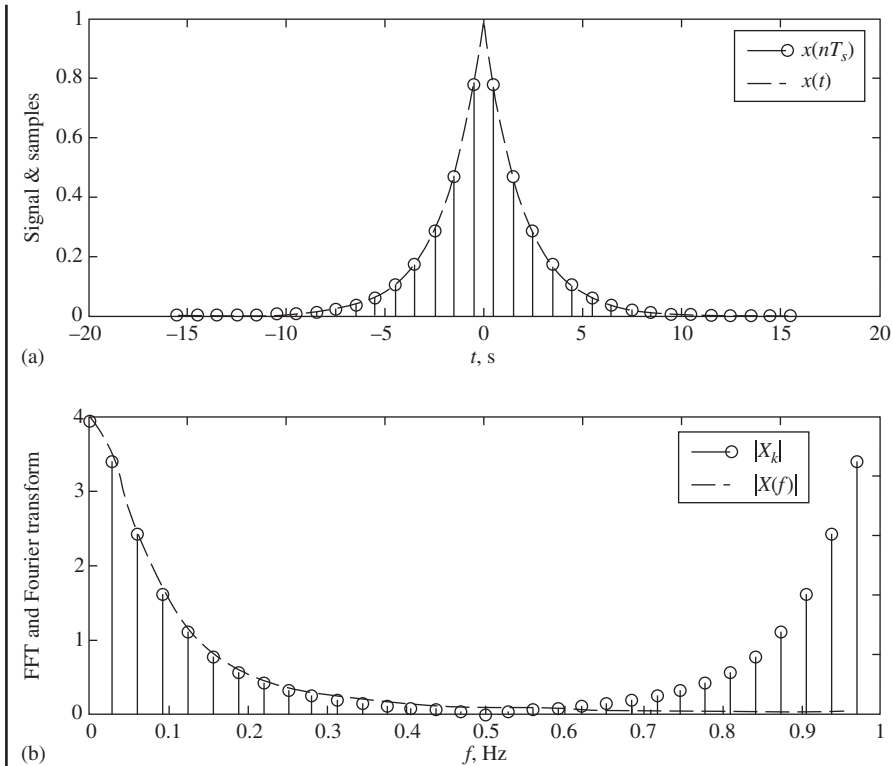


Figure 2.32

(a) $x(t) = \exp(-|t|/\tau)$ and samples taken each $T_s = 1$ s for $\tau = 2$ s; (b) Magnitude of the 32-point FFT of the sampled signal compared with the Fourier transform of $x(t)$. The spectral plots deviate from each other around $f_s/2$ most due to aliasing.

Further Reading

Bracewell (1986) is a text concerned exclusively with Fourier theory and applications. Ziemer, Tranter, and Fannin (1998) and Kamen and Heck (2007) are devoted to continuous- and discrete-time signal and system theory and provide background for this chapter. More elementary books are McClellan, Schafer, and Yoder (2003), Mersereau and Jackson (2006), and Wickert (2013).

Summary

1. Two general classes of signals are deterministic and random. The former can be expressed as completely known functions of time, whereas the amplitudes of random signals must be described probabilistically.

2. A periodic signal of period T_0 is one for which $x(t) = x(t + T_0)$, all t .

3. A single-sided spectrum for a rotating phasor $\tilde{x}(t) = Ae^{j(2\pi f_0 t + \theta)}$ shows A (amplitude) and θ (phase) versus f

(frequency). The real, sinusoidal signal corresponding to this phasor is obtained by taking the real part of $\tilde{x}(t)$. A double-sided spectrum results if we think of forming $x(t) = \frac{1}{2}\tilde{x}(t) + \frac{1}{2}\tilde{x}^*(t)$. Graphs of amplitude and phase (two plots) of this rotating phasor sum versus f are known as two-sided amplitude and phase spectra, respectively. Such spectral plots are referred to as frequency-domain representations of the signal $A \cos(2\pi f_0 t + \theta)$.

4. The unit impulse function, $\delta(t)$, can be thought of as a zero-width, infinite-height pulse with unity area. The sifting property, $\int_{-\infty}^{\infty} x(\lambda)\delta(\lambda - t_0) d\lambda = x(t_0)$, where $x(t)$ is continuous at $t = t_0$, is a generalization of the defining relation for a unit impulse. The unit step function, $u(t)$, is the integral of a unit impulse.

5. A signal $x(t)$ for which $E = \int_{-\infty}^{\infty} |x(t)|^2 dt$ is finite is called an *energy signal*. If $x(t)$ is such that $P = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |x(t)|^2 dt$ is finite, the signal is known as a *power signal*. Example signals may be either or neither.

6. The complex exponential Fourier series is $x(t) = \sum_{n=-\infty}^{\infty} X_n \exp(j2\pi n f_0 t)$ where $f_0 = 1/T_0$ and $(t_0, t_0 + T_0)$ is the expansion interval. The expansion coefficients are given by $X_n = \frac{1}{T_0} \int_{t_0}^{t_0+T_0} x(t) \exp(-j2\pi n f_0 t) dt$. If $x(t)$ is periodic with period T_0 , the exponential Fourier series represents $x(t)$ exactly for all t , except at points of discontinuity where the Fourier sum converges to the mean of the right- and left-handed limits of the signal at the discontinuity.

7. For exponential Fourier series of real signals, the Fourier coefficients obey $X_n = X_{-n}^*$, which implies that $|X_n| = |X_{-n}|$ and $\angle X_n = -\angle X_{-n}$. Plots of $|X_n|$ and $\angle X_n$ versus $n f_0$ are referred to as the discrete, double-sided amplitude and phase spectra, respectively, of $x(t)$. If $x(t)$ is real, the amplitude spectrum is even and the phase spectrum is odd as functions of $n f_0$.

8. Parseval's theorem for periodic signals is

$$\frac{1}{T_0} \int_{T_0} |x(t)|^2 dt = \sum_{n=-\infty}^{\infty} |X_n|^2$$

9. The Fourier transform of a signal $x(t)$ is

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt$$

and the inverse Fourier transform is

$$x(t) = \int_{-\infty}^{\infty} X(f) e^{j2\pi f t} df$$

For real signals, $|X(f)| = |X(-f)|$ and $\angle X(f) = -\angle X(-f)$.

10. Plots of $|X(f)|$ and $\angle X(f)$ versus f are referred to as the double-sided amplitude and phase spectra, respectively, of $x(t)$. As functions of frequency, the amplitude spectrum of a real signal is even and its phase spectrum is odd.

11. The energy of a signal is

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df$$

This is known as *Rayleigh's energy theorem*. The energy spectral density of a signal is $G(f) = |X(f)|^2$. It is the density of energy with frequency of the signal.

12. The convolution of two signals, $x_1(t)$ and $x_2(t)$, is

$$\begin{aligned} x(t) &= x_1 * x_2 = \int_{-\infty}^{\infty} x_1(\lambda) x_2(t - \lambda) d\lambda \\ &= \int_{-\infty}^{\infty} x_1(t - \lambda) x_2(\lambda) d\lambda \end{aligned}$$

The convolution theorem of Fourier transforms states that $X(f) = X_1(f)X_2(f)$, where $X(f)$, $X_1(f)$, and $X_2(f)$ are the Fourier transforms of $x(t)$, $x_1(t)$, and $x_2(t)$, respectively.

13. The Fourier transform of a periodic signal can be obtained formally by Fourier-transforming its exponential Fourier series term by term using $Ae^{j2\pi f_0 t} \leftrightarrow A\delta(f - f_0)$, even though, mathematically speaking, Fourier transforms of power signals do not exist. A more convenient approach is to convolve a pulse-type signal, $p(t)$, with the ideal sampling waveform to get a periodic signal of the form $x(t) = p(t) * \sum_{m=-\infty}^{\infty} \delta(t - mT_s)$; it follows that its Fourier transform is $X(f) = \sum_{n=-\infty}^{\infty} f_s P(nf_s) \delta(f - nf_s)$ where $P(f)$ is the Fourier transform of $p(t)$ and $f_s = 1/T_s$. It follows that the Fourier coefficients are $X_n = f_s P(nf_s)$.

14. The power spectrum $S(f)$ of a power signal $x(t)$ is a real, even, nonnegative function that integrates to give total average power: $\langle x^2(t) \rangle = \int_{-\infty}^{\infty} S(f) df$ where $\langle w(t) \rangle \triangleq \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T w(t) dt$. The time-average autocorrelation function of a power signal is defined as $R(\tau) = \langle x(t)x(t + \tau) \rangle$. The Wiener-Khinchine theorem states that $S(f)$ and $R(\tau)$ are Fourier-transform pairs.

15. A linear system, denoted operationally as $\mathcal{H}(\cdot)$, is one for which superposition holds; that is, if $y_1 = \mathcal{H}(x_1)$ and $y_2 = \mathcal{H}(x_2)$, then $\mathcal{H}(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 y_1 + \alpha_2 y_2$, where x_1 and x_2 are inputs, y_1 and y_2 are outputs (the time variable t is suppressed for simplicity), and α_1 and α_2 are arbitrary constants. A system is fixed, or time-invariant, if, given $y(t) = \mathcal{H}[x(t)]$, the input $x(t - t_0)$ results in the output $y(t - t_0)$.

16. The impulse response $h(t)$ of a linear time-invariant (LTI) system is its response to an impulse applied at $t = 0$: $h(t) = \mathcal{H}[\delta(t)]$. The output of an LTI system to an input $x(t)$ is given by $y(t) = h(t) * x(t) = \int_{-\infty}^{\infty} h(\tau) x(t - \tau) d\tau$.

17. A *causal system* is one that does not anticipate its input. For such an LTI system, $h(t) = 0$ for $t < 0$. A *stable system* is one for which every bounded input results in a bounded output. An LTI system is stable if and only if $\int_{-\infty}^{\infty} |h(t)| dt < \infty$.

18. The frequency response function $H(f)$ of an LTI system is the Fourier transform of $h(t)$. The Fourier transform of the system output $y(t)$ due to an input $x(t)$ is $Y(f) = H(f)X(f)$, where $X(f)$ is the Fourier transform of the input. $|H(f)| = |H(-f)|$ is called the *amplitude response* of the system and $\angle H(f) = -\angle H(-f)$ is called the *phase response*.

19. For a fixed linear system with a periodic input, the Fourier coefficients of the output are given by $Y_n = H(nf_0)X_n$, where $\{X_n\}$ represents the Fourier coefficients of the input.

20. Input and output spectral densities for a fixed linear system are related by

$$G_y(f) = |H(f)|^2 G_x(f) \quad (\text{energy signals})$$

$$S_y(f) = |H(f)|^2 S_x(f) \quad (\text{power signals})$$

21. A system is distortionless if its output looks like its input except for a time delay and amplitude scaling: $y(t) = H_0 x(t - t_0)$. The frequency response function of a distortionless system is $H(f) = H_0 e^{-j2\pi f t_0}$. Such a system's amplitude response is $|H(f)| = H_0$ and its phase response is $\angle H(f) = -2\pi f t_0$ over the band of frequencies occupied by the input. Three types of distortion that a system may introduce are amplitude, phase (or delay), and nonlinear, depending on whether $|H(f)| \neq \text{constant}$, $\angle H(f) \neq -\text{constant} \times f$, or the system is nonlinear, respectively. Two other important properties of a linear system are the group and phase delays. These are defined by

$$T_g(f) = -\frac{1}{2\pi} \frac{d\theta(f)}{df} \quad \text{and} \quad T_p(f) = -\frac{\theta(f)}{2\pi f}$$

respectively, in which $\theta(f)$ is the phase response of the LTI system. Phase distortionless systems have equal group and phase delays (constant).

22. Ideal filters are convenient in communication system analysis, even though they are noncausal. Three types of ideal filters are lowpass, bandpass, and highpass. Throughout their passbands, ideal filters have constant amplitude response and linear phase response. Outside their passbands, ideal filters perfectly reject all spectral components of the input.

23. Approximations to ideal filters are Butterworth, Chebyshev, and Bessel filters. The first two are attempts at approximating the amplitude response of an ideal filter, and the latter is an attempt to approximate the linear phase response of an ideal filter.

24. An inequality relating the duration T of a pulse and its single-sided bandwidth W is $W \geq 1/(2T)$. Pulse risetime T_R and signal bandwidth are related approximately by $W = 1/(2T_R)$. These relationships hold for the lowpass case. For bandpass filters and signals, the required bandwidth is doubled, and the risetime is that of the envelope of the signal.

25. The sampling theorem for lowpass signals of bandwidth W states that a signal can be perfectly recovered by lowpass filtering from sample values taken at a rate of $f_s > 2W$ samples per second. The spectrum of an impulse-sampled signal is

$$X_\delta(f) = f_s \sum_{n=-\infty}^{\infty} X(f - nf_s)$$

where $X(f)$ is the spectrum of the original signal. For bandpass signals, lower sampling rates than specified by the lowpass sampling theorem may be possible.

26. The Hilbert transform $\hat{x}(t)$ of a signal $x(t)$ corresponds to a -90° phase shift of all the signal's positive-frequency components. Mathematically,

$$\hat{x}(t) = \int_{-\infty}^{\infty} \frac{x(\lambda)}{\pi(t - \lambda)} d\lambda$$

In the frequency domain, $\hat{X}(f) = -j \operatorname{sgn}(f) X(f)$, where $\operatorname{sgn}(f)$ is the signum function, $X(f) = \mathfrak{F}[x(t)]$, and $\hat{X}(f) = \mathfrak{F}[\hat{x}(t)]$. The Hilbert transform of $\cos \omega_0 t$ is $\sin \omega_0 t$, and the Hilbert transform of $\sin \omega_0 t$ is $-\cos \omega_0 t$. The power (or energy) in a signal and its Hilbert transform are equal. A signal and its Hilbert transform are orthogonal in the range $(-\infty, \infty)$. If $m(t)$ is a lowpass signal and $c(t)$ is a highpass signal with nonoverlapping spectra,

$$m(t)\widehat{c}(t) = m(t)\hat{c}(t)$$

The Hilbert transform can be used to define the analytic signal

$$z(t) = x(t) \pm j\hat{x}(t)$$

The magnitude of the analytic signal, $|z(t)|$, is the envelope of the real signal $x(t)$. The Fourier transform of an analytic signal, $Z(f)$, is identically zero for $f < 0$ or $f > 0$, respectively, depending on whether the $+$ sign or $-$ sign is chosen for the imaginary part of $z(t)$.

27. The complex envelope $\tilde{x}(t)$ of a bandpass signal is defined by

$$x(t) + j\hat{x}(t) = \tilde{x}(t)e^{j2\pi f_0 t}$$

where f_0 is the reference frequency for the signal. Similarly, the complex envelope $\tilde{h}(t)$ of the impulse response of a bandpass system is defined by

$$h(t) + j\hat{h}(t) = \tilde{h}(t)e^{j2\pi f_0 t}$$

The complex envelope of the bandpass system output is conveniently obtained in terms of the complex envelope of the output, which can be found from either of the operations

$$\tilde{y}(t) = \tilde{h}(t) * \tilde{x}(t)$$

or

$$\tilde{y}(t) = \mathfrak{F}^{-1} \left[\tilde{H}(f)\tilde{X}(f) \right]$$

where $\tilde{H}(f)$ and $\tilde{X}(f)$ are the Fourier transforms of $\tilde{h}(t)$ and $\tilde{x}(t)$, respectively. The actual (real) output is then given by

$$y(t) = \frac{1}{2} \operatorname{Re} \left[\tilde{y}(t)e^{j2\pi f_0 t} \right]$$

28. The discrete Fourier transform (DFT) of a signal sequence $\{x_n\}$ is defined as

$$X_k = \sum_{n=0}^{N-1} x_n e^{j2\pi nk/N} = \text{DFT} \left[\{x_n\} \right], \quad k = 0, 1, \dots, N-1$$

and the inverse DFT can be found from

$$x_n = \frac{1}{N} \text{DFT} \left[\{X_k\} \right]^*, \quad k = 0, 1, \dots, N-1$$

The DFT can be used to digitally compute spectra of sampled signals and to approximate operations carried out by the normal Fourier transform, for example, filtering.

Drill Problems

2.1 Find the fundamental periods of the following signals:

- (a) $x_1(t) = 10 \cos(5\pi t)$
- (b) $x_2(t) = 10 \cos(5\pi t) + 2 \sin(7\pi t)$
- (c) $x_3(t) = 10 \cos(5\pi t) + 2 \sin(7\pi t) + 3 \cos(6.5\pi t)$
- (d) $x_4(t) = \exp(j6\pi t)$
- (e) $x_5(t) = \exp(j6\pi t) + \exp(-j6\pi t)$
- (f) $x_6(t) = \exp(j6\pi t) + \exp(j7\pi t)$

2.2 Plot the double-sided amplitude and phase spectra of the periodic signals given in Drill Problem 2.1.

2.3 Plot the single-sided amplitude and phase spectra of the periodic signals given in Drill Problem 2.1.

2.4 Evaluate the following integrals:

- (a) $I_1 = \int_{-10}^{10} u(t) dt$
- (b) $I_2 = \int_{-10}^{10} \delta(t-1)u(t) dt$
- (c) $I_3 = \int_{-10}^{10} \delta(t+1)u(t) dt$
- (d) $I_4 = \int_{-10}^{10} \delta(t-1)t^2 dt$
- (e) $I_5 = \int_{-10}^{10} \delta(t+1)t^2 dt$
- (f) $I_6 = \int_{-10}^{10} t^2 u(t-1) dt$

2.5 Find the powers and energies of the following signals (0 and ∞ are possible answers):

- (a) $x_1(t) = 2u(t)$
- (b) $x_2(t) = 3\Pi\left(\frac{t-1}{2}\right)$
- (c) $x_3(t) = 2\Pi\left(\frac{t-3}{4}\right)$
- (d) $x_4(t) = \cos(2\pi t)$
- (e) $x_5(t) = \cos(2\pi t)u(t)$
- (f) $x_6(t) = \cos^2(2\pi t) + \sin^2(2\pi t)$

2.6 Tell whether or not the following can be Fourier coefficients of real signals (give reasons for your answers):

- (a) $X_1 = 1 + j$; $X_{-1} = 1 - j$; all other Fourier coefficients are 0
- (b) $X_1 = 1 + j$; $X_{-1} = 2 - j$; all other Fourier coefficients are 0
- (c) $X_1 = \exp(-j\pi/2)$; $X_{-1} = \exp(j\pi/2)$; all other Fourier coefficients are 0
- (d) $X_1 = \exp(j3\pi/2)$; $X_{-1} = \exp(j\pi/2)$; all other Fourier coefficients are 0
- (e) $X_1 = \exp(j3\pi/2)$; $X_{-1} = \exp(j5\pi/2)$; all other Fourier coefficients are 0

2.7 By invoking uniqueness of the Fourier series, give the complex exponential Fourier series coefficients for the following signals:

- (a) $x_1(t) = 1 + \cos(2\pi t)$
 (b) $x_2(t) = 2 \sin(2\pi t)$
 (c) $x_3(t) = 2 \cos(2\pi t) + 2 \sin(2\pi t)$
 (d) $x_4(t) = 2 \cos(2\pi t) + 2 \sin(4\pi t)$
 (e) $x_5(t) = 2 \cos(2\pi t) + 2 \sin(4\pi t) + 3 \cos(6\pi t)$

2.8 Tell whether the following statements are true or false and why:

- (a) A triangular wave has only odd harmonics in its Fourier series.
 (b) The spectral content of a pulse train has more higher-frequency content the longer the pulse width.
 (c) A full rectified sine wave has a fundamental frequency, which is half that of the original sinusoid that was rectified.
 (d) The harmonics of a square wave decrease faster with the harmonic number n than those of a triangular wave.
 (e) The delay of a pulse train affects its amplitude spectrum.
 (f) The amplitude spectra of a half-rectified sine wave and a half-rectified cosine wave are identical.

2.9 Given the Fourier-transform pairs $\Pi(t) \leftrightarrow \text{sinc}(f)$ and $\Lambda(t) \leftrightarrow \text{sinc}^2(f)$, use appropriate Fourier-transform theorems to find Fourier transforms of the following signals. Tell which theorem(s) you used in each case. Sketch signals and transforms.

- (a) $x_1(t) = \Pi(2t)$
 (b) $x_2(t) = \text{sinc}^2(4t)$
 (c) $x_3(t) = \Pi(2t) \cos(6\pi t)$
 (d) $x_4(t) = \Lambda\left(\frac{t-3}{2}\right)$
 (e) $x_5(t) = \Pi(2t) \star \Pi(2t)$
 (f) $x_6(t) = \Pi(2t) \exp(j4\pi t)$
 (g) $x_7(t) = \Pi\left(\frac{t}{2}\right) + \Lambda(t)$
 (h) $x_8(t) = \frac{d\Lambda(t)}{dt}$
 (i) $x_9(t) = \Pi\left(\frac{t}{2}\right) \Lambda(t)$

2.10 Obtain the Fourier transform of the signal $x(t) = \sum_{m=-\infty}^{\infty} \Lambda(t - 3m)$. Sketch the signal and its transform.

2.11 Obtain the power spectral densities corresponding to the autocorrelation functions given below. Verify in each case that the power spectral density integrates to the

total average power [i.e., $R(0)$]. Provide a sketch of each autocorrelation function and corresponding power spectral density.

- (a) $R_1(\tau) = 3\Lambda(\tau/2)$
 (b) $R_2(\tau) = 2 \cos(4\pi\tau)$
 (c) $R_3(\tau) = 2\Lambda(\tau/2) \cos(4\pi\tau)$
 (d) $R_4(\tau) = \exp(-2|\tau|)$
 (e) $R_5(\tau) = 1 + \cos(2\pi\tau)$

2.12 Obtain the impulse response of a system with frequency response function $H(f) = 2/(3 + j2\pi f) + 1/(2 + j2\pi f)$. Plot the impulse response and the amplitude and phase responses.

2.13 Tell whether or not the following systems are (1) stable and (2) causal. Give reasons for your answers.

- (a) $h_1(t) = 3/(4 + |t|)$
 (b) $H_2(f) = 1 + j2\pi f$
 (c) $H_3(f) = 1/(1 + j2\pi f)$
 (d) $h_4(t) = \exp(-2|t|)$
 (e) $h_5(t) = [2 \exp(-3t) + \exp(-2t)] u(t)$

2.14 Find the phase and group delays for the following systems.

- (a) $h_1(t) = \exp(-2t) u(t)$
 (b) $H_2(f) = 1 + j2\pi f$
 (c) $H_3(f) = 1/(1 + j2\pi f)$
 (d) $h_4(t) = 2t \exp(-3t) u(t)$

2.15 A filter has frequency response function

$$H(f) = \left[\Pi\left(\frac{f}{30}\right) + \Pi\left(\frac{f}{10}\right) \right] \exp[-j\pi f \Pi(f/15)/20]$$

The input is $x(t) = 2 \cos(2\pi f_1 t) + \cos(2\pi f_2 t)$. For the values of f_1 and f_2 given below tell whether there is (1) no distortion, (2) amplitude distortion, (3) phase or delay distortion, or (4) both amplitude and phase (delay) distortion.

- (a) $f_1 = 2$ Hz and $f_2 = 4$ Hz
 (b) $f_1 = 2$ Hz and $f_2 = 6$ Hz
 (c) $f_1 = 2$ Hz and $f_2 = 8$ Hz
 (d) $f_1 = 6$ Hz and $f_2 = 7$ Hz
 (e) $f_1 = 6$ Hz and $f_2 = 8$ Hz
 (f) $f_1 = 8$ Hz and $f_2 = 16$ Hz

2.16 A filter has input-output transfer characteristic given by $y(t) = x(t) + x^2(t)$. With the input $x(t) = \cos(2\pi f_1 t) + \cos(2\pi f_2 t)$ tell what frequency components will appear at the output. Which are distortion terms?

2.17 A filter has frequency response function $H(j2\pi f) = \frac{2}{-(2\pi f)^2 + j4\pi f + 1}$. Find its 10% to 90% risetime.

2.18 The signal $x(t) = \cos(2\pi f_1 t)$ is sampled at $f_s = 9$ samples per second. Give the lowest frequency present in the sampled signal spectrum for the following values of f_1 :

- (a) $f_1 = 2$ Hz
- (b) $f_1 = 4$ Hz
- (c) $f_1 = 6$ Hz
- (d) $f_1 = 8$ Hz

(e) $f_1 = 10$ Hz

(f) $f_1 = 12$ Hz

2.19 Give the Hilbert transforms of the following signals:

(a) $x_1(t) = \cos(4\pi t)$

(b) $x_2(t) = \sin(6\pi t)$

(c) $x_3(t) = \exp(j5\pi t)$

(d) $x_4(t) = \exp(-j8\pi t)$

(e) $x_5(t) = 2 \cos^2(4\pi t)$

(f) $x_6(t) = \cos(2\pi t) \cos(10\pi t)$

(g) $x_7(t) = 2 \sin(4\pi t) \cos(4\pi t)$

2.20 Obtain the analytic signal and complex envelope of the signal $x(t) = \cos(10\pi t)$, where $f_0 = 6$ Hz.

Problems

Section 2.1

2.1 Sketch the single-sided and double-sided amplitude and phase spectra of the following signals:

- (a) $x_1(t) = 10 \cos(4\pi t + \pi/8) + 6 \sin(8\pi t + 3\pi/4)$
- (b) $x_2(t) = 8 \cos(2\pi t + \pi/3) + 4 \cos(6\pi t + \pi/4)$
- (c) $x_3(t) = 2 \sin(4\pi t + \pi/8) + 12 \sin(10\pi t)$
- (d) $x_4(t) = 2 \cos(7\pi t + \pi/4) + 3 \sin(18\pi t + \pi/2)$
- (e) $x_5(t) = 5 \sin(2\pi t) + 4 \cos(5\pi t + \pi/4)$
- (f) $x_6(t) = 3 \cos(4\pi t + \pi/8) + 4 \sin(10\pi t + \pi/6)$

2.2 A signal has the double-sided amplitude and phase spectra shown in Figure 2.33. Write a time-domain expression for the signal.

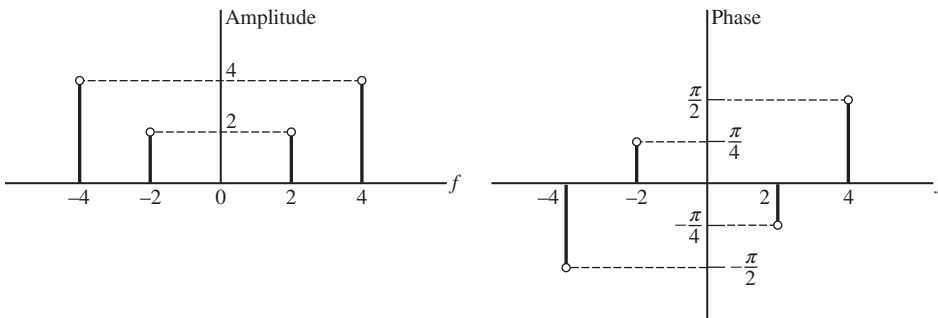


Figure 2.33

2.3 The sum of two or more sinusoids may or may not be periodic depending on the relationship of their separate frequencies. For the sum of two sinusoids, let the frequencies of the individual terms be f_1 and f_2 , respectively. For the sum to be periodic, f_1 and f_2 must be commensurable; i.e., there must be a number f_0 contained in each an integral number of times. Thus, if f_0 is the largest such number,

$$f_1 = n_1 f_0 \text{ and } f_2 = n_2 f_0$$

where n_1 and n_2 are integers; f_0 is the fundamental frequency. Which of the signals given below are periodic? Find the periods of those that are periodic.

(a) $x_1(t) = 2 \cos(2t) + 4 \sin(6\pi t)$

(b) $x_2(t) = \cos(6\pi t) + 7 \cos(30\pi t)$

- (c) $x_3(t) = \cos(4\pi t) + 9 \sin(21\pi t)$
 (d) $x_4(t) = 3 \sin(4\pi t) + 5 \cos(7\pi t) + 6 \sin(11\pi t)$
 (e) $x_5(t) = \cos(17\pi t) + 5 \cos(18\pi t)$
 (f) $x_6(t) = \cos(2\pi t) + 7 \sin(3\pi t)$
 (g) $x_7(t) = 4 \cos(7\pi t) + 5 \cos(11\pi t)$
 (h) $x_8(t) = \cos(120\pi t) + 3 \cos(377t)$
 (i) $x_9(t) = \cos(19\pi t) + 2 \sin(21\pi t)$
 (j) $x_{10}(t) = 5 \cos(6\pi t) + 6 \sin(7\pi t)$

2.4 Sketch the single-sided and double-sided amplitude and phase spectra of

- (a) $x_1(t) = 5 \cos(12\pi t - \pi/6)$
 (b) $x_2(t) = 3 \sin(12\pi t) + 4 \cos(16\pi t)$
 (c) $x_3(t) = 4 \cos(8\pi t) \cos(12\pi t)$

(Hint: Use an appropriate trigonometric identity.)

- (d) $x_4(t) = 8 \sin(2\pi t) \cos^2(5\pi t)$

(Hint: Use appropriate trigonometric identities.)

- (e) $x_5(t) = \cos(6\pi t) + 7 \cos(30\pi t)$
 (f) $x_6(t) = \cos(4\pi t) + 9 \sin(21\pi t)$
 (g) $x_7(t) = 2 \cos(4\pi t) + \cos(6\pi t) + 6 \sin(17\pi t)$

2.5

- (a) Show that the function $\delta_\epsilon(t)$ sketched in Figure 2.4(b) has unity area.
 (b) Show that

$$\delta_\epsilon(t) = \epsilon^{-1} e^{-t/\epsilon} u(t)$$

has unity area. Sketch this function for $\epsilon = 1, \frac{1}{2},$ and $\frac{1}{4}$. Comment on its suitability as an approximation for the unit impulse function.

- (c) Show that a suitable approximation for the unit impulse function as $\epsilon \rightarrow 0$ is given by

$$\delta_\epsilon(t) = \begin{cases} \epsilon^{-1} (1 - |t|/\epsilon), & |t| \leq \epsilon \\ 0, & \text{otherwise} \end{cases}$$

2.6 Use the properties of the unit impulse function given after (2.14) to evaluate the following relations.

- (a) $\int_{-\infty}^{\infty} [t^2 + \exp(-2t)] \delta(2t - 5) dt$
 (b) $\int_{-10^-}^{10^+} (t^2 + 1) \left[\sum_{n=-\infty}^{\infty} \delta(t - 5n) \right] dt$ (Note: 10^+ means just to the right of 10; -10^- means just

to the left of -10)

- (c) $10\delta(t) + A \frac{d\delta(t)}{dt} + 3 \frac{d^2\delta(t)}{dt^2} = B\delta(t) + 5 \frac{d\delta(t)}{dt} + C \frac{d^2\delta(t)}{dt^2}$; find $A, B,$ and C
 (d) $\int_{-2}^{11} [e^{-4\pi t} + \tan(10\pi t)] \delta(4t + 3) dt$
 (e) $\int_{-\infty}^{\infty} [\cos(5\pi t) + e^{-3t}] \frac{d^2\delta(t-2)}{dt^2} dt$

2.7 Which of the following signals are periodic and which are aperiodic? Find the periods of those that are periodic. Sketch all signals.

- (a) $x_a(t) = \cos(5\pi t) + \sin(7\pi t)$
 (b) $x_b(t) = \sum_{n=0}^{\infty} \Lambda(t - 2n)$
 (c) $x_c(t) = \sum_{n=-\infty}^{\infty} \Lambda(t - 2n)$
 (d) $x_d(t) = \sin(3t) + \cos(2\pi t)$
 (e) $x_e(t) = \sum_{n=-\infty}^{\infty} \Pi(t - 3n)$
 (f) $x_f(t) = \sum_{n=0}^{\infty} \Pi(t - 3n)$

2.8 Write the signal $x(t) = \cos(6\pi t) + 2 \sin(10\pi t)$ as

- (a) The real part of a sum of rotating phasors.
 (b) A sum of rotating phasors plus their complex conjugates.
 (c) From your results in parts (a) and (b), sketch the single-sided and double-sided amplitude and phase spectra of $x(t)$.

Section 2.2

2.9 Find the normalized power for each signal below that is a power signal and the normalized energy for each signal that is an energy signal. If a signal is neither a power signal nor an energy signal, so designate it. Sketch each signal (α is a positive constant).

- (a) $x_1(t) = 2 \cos(4\pi t + 2\pi/3)$
 (b) $x_2(t) = e^{-\alpha t} u(t)$
 (c) $x_3(t) = e^{\alpha t} u(-t)$
 (d) $x_4(t) = (\alpha^2 + t^2)^{-1/2}$
 (e) $x_5(t) = e^{-\alpha|t|}$
 (f) $x_6(t) = e^{-\alpha t} u(t) - e^{-\alpha(t-1)} u(t-1)$

2.10 Classify each of the following signals as an energy signal or as a power signal by calculating E , the energy, or P , the power ($A, B, \theta, \omega,$ and τ are positive constants).

- (a) $x_1(t) = A |\sin(\omega t + \theta)|$
 (b) $x_2(t) = A\tau / \sqrt{\tau + jt}, j = \sqrt{-1}$
 (c) $x_3(t) = Ate^{-t/\tau} u(t)$
 (d) $x_4(t) = \Pi(t/\tau) + \Pi(t/2\tau)$

102 Chapter 2 • Signal and Linear System Analysis

(e) $x_5(t) = \Pi(t/2) + \Lambda(t)$

(f) $x_6(t) = A \cos(\omega t) + B \sin(2\omega t)$

2.11 Find the powers of the following periodic signals. In each case provide a sketch of the signal and give its period.

(a) $x_1(t) = 2 \cos(4\pi t - \pi/3)$

(b) $x_2(t) = \sum_{n=-\infty}^{\infty} 3\Pi\left(\frac{t-4n}{2}\right)$

(c) $x_3(t) = \sum_{n=-\infty}^{\infty} \Lambda\left(\frac{t-6n}{2}\right)$

(d) $x_4(t) = \sum_{n=-\infty}^{\infty} \left[\Lambda(t - 4n) + \Pi\left(\frac{t-4n}{2}\right) \right]$

2.12 For each of the following signals, determine both the normalized energy and power. Tell which are power signals, which are energy signals, and which are neither. (Note: 0 and ∞ are possible answers.)

(a) $x_1(t) = 6e^{(-3+j4\pi)t} u(t)$

(b) $x_2(t) = \Pi[(t-3)/2] + \Pi\left(\frac{t-3}{6}\right)$

(c) $x_3(t) = 7e^{j6\pi t} u(t)$

(d) $x_4(t) = 2 \cos(4\pi t)$

(e) $x_5(t) = |t|$

(f) $x_6(t) = t^{-1/2} u(t-1)$

2.13 Show that the following are energy signals. Sketch each signal.

(a) $x_1(t) = \Pi(t/12) \cos(6\pi t)$

(b) $x_2(t) = e^{-|t|/3}$

(c) $x_3(t) = 2u(t) - 2u(t-8)$

(d) $x_4(t) = \int_{-\infty}^t u(\lambda) d\lambda - 2 \int_{-\infty}^{t-10} u(\lambda) d\lambda + \int_{-\infty}^{t-20} u(\lambda) d\lambda$

(Hint: Consider what the indefinite integral of a step function is first.)

2.14 Find the energies and powers of the following signals (note that 0 and ∞ are possible answers). Tell which are energy signals and which are power signals.

(a) $x_1(t) = \cos(10\pi t) u(t) u(2-t)$

(b) $x_2(t) = \sum_{n=-\infty}^{\infty} \Lambda\left(\frac{t-3n}{2}\right)$

(c) $x_3(t) = e^{-|t|} \cos(2\pi t)$

(d) $x_4(t) = \Pi\left(\frac{t}{2}\right) + \Lambda(t)$

Section 2.3

2.15 Using the uniqueness property of the Fourier series, find exponential Fourier series for the following signals (f_0 is an arbitrary frequency):

(a) $x_1(t) = \sin^2(2\pi f_0 t)$

(b) $x_2(t) = \cos(2\pi f_0 t) + \sin(4\pi f_0 t)$

(c) $x_3(t) = \sin(4\pi f_0 t) \cos(4\pi f_0 t)$

(d) $x_4(t) = \cos^3(2\pi f_0 t)$

(e) $x_5(t) = \sin(2\pi f_0 t) \cos^2(4\pi f_0 t)$

(f) $x_6(t) = \sin^2(3\pi f_0 t) \cos(5\pi f_0 t)$

(Hint: Use appropriate trigonometric identities and Euler's theorem.)

2.16 Expand the signal $x(t) = 2t^2$ in a complex exponential Fourier series over the interval $|t| \leq 2$. Sketch the signal to which the Fourier series converges for all t .

2.17 If $X_n = |X_n| \exp[j\angle X_n]$ are the Fourier coefficients of a real signal, $x(t)$, fill in all the steps to show that:

(a) $|X_n| = |X_{-n}|$ and $\angle X_n = -\angle X_{-n}$.

(b) X_n is a real, even function of n for $x(t)$ even.

(c) X_n is imaginary and an odd function of n for $x(t)$ odd.

(d) $x(t) = -x(t + T_0/2)$ (halfwave odd symmetry) implies that $X_n = 0$, n even.

2.18 Obtain the complex exponential Fourier series coefficients for the (a) pulse train, (b) half-rectified sinewave, (c) full-rectified sinewave, and (d) triangular waveform as given in Table 2.1.

2.19 Find the ratio of the power contained in a rectangular pulse train for $|nf_0| \leq \tau^{-1}$ to the total power for each of the following cases:

(a) $\tau/T_0 = \frac{1}{2}$ (c) $\tau/T_0 = \frac{1}{10}$

(b) $\tau/T_0 = \frac{1}{5}$ (d) $\tau/T_0 = \frac{1}{20}$

(Hint: You can save work by noting the spectra are even about $f = 0$.)

2.20

(a) If $x(t)$ has the Fourier series

$$x(t) = \sum_{n=-\infty}^{\infty} X_n e^{j2\pi n f_0 t}$$

and $y(t) = x(t - t_0)$, show that

$$Y_n = X_n e^{-j2\pi n f_0 t_0}$$

where the Y_n 's are the Fourier coefficients for $y(t)$.

(b) Verify the theorem proved in part (a) by examining the Fourier coefficients for $x(t) = \cos(\omega_0 t)$ and $y(t) = \sin(\omega_0 t)$.

(Hint: What delay, t_0 , will convert a cosine into a sine. Use the uniqueness property to write down the corresponding Fourier series.)

2.21 Use the Fourier series expansions of periodic square wave and triangular wave signals to find the sum of the following series:

(a) $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$

(b) $1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \dots$

(Hint: Write down the Fourier series in each case and evaluate it for a particular, appropriately chosen value of t .)

2.22 Using the results given in Table 2.1 for the Fourier coefficients of a pulse train, plot the double-sided amplitude and phase spectra for the waveforms shown in Figure 2.34.

(Hint: Note that $x_b(t) = -x_a(t) + A$. How is a sign change and DC level shift manifested in the spectrum of the waveform?)

2.23

(a) Plot the single-sided and double-sided amplitude and phase spectra of the square wave shown in Figure 2.35(a).

(b) Obtain an expression relating the complex exponential Fourier series coefficients of the triangular waveform shown in Figure 2.35(b) and those of $x_a(t)$ shown in Figure 2.35(a).

(Hint: Note that $x_a(t) = K[dx_b(t)/dt]$, where K is an appropriate scale change.)

(c) Plot the double-sided amplitude and phase spectra for $x_b(t)$.

Section 2.4

2.24 Sketch each signal given below and find its Fourier transform. Plot the amplitude and phase spectra of each signal (A and τ are positive constants).

(a) $x_1(t) = A \exp(-t/\tau) u(t)$

(b) $x_2(t) = A \exp(t/\tau) u(-t)$

(c) $x_3(t) = x_1(t) - x_2(t)$

(d) $x_4(t) = x_1(t) + x_2(t)$. Does the result check with the answer found using Fourier-transform tables?

(e) $x_5(t) = x_1(t - 5)$

(f) $x_6(t) = x_1(t) - x_1(t - 5)$

2.25

(a) Use the Fourier transform of

$$x(t) = \exp(-\alpha t) u(t) - \exp(\alpha t) u(-t)$$

where $\alpha > 0$ to find the Fourier transform of the signum function defined as

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

(Hint: Take the limit as $\alpha \rightarrow 0$ of the Fourier transform found.)

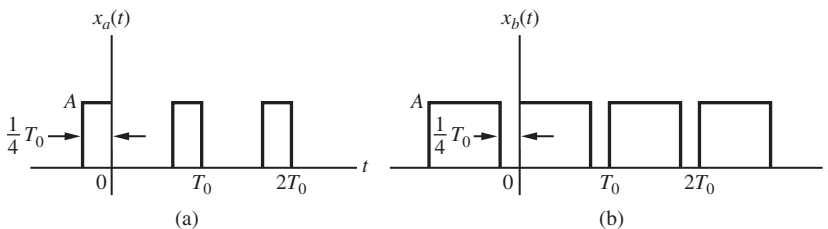


Figure 2.34

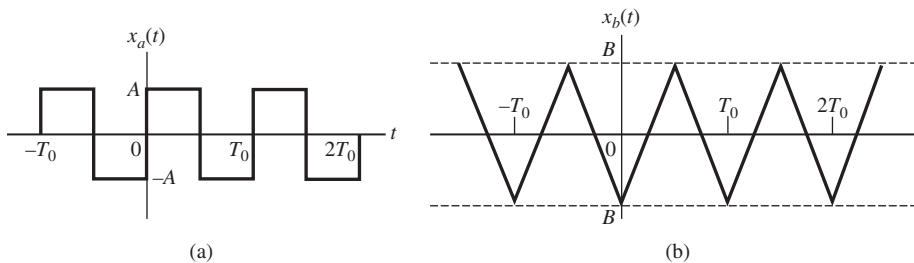


Figure 2.35

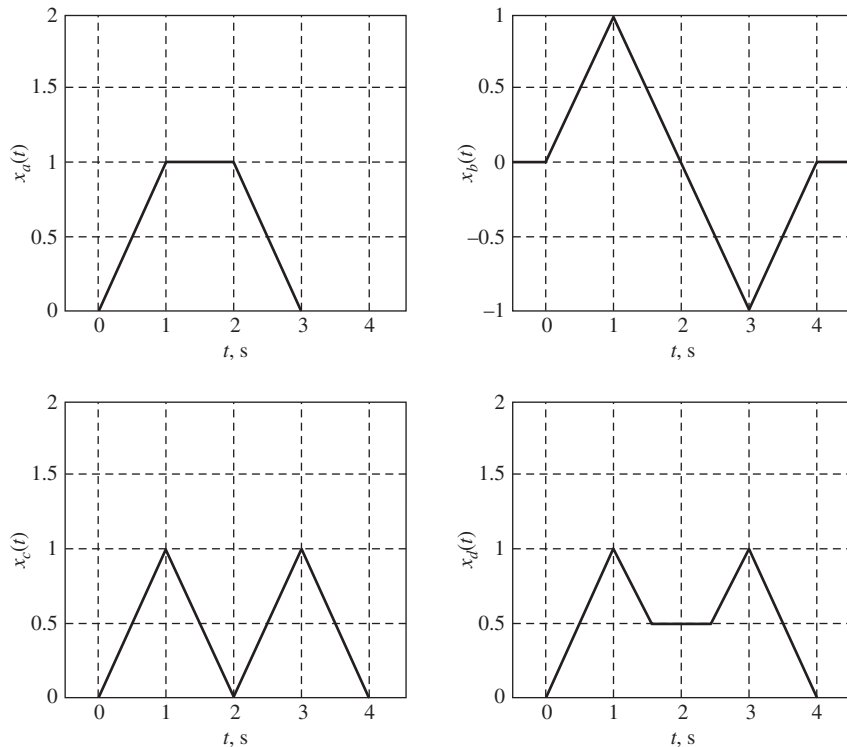


Figure 2.36

- (b) Use the result above and the relation $u(t) = \frac{1}{2} [\text{sgn}(t) + 1]$ to find the Fourier transform of the unit step.
- (c) Use the integration theorem and the Fourier transform of the unit impulse function to find the Fourier transform of the unit step. Compare the result with part (b).

2.26 Using only the Fourier transform of the unit impulse function and the differentiation theorem, find the Fourier transforms of the signals shown in Figure 2.36.

2.27

- (a) Write the signals of Figure 2.37 as the linear combination of two delayed triangular functions. That is, write $x_a(t) = a_1 \Lambda((t - t_1)/T_1) + a_2 \Lambda((t - t_2)/T_2)$ by finding appropriate values for $a_1, a_2, t_1, t_2, T_1,$ and T_2 . Do similar expressions for all four signals shown in Figure 2.36.
- (b) Given the Fourier-transform pair $\Lambda(t) \leftrightarrow \text{sinc}^2(f)$, find their Fourier transforms using the superposition, scale-change, and time-delay the-

orems. Compare your results with the answers obtained in Problem 2.26.

2.28

- (a) Given $\Pi(t) \leftrightarrow \text{sinc}(f)$, find the Fourier transforms of the following signals using the frequency-translation followed by the time-delay theorem.
- (i) $x_1(t) = \Pi(t - 1) \exp[j4\pi(t - 1)]$
- (ii) $x_2(t) = \Pi(t + 1) \exp[j4\pi(t + 1)]$
- (b) Repeat the above, but now applying the time-delay theorem followed by the frequency-translation theorem.

2.29 By applying appropriate theorems and using the signals defined in Problem 2.28, find Fourier transforms of the following signals:

- (a) $x_a(t) = \frac{1}{2}x_1(t) + \frac{1}{2}x_1(-t)$
- (b) $x_b(t) = \frac{1}{2}x_2(t) + \frac{1}{2}x_2(-t)$

2.30 Use the superposition, scale-change, and time-delay theorems along with the transform pairs $\Pi(t) \leftrightarrow \text{sinc}(f)$, $\text{sinc}(t) \leftrightarrow \Pi(f)$, $\Lambda(t) \leftrightarrow \text{sinc}^2(f)$, and $\text{sinc}^2(t) \leftrightarrow \Lambda(f)$ to find Fourier transforms of the following:

- (a) $x_1(t) = \Pi\left(\frac{t-1}{2}\right)$
- (b) $x_2(t) = 2 \text{sinc}[2(t-1)]$
- (c) $x_3(t) = \Lambda\left(\frac{t-2}{8}\right)$
- (d) $x_4(t) = \text{sinc}^2\left(\frac{t-3}{4}\right)$
- (e) $x_5(t) = 5 \text{sinc}[2(t-1)] + 5 \text{sinc}[2(t+1)]$
- (f) $x_6(t) = 2\Lambda\left(\frac{t-2}{8}\right) + 2\Lambda\left(\frac{t+2}{8}\right)$

2.31 Without actually computing them, but using appropriate sketches, tell if the Fourier transforms of the signals given below are real, imaginary, or neither; even, odd, or neither. Give your reasoning in each case.

- (a) $x_1(t) = \Pi(t+1/2) - \Pi(t-1/2)$
- (b) $x_2(t) = \Pi(t/2) + \Pi(t)$
- (c) $x_3(t) = \sin(2\pi t)\Pi(t)$
- (d) $x_4(t) = \sin(2\pi t + \pi/4)\Pi(t)$
- (e) $x_5(t) = \cos(2\pi t)\Pi(t)$
- (f) $x_6(t) = 1/[1+(t/5)^4]$

2.32 Use the Poisson sum formula to obtain the Fourier series of the signal

$$x(t) = \sum_{m=-\infty}^{\infty} \Pi\left(\frac{t-4m}{2}\right)$$

2.33 Find and plot the energy spectral densities of the following signals. Dimension your plots fully. Use appropriate Fourier-transforms pairs and theorems.

- (a) $x_1(t) = 10e^{-5t}u(t)$
- (b) $x_2(t) = 10 \text{sinc}(2t)$
- (c) $x_3(t) = 3\Pi(2t)$
- (d) $x_4(t) = 3\Pi(2t)\cos(10\pi t)$

2.34 Evaluate the following integrals using Rayleigh's energy theorem (Parseval's theorem for Fourier transforms).

(a) $I_1 = \int_{-\infty}^{\infty} \frac{df}{a^2 + (2\pi f)^2}$

[Hint: Consider the Fourier transform of $\exp(-at)u(t)$.]

- (b) $I_2 = \int_{-\infty}^{\infty} \text{sinc}^2(\tau f) df$
- (c) $I_3 = \int_{-\infty}^{\infty} \frac{df}{[a^2 + (2\pi f)^2]^2}$
- (d) $I_4 = \int_{-\infty}^{\infty} \text{sinc}^4(\tau f) df$

2.35 Obtain and sketch the convolutions of the following signals.

- (a) $y_1(t) = e^{-at}u(t) * \Pi(t-\tau)$, a and τ positive constants
- (b) $y_2(t) = [\Pi(t/2) + \Pi(t)] * \Pi(t)$
- (c) $y_3(t) = e^{-\alpha|t|} * \Pi(t)$, $\alpha > 0$
- (d) $y_4(t) = x(t) * u(t)$, where $x(t)$ is any energy signal [you will have to assume a particular form for $x(t)$ to sketch this one, but obtain the general result before doing so].

2.36 Find the signals corresponding to the following spectra. Make use of appropriate Fourier-transform theorems.

- (a) $X_1(f) = 2 \cos(2\pi f)\Pi(f) \exp(-j4\pi f)$
- (b) $X_2(f) = \Lambda(f/2) \exp(-j5\pi f)$
- (c) $X_3(f) = \left[\Pi\left(\frac{f+4}{2}\right) + \Pi\left(\frac{f-4}{2}\right)\right] \exp(-j8\pi f)$

2.37 Given the following signals, suppose that all energy spectral components outside the bandwidth $|f| \leq W$ are removed by an ideal filter, while all energy spectral components within this bandwidth are kept. Find the ratio of energy kept to total energy in each case. (α , β , and τ are positive constants.)

- (a) $x_1(t) = e^{-at}u(t)$
- (b) $x_2(t) = \Pi(t/\tau)$ (requires numerical integration)
- (c) $x_3(t) = e^{-at}u(t) - e^{-\beta t}u(t)$

2.38

- (a) Find the Fourier transform of the cosine pulse

$$x(t) = A\Pi\left(\frac{2t}{T_0}\right) \cos(\omega_0 t), \text{ where } \omega_0 = \frac{2\pi}{T_0}$$

Express your answer in terms of a sum of sinc functions. Provide MATLAB plots of $x(t)$ and $X(f)$ [note that $X(f)$ is real].

- (b) Obtain the Fourier transform of the raised cosine pulse

$$y(t) = \frac{1}{2}A\Pi\left(\frac{2t}{T_0}\right) [1 + \cos(2\omega_0 t)]$$

106 Chapter 2 • Signal and Linear System Analysis

Provide MATLAB plots of $y(t)$ and $Y(f)$ [note that $Y(f)$ is real]. Compare with part (a).

- (c) Use Equation (2.134) with the result of part (a) to find the Fourier transform of the half-rectified cosine wave.

2.39 Provide plots of the following functions of time and find their Fourier transforms. Tell which ones should be real and even functions of f and which ones should be imaginary and odd functions of f . Do your results bear this out?

- (a) $x_1(t) = \Lambda\left(\frac{t}{2}\right) + \Pi\left(\frac{t}{2}\right)$
 (b) $x_2(t) = \Pi\left(\frac{t}{2}\right) - \Lambda(t)$
 (c) $x_3(t) = \Pi\left(t + \frac{1}{2}\right) - \Pi\left(t - \frac{1}{2}\right)$
 (d) $x_4(t) = \Lambda(t - 1) - \Lambda(t + 1)$
 (e) $x_5(t) = \Lambda(t)\text{sgn}(t)$
 (f) $x_6(t) = \Lambda(t)\cos(2\pi t)$

Section 2.5

2.40

- (a) Obtain the time-average autocorrelation function of $x(t) = 3 + 6\cos(20\pi t) + 3\sin(20\pi t)$.

(Hint: Combine the cosine and sine terms into a single cosine with a phase angle.)

- (b) Obtain the power spectral density of the signal of part (a). What is its total average power?

2.41 Find the power spectral densities and average powers of the following signals.

- (a) $x_1(t) = 2\cos(20\pi t + \pi/3)$
 (b) $x_2(t) = 3\sin(30\pi t)$
 (c) $x_3(t) = 5\sin(10\pi t - \pi/6)$
 (d) $x_4(t) = 3\sin(30\pi t) + 5\sin(10\pi t - \pi/6)$

2.42 Find the autocorrelation functions of the signals having the following power spectral densities. Also give their average powers.

- (a) $S_1(f) = 4\delta(f - 15) + 4\delta(f + 15)$
 (b) $S_2(f) = 9\delta(f - 20) + 9\delta(f + 20)$
 (c) $S_3(f) = 16\delta(f - 5) + 16\delta(f + 5)$
 (d) $S_4(f) = 9\delta(f - 20) + 9\delta(f + 20) + 16\delta(f - 5) + 16\delta(f + 5)$

2.43 By applying the properties of the autocorrelation function, determine whether the following are acceptable

for autocorrelation functions. In each case, tell why or why not.

- (a) $R_1(\tau) = 2\cos(10\pi\tau) + \cos(30\pi\tau)$
 (b) $R_2(\tau) = 1 + 3\cos(30\pi\tau)$
 (c) $R_3(\tau) = 3\cos(20\pi\tau + \pi/3)$
 (d) $R_4(\tau) = 4\Lambda(\tau/2)$
 (e) $R_5(\tau) = 3\Pi(\tau/6)$
 (f) $R_6(\tau) = 2\sin(10\pi\tau)$

2.44 Find the autocorrelation functions corresponding to the following signals.

- (a) $x_1(t) = 2\cos(10\pi t + \pi/3)$
 (b) $x_2(t) = 2\sin(10\pi t + \pi/3)$
 (c) $x_3(t) = \text{Re} [3\exp(j10\pi t) + 4j\exp(j10\pi t)]$
 (d) $x_4(t) = x_1(t) + x_2(t)$

2.45 Show that the $R(\tau)$ of Example 2.20 has the Fourier transform $S(f)$ given there. Plot the power spectral density.

Section 2.6

2.46 A system is governed by the differential equation (a , b , and c are nonnegative constants)

$$\frac{dy(t)}{dt} + ay(t) = b\frac{dx(t)}{dt} + cx(t)$$

- (a) Find $H(f)$.
 (b) Find and plot $|H(f)|$ and $\angle H(f)$ for $c = 0$.
 (c) Find and plot $|H(f)|$ and $\angle H(f)$ for $b = 0$.

2.47 For each of the following transfer functions, determine the unit impulse response of the system.

- (a) $H_1(f) = \frac{1}{7 + j2\pi f}$
 (b) $H_2(f) = \frac{j2\pi f}{7 + j2\pi f}$

(Hint: Use long division first.)

- (c) $H_3(f) = \frac{e^{-j6\pi f}}{7 + j2\pi f}$
 (d) $H_4(f) = \frac{1 - e^{-j6\pi f}}{7 + j2\pi f}$

2.48 A filter has frequency response function $H(f) = \Pi(f/2B)$ and input $x(t) = 2W\text{sinc}(2Wt)$.

- (a) Find the output $y(t)$ for $W < B$.
 (b) Find the output $y(t)$ for $W > B$.
 (c) In which case does the output suffer distortion? What influenced your answer?

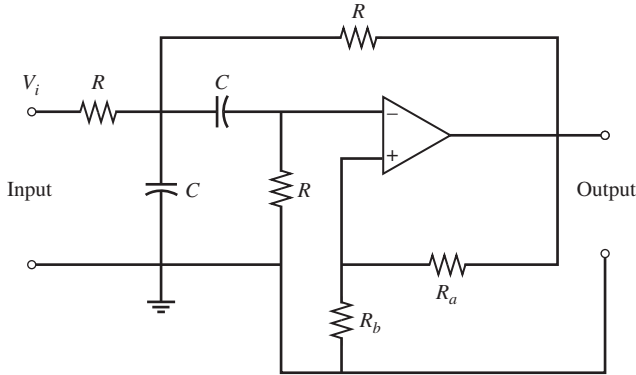


Figure 2.37

2.49 A second-order active bandpass filter (BPF), known as a bandpass Sallen–Key circuit, is shown in Figure 2.37.

- (a) Show that the frequency response function of this filter is given by

$$H(j\omega) = \frac{(K\omega_0/\sqrt{2})(j\omega)}{-\omega^2 + (\omega_0/Q)(j\omega) + \omega_0^2}, \quad \omega = 2\pi f$$

where

$$\omega_0 = \sqrt{2}(RC)^{-1}$$

$$Q = \frac{\sqrt{2}}{4 - K}$$

$$K = 1 + \frac{R_a}{R_b}$$

- (b) Plot $|H(f)|$.
 (c) Show that the 3-dB bandwidth of the filter can be expressed as $B = f_0/Q$, where $f_0 = \omega_0/2\pi$.

- (d) Design a BPF using this circuit with center frequency $f_0 = 1000$ Hz and 3-dB bandwidth of 300 Hz. Find values of R_a , R_b , R , and C that will give these desired specifications.

2.50 For the two circuits shown in Figure 2.38, determine $H(f)$ and $h(t)$. Sketch accurately the amplitude and phase responses. Plot the amplitude response in decibels. Use a logarithmic frequency axis.

2.51 Using the Paley-Wiener criterion, show that

$$|H(f)| = \exp(-\beta f^2)$$

is not a suitable amplitude response for a causal, linear time-invariant filter.

2.52 Determine whether or not the filters with impulse responses given below are BIBO stable. α and f_0 are positive constants.

(a) $h_1(t) = \exp(-\alpha|t|) \cos(2\pi f_0 t)$

(b) $h_2(t) = \cos(2\pi f_0 t) u(t)$

(c) $h_3(t) = t^{-1} u(t - 1)$

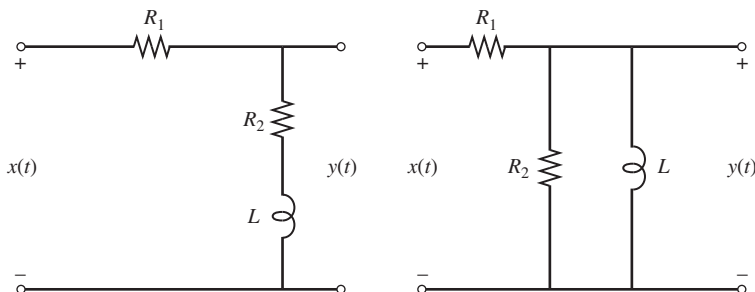


Figure 2.38

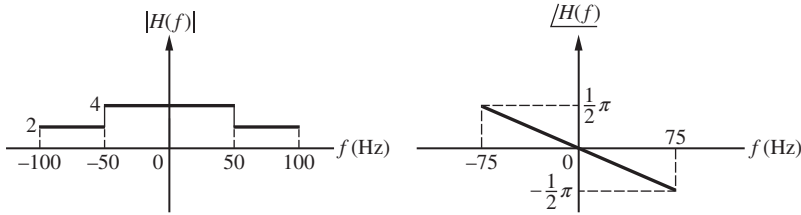


Figure 2.39

- (d) $h_4(t) = e^{-t}u(t) - e^{-(t-1)}u(t-1)$
- (e) $h_5(t) = t^{-2}u(t-1)$
- (f) $h_6(t) = \text{sinc}(2t)$

2.53 Given a filter with frequency response function

$$H(f) = \frac{5}{4 + j(2\pi f)}$$

and input $x(t) = e^{-3t}u(t)$, obtain and plot accurately the energy spectral densities of the input and output.

2.54 A filter with frequency response function

$$H(f) = 3\Pi\left(\frac{f}{62}\right)$$

has as an input a half-rectified cosine waveform of fundamental frequency 10 Hz. Determine an analytical expression for the output of the filter. Plot the output using MATLAB.

2.55 Another definition of bandwidth for a signal is the 90% energy containment bandwidth. For a signal with energy spectral density $G(f) = |X(f)|^2$, it is given by B_{90} in the relation

$$0.9E_{\text{Total}} = \int_{-B_{90}}^{B_{90}} G(f) df = 2 \int_0^{B_{90}} G(f) df;$$

$$E_{\text{Total}} = \int_{-\infty}^{\infty} G(f) df = 2 \int_0^{\infty} G(f) df$$

Obtain B_{90} for the following signals if it is defined. If it is not defined for a particular signal, state why it is not.

- (a) $x_1(t) = e^{-\alpha t}u(t)$, where α is a positive constant
- (b) $x_2(t) = 2W \text{sinc}(2Wt)$ where W is a positive constant
- (c) $x_3(t) = \Pi(t/\tau)$ (requires numerical integration)
- (d) $x_4(t) = \Lambda(t/\tau)$ (requires numerical integration)
- (e) $x_5(t) = e^{-\alpha|t|}$

2.56 An ideal quadrature phase shifter has frequency response function

$$H(f) = \begin{cases} e^{-j\pi/2}, & f > 0 \\ e^{+j\pi/2}, & f < 0 \end{cases}$$

Find the outputs for the following inputs:

- (a) $x_1(t) = \exp(j100\pi t)$
- (b) $x_2(t) = \cos(100\pi t)$
- (c) $x_3(t) = \sin(100\pi t)$
- (d) $x_4(t) = \Pi(t/2)$

2.57 A filter has amplitude response and phase shift shown in Figure 2.39. Find the output for each of the inputs given below. For which cases is the transmission distortionless? Tell what type of distortion is imposed for the others.

- (a) $\cos(48\pi t) + 5 \cos(126\pi t)$
- (b) $\cos(126\pi t) + 0.5 \cos(170\pi t)$
- (c) $\cos(126\pi t) + 3 \cos(144\pi t)$
- (d) $\cos(10\pi t) + 4 \cos(50\pi t)$

2.58 Determine and accurately plot, on the same set of axes, the group delay and the phase delay for the systems with unit impulse responses:

- (a) $h_1(t) = 3e^{-5t}u(t)$
- (b) $h_2(t) = 5e^{-3t}u(t) - 2e^{-5t}u(t)$
- (c) $h_3(t) = \text{sinc}[2B(t-t_0)]$ where B and t_0 are positive constants
- (d) $h_4(t) = 5e^{-3t}u(t) - 2e^{-3(t-t_0)}u(t-t_0)$ where t_0 is a positive constant

2.59 A system has the frequency response function

$$H(f) = \frac{j2\pi f}{(8 + j2\pi f)(3 + j2\pi f)}$$

Determine and accurately plot the following: (a) The amplitude response; (b) The phase response; (c) The phase delay; (d) The group delay.

2.60 The nonlinear system defined by

$$y(t) = x(t) + 0.1x^2(t)$$

has an input signal with the bandpass spectrum

$$X(f) = 2\Pi\left(\frac{f-10}{4}\right) + 2\Pi\left(\frac{f+10}{4}\right)$$

Sketch the spectrum of the output, labeling all important frequencies and amplitudes.

2.61 Given a filter with frequency response function

$$H(f) = \frac{j2\pi f}{(9 - 4\pi^2 f^2) + j0.3\pi f}$$

Determine and accurately plot the following: (a) The amplitude response; (b) The phase response; (c) The phase delay; (d) The group delay.

2.62 Given a nonlinear, zero-memory device with transfer characteristic

$$y(t) = x^3(t),$$

find its output due to the input

$$x(t) = \cos(2\pi t) + \cos(6\pi t)$$

List all frequency components and tell whether they are due to harmonic generation or intermodulation terms.

2.63 Find the impulse response of an ideal highpass filter with the frequency response function

$$H_{HP}(f) = H_0 \left[1 - \Pi\left(\frac{f}{2W}\right) \right] e^{-j2\pi f t_0}$$

2.64 Verify the pulsewidth-bandwidth relationship of Equation (2.234) for the following signals. Sketch each signal and its spectrum.

- (a) $x(t) = A \exp(-t^2/2\tau^2)$ (Gaussian pulse)
- (b) $x(t) = A \exp(-\alpha |t|)$, $\alpha > 0$ (double-sided exponential)

2.65

- (a) Show that the frequency response function of a second-order Butterworth filter is

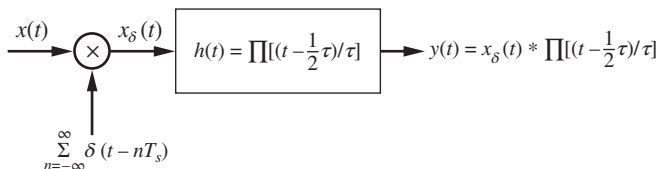


Figure 2.40

$$H(f) = \frac{f_3^2}{f_3^2 + j\sqrt{2}f_3f - f^2}$$

where f_3 is the 3-dB frequency in hertz.

- (b) Find an expression for the group delay of this filter. Plot the group delay as a function of f/f_3 .
- (c) Given that the step response for a second-order Butterworth filter is

$$y_s(t) = \left[1 - \exp\left(-\frac{2\pi f_3 t}{\sqrt{2}}\right) \times \left(\cos \frac{2\pi f_3 t}{\sqrt{2}} + \sin \frac{2\pi f_3 t}{\sqrt{2}} \right) \right] u(t)$$

where $u(t)$ is the unit step function, find the 10% to 90% risetime in terms of f_3 .

Section 2.7

2.66 A sinusoidal signal of frequency 1 Hz is to be sampled periodically.

- (a) Find the maximum allowable time interval between samples.
- (b) Samples are taken at $\frac{1}{3}$ -s intervals (i.e., at a rate of $f_s = 3$ sps). Construct a plot of the sampled signal spectrum that illustrates that this is an acceptable sampling rate to allow recovery of the original sinusoid.
- (c) The samples are spaced $\frac{2}{3}$ s apart. Construct a plot of the sampled signal spectrum that shows what the recovered signal will be if the samples are passed through a lowpass filter such that only the lowest frequency spectral lines are passed.

2.67 A flat-top sampler can be represented as the block diagram of Figure 2.40.

- (a) Assuming $\tau \ll T_s$, sketch the output for a typical $x(t)$.
- (b) Find the spectrum of the output, $Y(f)$, in terms of the spectrum of the input, $X(f)$. Determine the relationship between τ and T_s required to minimize distortion in the recovered waveform?

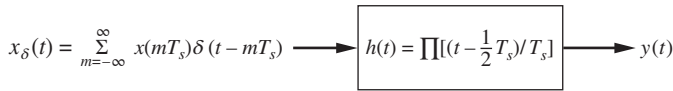


Figure 2.41

2.68 Figure 2.41 illustrates so-called *zero-order-hold reconstruction*.

- (a) Sketch $y(t)$ for a typical $x(t)$. Under what conditions is $y(t)$ a good approximation to $x(t)$?
- (b) Find the spectrum of $y(t)$ in terms of the spectrum of $x(t)$. Discuss the approximation of $y(t)$ to $x(t)$ in terms of frequency-domain arguments.

2.69 Determine the range of permissible cutoff frequencies for the ideal lowpass filter used to reconstruct the signal

$$x(t) = 10 \cos^2(600\pi t) \cos(2400\pi t)$$

which is sampled at 4500 samples per second. Sketch $X(f)$ and $X_\delta(f)$. Find the minimum allowable sampling frequency.

2.70 Given the bandpass signal spectrum shown in Figure 2.42, sketch spectra for the following sampling rates f_s and indicate which ones are suitable.

- (a) $2B$ (b) $2.5B$ (c) $3B$ (d) $4B$ (e) $5B$ (f) $6B$

Section 2.8

2.71 Using appropriate Fourier-transform theorems and pairs, express the spectrum $Y(f)$ of

$$y(t) = x(t) \cos(\omega_0 t) + \hat{x}(t) \sin(\omega_0 t)$$

in terms of the spectrum $X(f)$ of $x(t)$, where $X(f)$ is lowpass with bandwidth

$$B < f_0 = \frac{\omega_0}{2\pi}$$

Sketch $Y(f)$ for a typical $X(f)$.

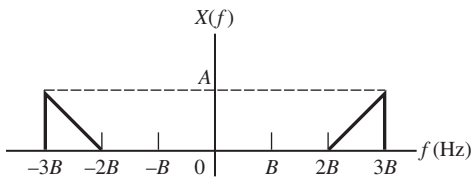


Figure 2.42

2.72 Show that $x(t)$ and $\hat{x}(t)$ are orthogonal for the following signals:

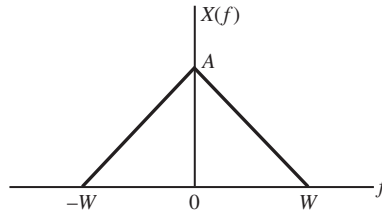


Figure 2.43

- (a) $x_1(t) = \sin(\omega_0 t)$
- (b) $x_2(t) = 2 \cos(\omega_0 t) + \sin(\omega_0 t) \cos(2\omega_0 t)$
- (c) $x_3(t) = A \exp(j\omega_0 t)$

2.73 Assume that the Fourier transform of $x(t)$ is real and has the shape shown in Figure 2.43. Determine and plot the spectrum of each of the following signals:

- (a) $x_1(t) = \frac{2}{3}x(t) + \frac{1}{3}j\hat{x}(t)$
- (b) $x_2(t) = \left[\frac{3}{4}x(t) + \frac{3}{4}j\hat{x}(t)\right] e^{j2\pi f_0 t}$, $f_0 \gg W$
- (c) $x_3(t) = \left[\frac{2}{3}x(t) + \frac{1}{3}j\hat{x}(t)\right] e^{j2\pi W t}$
- (d) $x_4(t) = \left[\frac{2}{3}x(t) - \frac{1}{3}j\hat{x}(t)\right] e^{j\pi W t}$

2.74 Following Example 2.30, consider

$$x(t) = 2 \cos(52\pi t)$$

Find $\hat{x}(t)$, $x_p(t)$, $\tilde{x}(t)$, $x_R(t)$, and $x_I(t)$ for the following cases: (a) $f_0 = 25$ Hz; (b) $f_0 = 27$ Hz; (c) $f_0 = 10$ Hz; (d) $f_0 = 15$ Hz; (e) $f_0 = 30$ Hz; (f) $f_0 = 20$ Hz.

2.75 Consider the input

$$x(t) = \Pi(t/\tau) \cos[2\pi(f_0 + \Delta f)t], \quad \Delta f \ll f_0$$

to a filter with impulse response

$$h(t) = \alpha e^{-at} \cos(2\pi f_0 t) u(t)$$

Find the output using complex envelope techniques.

Computer Exercises

2.1 Write²⁰ a computer program to sum the Fourier series for the signals given in Table 2.1. The number of terms in the Fourier sum should be adjustable so that one may study the convergence of each Fourier series.

2.2 Generalize the computer program of Computer Example 2.1 to evaluate the coefficients of the complex exponential Fourier series of several signals. Include a plot of the amplitude and phase spectrum of the signal for which the Fourier series coefficients are evaluated. Check by evaluating the Fourier series coefficients of a squarewave.

2.3 Write a computer program to evaluate the coefficients of the complex exponential Fourier series of a signal by using the fast Fourier transform (FFT). Check it by evaluating the Fourier series coefficients of a squarewave and comparing your results with Computer Exercise 2.2.

2.4 How would you use the same approach as in Computer Exercise 2.3 to evaluate the Fourier transform of a pulse-type signal. How do the two outputs differ? Compute an approximation to the Fourier transform of a square pulse signal 1 unit wide and compare with the theoretical result.

2.5 Write a computer program to find the bandwidth of a lowpass energy signal that contains a certain specified percentage of its total energy, for example, 95%. In other words, write a program to find W in the equation

$$E_W = \frac{\int_0^W G_x(f) df}{\int_0^\infty G_x(f) df} \times 100\%$$

with E_W set equal to a specified value, where $G_x(f)$ is the energy spectral density of the signal.

2.6 Write a computer program to find the time duration of a lowpass energy signal that contains a certain specified percentage of its total energy, for example, 95%. In other words, write a program to find T in the equation

$$E_T = \frac{\int_0^T |x(t)|^2 dt}{\int_0^\infty |x(t)|^2 dt} \times 100\%$$

with E_T set equal to a specified value, where it is assumed that the signal is zero for $t < 0$.

2.7 Use a MATLAB program like Computer Example 2.2 to investigate the frequency response of the Sallen-Key circuit for various Q -values.

²⁰When doing these computer exercises, we suggest that the student make use of a mathematics package such as MATLAB™. Considerable time will be saved in being able to use the plotting capability of MATLAB™. You should strive to use the vector capability of MATLAB™ as well.

CHAPTER 3

LINEAR MODULATION TECHNIQUES

Before an information-bearing signal is transmitted through a communication channel, some type of modulation process is typically utilized to produce a signal that can easily be accommodated by the channel. In this chapter we will discuss various types of *linear* modulation techniques. The modulation process commonly translates an information-bearing signal, usually referred to as the message signal, to a new spectral location depending upon the intended frequency for transmission. For example, if the signal is to be transmitted through the atmosphere or free space, frequency translation is necessary to raise the signal spectrum to a frequency that can be radiated efficiently with antennas of reasonable size. If more than one signal utilizes a channel, modulation allows translation of different signals to different spectral locations, thus allowing the receiver to select the desired signal. Multiplexing allows two or more message signals to be transmitted by a single transmitter and received by a single receiver simultaneously. The logical choice of a modulation technique for a specific application is influenced by the characteristics of the message signal, the characteristics of the channel, the performance desired from the overall communication system, the use to be made of the transmitted data, and the economic factors that are always important in practical applications.

The two basic types of analog modulation are continuous-wave modulation and pulse modulation. In continuous-wave modulation, which is the main focus of this chapter, a parameter of a high-frequency carrier is varied proportionally to the message signal such that a one-to-one correspondence exists between the parameter and the message signal. The carrier is usually assumed to be sinusoidal, but as will be illustrated, this is not a necessary restriction. However, for a sinusoidal carrier, a general modulated carrier can be represented mathematically as

$$x_c(t) = A(t) \cos[2\pi f_c t + \phi(t)] \quad (3.1)$$

where f_c is the carrier frequency. Since a sinusoid is completely specified by its amplitude, $A(t)$, and instantaneous phase, $2\pi f_c t + \phi(t)$, it follows that once the carrier frequency, f_c , is specified, only two parameters are candidates to be varied in the modulation process: the instantaneous amplitude $A(t)$ and the phase deviation $\phi(t)$. When the amplitude $A(t)$ is linearly related to the modulating signal, the result is linear modulation. Letting $\phi(t)$ or the time derivative of $\phi(t)$ be linearly related to the modulating signal yields phase or frequency modulation, respectively. Collectively, phase and frequency modulation are referred to as angle modulation, since the instantaneous phase angle of the modulated carrier conveys the information.

In this chapter we focus on continuous-wave linear modulation. However, at the end of this chapter, we briefly consider pulse amplitude modulation, which is a linear process and a simple application of the sampling theorem studied in the preceding chapter. In the following chapter we consider angle modulation, both continuous wave and pulse.

3.1 DOUBLE-SIDEBAND MODULATION

A general linearly modulated carrier is represented by setting the instantaneous phase deviation $\phi(t)$ in (3.1) equal to zero. Thus, a linearly modulated carrier is represented by

$$x_c(t) = A(t) \cos(2\pi f_c t) \quad (3.2)$$

in which the carrier amplitude $A(t)$ varies in one-to-one correspondence with the message signal. We next discuss several different types of linear modulation as well as techniques that can be used for demodulation.

Double-sideband (DSB) modulation results when $A(t)$ is proportional to the message signal $m(t)$. Thus, the output of a DSB modulator can be represented as

$$x_c(t) = A_c m(t) \cos(2\pi f_c t) \quad (3.3)$$

which illustrates that DSB modulation is simply the multiplication of a carrier, $A_c \cos(2\pi f_c t)$, by the message signal. It follows from the modulation theorem for Fourier transforms that the spectrum of a DSB signal is given by

$$X_c(f) = \frac{1}{2} A_c M(f + f_c) + \frac{1}{2} A_c M(f - f_c) \quad (3.4)$$

The process of DSB modulation is illustrated in Figure 3.1. Figure 3.1(a) illustrates a DSB system and shows that a DSB signal is demodulated by multiplying the received signal, denoted by $x_r(t)$, by the demodulation carrier $2 \cos(2\pi f_c t)$ and lowpass filtering. For the idealized system that we are considering here, the received signal $x_r(t)$ is identical to the transmitted signal $x_c(t)$. The output of the multiplier is

$$d(t) = 2A_c [m(t) \cos(2\pi f_c t)] \cos(2\pi f_c t) \quad (3.5)$$

or

$$d(t) = A_c m(t) + A_c m(t) \cos(4\pi f_c t) \quad (3.6)$$

where we have used the trigonometric identity $2 \cos^2 x = 1 + \cos 2x$.

The time-domain signals are shown in Figure 3.1(b) for an assumed $m(t)$. The message signal $m(t)$ forms the envelope, or instantaneous magnitude, of $x_c(t)$. The waveform for $d(t)$ can be best understood by realizing that since $\cos^2(2\pi f_c t)$ is nonnegative for all t , then $d(t)$ is positive if $m(t)$ is positive and $d(t)$ is negative if $m(t)$ is negative. Also note that $m(t)$ (appropriately scaled) forms the envelope of $d(t)$ and that the frequency of the sinusoid under the envelope is $2f_c$ rather than f_c .

The spectra of the signals $m(t)$, $x_c(t)$ and $d(t)$ are shown in Figure 3.1(c) for an assumed $M(f)$ having a bandwidth W . The spectra $M(f + f_c)$ and $M(f - f_c)$ are simply the message spectrum translated to $f = \pm f_c$. The portion of $M(f - f_c)$ above the carrier frequency is called the *upper sideband* (USB), and the portion below the carrier frequency is called the *lower sideband* (LSB). Since the carrier frequency f_c is typically much greater than the bandwidth

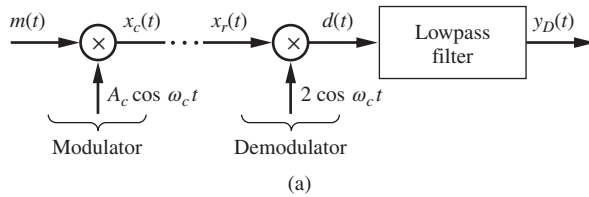
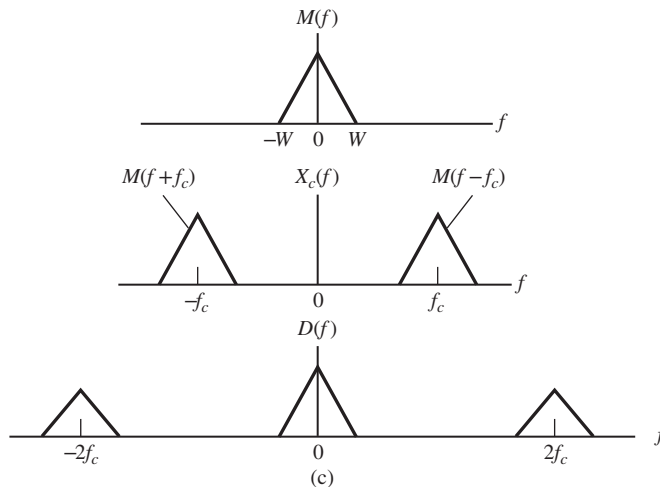
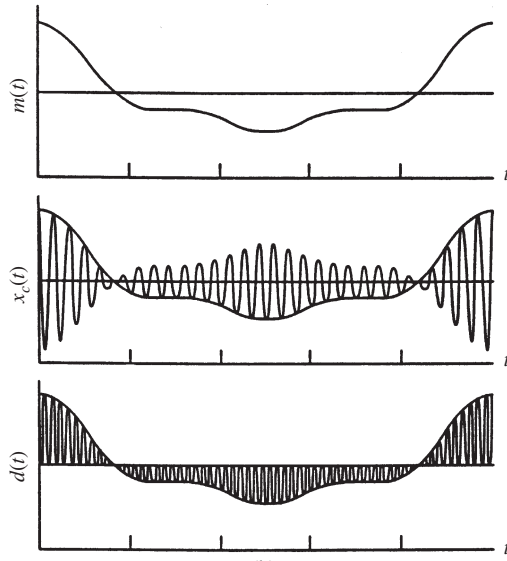


Figure 3.1
Double-sideband modulation. (a) System.
(b) Example waveforms.
(c) Spectra.



of the message signal W , the spectra of the two terms in $d(t)$ do not overlap. Thus, $d(t)$ can be lowpass filtered and amplitude scaled by A_c to yield the demodulated output $y_D(t)$. In practice, any amplitude scaling factor can be used since, as we saw in Chapter 2, multiplication by a constant does not induce amplitude distortion and the amplitude can be adjusted as desired. A volume control is an example. Thus, for convenience, A_c is usually set equal to unity at

the demodulator output. For this case, the demodulated output $y_D(t)$ will equal the message signal $m(t)$. The lowpass filter that removes the term at $2f_c$ must have a bandwidth greater than or equal to the bandwidth of the message signal W . We will see in Chapter 8 that when noise is present, this lowpass filter, known as the *postdetection* filter, should have the smallest possible bandwidth since minimizing the bandwidth of the postdetection filter is important for removing out-of-band noise or interference.

We will see later that DSB is 100% power efficient because all of the transmitted power lies in the sidebands and it is the sidebands that carry the message signal $m(t)$. This makes DSB modulation power efficient and therefore attractive, especially in power-limited applications. Demodulation of DSB is difficult, however, because the presence of a demodulation carrier, phase coherent with the carrier used for modulation at the transmitter, is required at the receiver. Demodulation utilizing a coherent reference is known as *synchronous* or *coherent demodulation*. The generation of a phase coherent demodulation carrier can be accomplished using a variety of techniques, including the use of a Costas phase-locked loop to be considered in the following chapter. The use of these techniques complicate receiver design. In addition, careful attention is required to ensure that phase errors in the demodulation carrier are minimized since even small phase errors can result in serious distortion of the demodulated message waveform. This effect will be thoroughly analyzed in Chapter 8, but a simplified analysis can be carried out by assuming a demodulation carrier in Figure 3.1(a) of the form $2 \cos[2\pi f_c t + \theta(t)]$, where $\theta(t)$ is a time-varying phase error. Applying the trigonometric identity

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

yields

$$d(t) = A_c m(t) \cos \theta(t) + A_c m(t) \cos[4\pi f_c t + \theta(t)] \quad (3.7)$$

which, after lowpass filtering and amplitude scaling to remove the carrier amplitude, becomes

$$y_D(t) = m(t) \cos \theta(t) \quad (3.8)$$

assuming, once again, that the spectra of the two terms of $d(t)$ do not overlap. If the phase error $\theta(t)$ is a constant, the effect of the phase error is an attenuation of the demodulated message signal. This does not represent distortion, since the effect of the phase error can be removed by amplitude scaling unless $\theta(t)$ is $\pi/2$. However, if $\theta(t)$ is time varying in an unknown and unpredictable manner, the effect of the phase error can be serious distortion of the demodulated output.

A simple technique for generating a phase coherent demodulation carrier is to square the received DSB signal, which yields

$$\begin{aligned} x_r^2(t) &= A_c^2 m^2(t) \cos^2(2\pi f_c t) \\ &= \frac{1}{2} A_c^2 m^2(t) + \frac{1}{2} A_c^2 m^2(t) \cos(4\pi f_c t) \end{aligned} \quad (3.9)$$

If $m(t)$ is a power signal, $m^2(t)$ has a nonzero DC value. Thus, by the modulation theorem, $x_r^2(t)$ has a discrete frequency component at $2f_c$, which can be extracted from the spectrum of $x_r^2(t)$ using a narrowband bandpass filter. The frequency of this component can be divided by 2 to yield the desired demodulation carrier. Later we will discuss a convenient technique for implementing the required frequency divider.

The analysis of DSB illustrates that the spectrum of a DSB signal does not contain a discrete spectral component at the carrier frequency unless $m(t)$ has a DC component. For this reason, DSB systems with no carrier frequency component present are often referred to as *suppressed carrier systems*. However, if a carrier component is transmitted along with the DSB signal, demodulation can be simplified. The received carrier component can be extracted using a narrowband bandpass filter and can be used as the demodulation carrier. If the carrier amplitude is sufficiently large, the need for generating a demodulation carrier can be completely avoided. This naturally leads to the subject of amplitude modulation.

■ 3.2 AMPLITUDE MODULATION (AM)

Amplitude modulation results when a carrier component is added to a DSB signal. Adding a carrier component, $A_c \cos(2\pi f_c t)$ to the DSB signal given by (3.3) and scaling the message signal gives

$$x_c(t) = A_c[1 + am_n(t)] \cos(2\pi f_c t) \quad (3.10)$$

in which a is the *modulation index*,¹ which typically takes on values in the range $0 < a \leq 1$, and $m_n(t)$ is a scaled version of the message signal $m(t)$. The scaling is applied to ensure that $m_n(t) \geq -1$ for all t . Mathematically

$$m_n(t) = \frac{m(t)}{|\min[m(t)]|} \quad (3.11)$$

We note that for $a \leq 1$, the condition $m_n(t) \geq -1$ for all t ensures that the *envelope* of the AM signal defined by $[1 + am_n(t)]$ is nonnegative for all t . We will understand the importance of this condition when we study *envelope detection* in the following section. The time-domain representation of AM is illustrated in Figure 3.2(a) and (b), and the block diagram of the modulator for producing AM is shown in Figure 3.2(c).

An AM signal can be demodulated using the same coherent demodulation technique that was used for DSB. However, the use of coherent demodulation negates the advantage of AM. The advantage of AM over DSB is that a very simple technique, known as envelope detection or envelope demodulation, can be used. An envelope demodulator is implemented as shown in Figure 3.3(a). It can be seen from Figure 3.2(b) that, as the carrier frequency is increased, the envelope, defined as $A_c[1 + am_n(t)]$, becomes well defined and easier to observe. More importantly, it also follows from observation of Figure 3.3(b) that, if the envelope of the AM signal $A_c[1 + am_n(t)]$ goes negative, distortion will result in the demodulated signal assuming that envelope demodulation is used. The normalized message signal is defined so that this distortion is prevented. Thus, for $a = 1$, the minimum value of $1 + am_n(t)$ is zero. In order to ensure that the envelope is nonnegative for all t we require that $1 + m_n(t) \geq 0$ or, equivalently, $m_n(t) \geq -1$ for all t . The normalized message signal $m_n(t)$ is therefore found by dividing $m(t)$ by a positive constant so that the condition $m_n(t) \geq -1$ is satisfied. This normalizing constant is $|\min m(t)|$. In many cases of practical interest, such as speech or music signals, the maximum and minimum values of the message signal

¹The parameter a as used here is sometimes called the *negative modulation factor*. Also, the quantity $a \times 100\%$ is often referred to as the *percent modulation*.

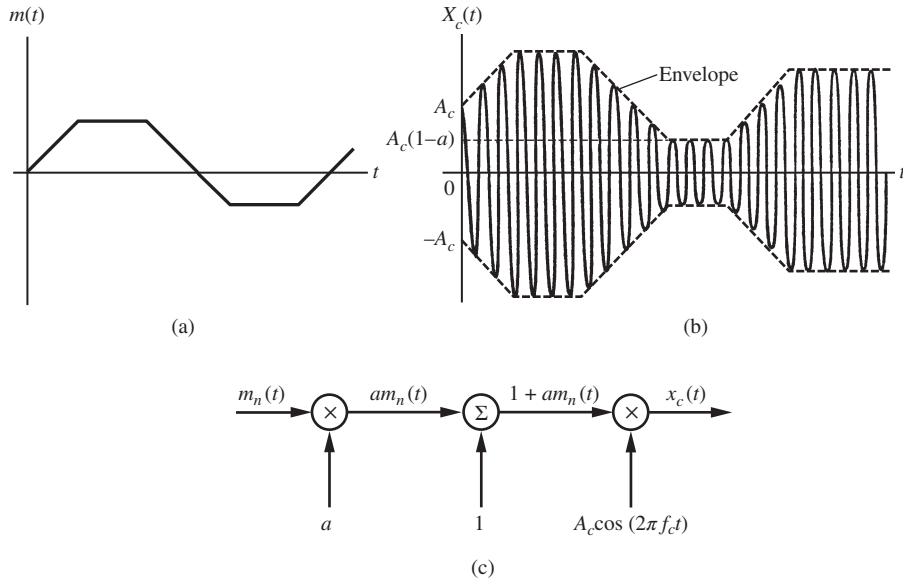


Figure 3.2 Amplitude modulation. (a) Message signal. (b) Modulator output for $a < 1$. (c) Modulator.

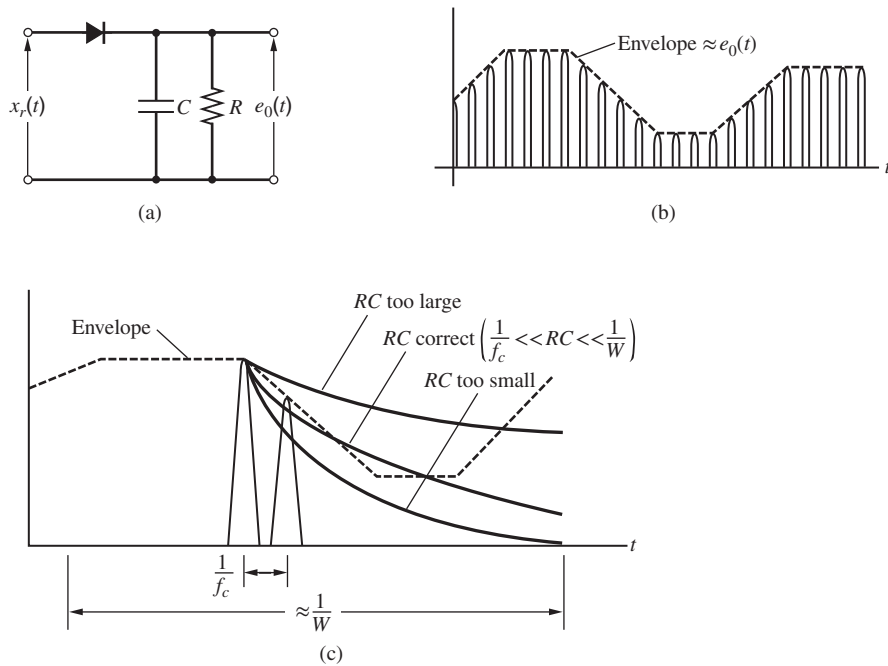


Figure 3.3 Envelope detection. (a) Circuit. (b) Waveform. (c) Effect of RC time constant.

are equal. We will see why this is true when we study probability and random signals in Chapters 6 and 7.

3.2.1 Envelope Detection

In order for the envelope detection process to operate properly, the RC time constant of the detector, shown in Figure 3.3(a), must be chosen carefully. The appropriate value for the time constant is related to the carrier frequency and to the bandwidth of $m(t)$. In practice, satisfactory operation requires a carrier frequency of at least 10 times the bandwidth of $m(t)$, which is designated W . Also, the cutoff frequency of the RC circuit must lie between f_c and W and must be well separated from both. This is illustrated in Figure 3.3(c).

All information in the modulator output is contained in the sidebands. Thus, the carrier component of (3.10), $A_c \cos \omega_c t$, is wasted power as far as information transfer is concerned. This fact can be of considerable importance in an environment where power is limited and can completely preclude the use of AM as a modulation technique in power-limited applications.

From (3.10) we see that the total power contained in the AM modulator output is

$$\langle x_c^2(t) \rangle = \langle A_c^2 [1 + am_n(t)]^2 \cos^2(2\pi f_c t) \rangle \quad (3.12)$$

where $\langle \cdot \rangle$ denotes the time average value. If $m_n(t)$ is *slowly* varying with respect to the carrier

$$\begin{aligned} \langle x_c^2 \rangle &= \left\langle A_c^2 [1 + am_n(t)]^2 \left[\frac{1}{2} + \frac{1}{2} \cos(4\pi f_c t) \right] \right\rangle \\ &= \left\langle \frac{1}{2} A_c^2 [1 + 2am_n(t) + a^2 m_n^2(t)] \right\rangle \end{aligned} \quad (3.13)$$

Assuming $m_n(t)$ to have zero average value and taking the time average term-by-term gives

$$\langle x_c^2(t) \rangle = \frac{1}{2} A_c^2 + \frac{1}{2} A_c^2 a^2 \langle m_n^2(t) \rangle \quad (3.14)$$

The first term in the preceding expression represents the carrier power, and the second term represents the sideband (information) power. The efficiency of the modulation process is defined as the ratio of the power in the information-bearing signal (the sideband power) to the total power in the transmitted signal. This is

$$E_{ff} = \frac{a^2 \langle m_n^2(t) \rangle}{1 + a^2 \langle m_n^2(t) \rangle} \quad (3.15)$$

The efficiency is typically multiplied by 100 so that efficiency can be expressed as a percent.

If the message signal has symmetrical maximum and minimum values, such that $|\min m(t)|$ and $|\max m(t)|$ are equal, then $\langle m_n^2(t) \rangle \leq 1$. It follows that for $a \leq 1$, the maximum efficiency is 50% and is achieved for square-wave-type message signals. If $m(t)$ is a sine wave, $\langle m_n^2(t) \rangle = \frac{1}{2}$ and the efficiency is 33.3% for $a = 1$. Note that if we allow the modulation index to exceed 1, efficiency can exceed 50% and that $E_{ff} \rightarrow 100\%$ as $a \rightarrow \infty$. Values of a greater than 1, as we have seen, preclude the use of envelope detection. Efficiency obviously

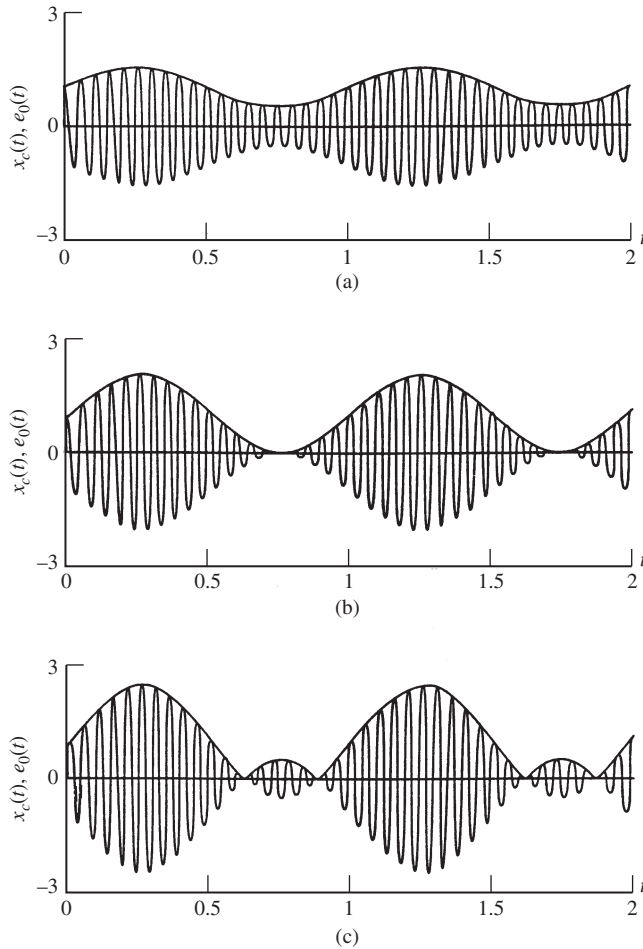


Figure 3.4
Modulated carrier and envelope detector outputs for various values of the modulation index.
(a) $a = 0.5$. (b) $a = 1.0$.
(c) $a = 1.5$.

declines rapidly as the index is reduced below unity. If the message signal does not have symmetrical maximum and minimum values, then higher values of efficiency can be achieved.

The main advantage of AM is that since a coherent reference is not needed for demodulation as long as $a \leq 1$, the demodulator becomes simple and inexpensive. In many applications, such as commercial radio, this fact alone is sufficient to justify its use.

The AM modulator output $x_c(t)$ is shown in Figure 3.4 for three values of the modulation index: $a = 0.5$, $a = 1.0$, and $a = 1.5$. The message signal $m(t)$ is assumed to be a unity amplitude sinusoid with a frequency of 1 Hz. A unity amplitude carrier is also assumed. The envelope detector output $e_0(t)$, as identified in Figure 3.3, is also shown for each value of the modulation index. Note that for $a = 0.5$ the envelope is always positive. For $a = 1.0$ the minimum value of the envelope is exactly zero. Thus, envelope detection can be used for both of these cases. For $a = 1.5$ the envelope goes negative and $e_0(t)$, which is the absolute value of the envelope, is a badly distorted version of the message signal.

EXAMPLE 3.1

In this example we determine the efficiency and the output spectrum for an AM modulator operating with a modulation index of 0.5. The carrier power is 50 W, and the message signal is

$$m(t) = 4 \cos\left(2\pi f_m t - \frac{\pi}{9}\right) + 2 \sin(4\pi f_m t) \quad (3.16)$$

The first step is to determine the minimum value of $m(t)$. There are a number of ways to accomplish this. Perhaps the easiest way is to simply plot $m(t)$ and pick off the minimum value. MATLAB is very useful for this purpose as shown in the following program. The only drawback to this approach is that $m(t)$ must be sampled at a sufficiently high frequency to ensure that the maximum value of $m(t)$ is determined with the required accuracy.

```
%File: c3ex1.m
fmt=0:0.0001:1;
m=4*cos(2*pi*fmt-pi/9) + 2*sin(4*pi*fmt);
[message,index]=min(m);
plot(fmt,m,'k'),
grid, xlabel('Normalized Time'), ylabel('Amplitude')
message, mintime=0.0001*(index-1)
%End of script file.
```

Executing the program yields the plot of the message signal, the minimum value of $m(t)$, and the occurrence time for the minimum value as follows:

```
message=-4.3642
mintime=0.4352
```

The message signal as generated by the MATLAB program is shown in Figure 3.5(a). Note that the time axis is normalized by multiplying by f_m . As shown, the minimum value of $m(t)$ is -4.364 and occurs at $f_m t = 0.435$, as shown. The normalized message signal is therefore given by

$$m_n(t) = \frac{1}{4.364} \left[4 \cos\left(2\pi f_m t - \frac{\pi}{9}\right) + 2 \sin(4\pi f_m t) \right] \quad (3.17)$$

or

$$m_n(t) = 0.9166 \cos\left(2\pi f_m t - \frac{\pi}{9}\right) + 0.4583 \sin(4\pi f_m t) \quad (3.18)$$

The mean-square value of $m_n(t)$ is

$$\langle m_n^2(t) \rangle = \frac{1}{2}(0.9166)^2 + \frac{1}{2}(0.4583)^2 = 0.5251 \quad (3.19)$$

Thus, the efficiency is

$$E_{ff} = \frac{(0.25)(0.5251)}{1 + (0.25)(0.5251)} = 0.116 \quad (3.20)$$

or 11.6%.

Since the carrier power is 50 W, we have

$$\frac{1}{2}(A_c)^2 = 50 \quad (3.21)$$

from which

$$A_c = 10 \quad (3.22)$$

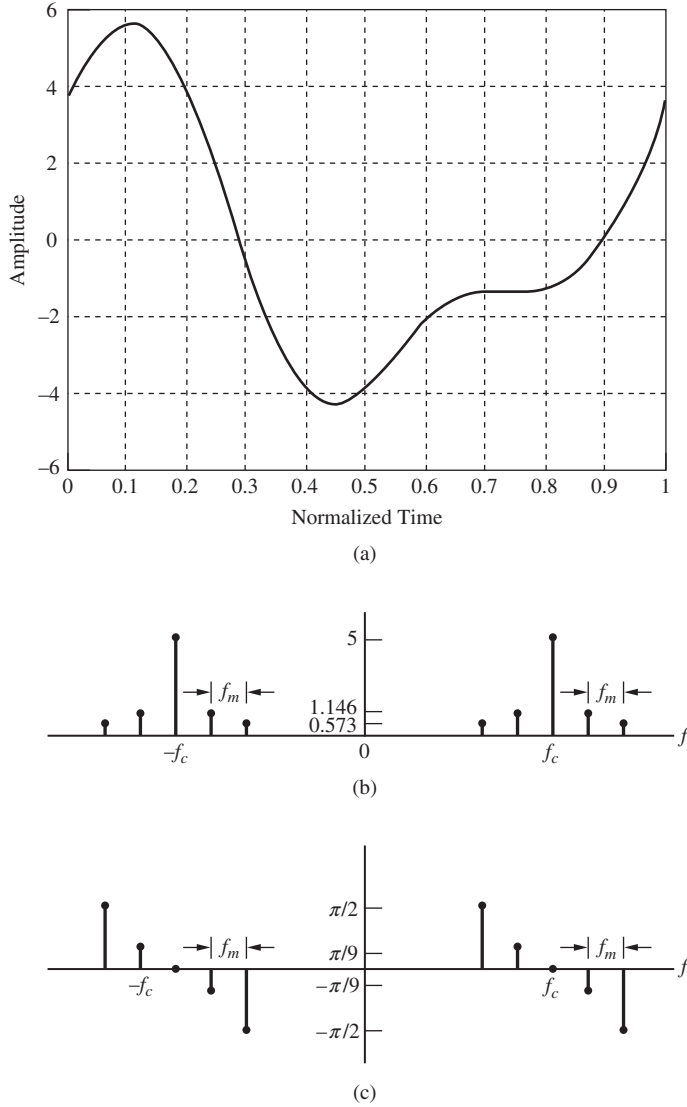


Figure 3.5 Waveform and spectra for Example 3.1. (a) Message signal. (b) Amplitude spectrum of the modulator output. (c) Phase spectrum of the modulator output.

Also, since $\sin x = \cos(x - \pi/2)$, we can write $x_c(t)$ as

$$x_c(t) = 10 \left\{ 1 + 0.5 \left[0.9166 \cos \left(2\pi f_m t - \frac{\pi}{9} \right) + 0.4583 \cos \left(4\pi f_m t - \frac{\pi}{2} \right) \right] \right\} \cos(2\pi f_c t) \quad (3.23)$$

In order to plot the spectrum of $x_c(t)$, we write the preceding equation as

$$x_c(t) = 10 \cos(2\pi f_c t)$$

$$\begin{aligned}
 &+2.292 \left\{ \cos \left[2\pi(f_c + f_m)t - \frac{\pi}{9} \right] + \cos \left[2\pi(f_c + f_m)t + \frac{\pi}{9} \right] \right\} \\
 &+1.146 \left\{ \cos \left[2\pi(f_c + 2f_m)t - \frac{\pi}{2} \right] + \cos \left[2\pi(f_c + 2f_m)t + \frac{\pi}{2} \right] \right\} \quad (3.24)
 \end{aligned}$$

Figures 3.5(b) and (c) show the amplitude and phase spectra of $x_c(t)$. Note that the amplitude spectrum has even symmetry about the carrier frequency and that the phase spectrum has odd symmetry about the carrier frequency. Of course, since $x_c(t)$ is a real signal, the overall amplitude spectrum is also even about $f = 0$, and the overall phase spectrum is odd about $f = 0$. ■

3.2.2 The Modulation Trapezoid

A nice tool for monitoring the modulation index of an AM signal is the *modulation trapezoid*. If the modulated carrier, $x_c(t)$, is placed on the vertical input to an oscilloscope and the message signal, $m(t)$, on the horizontal input, the envelope of the modulation trapezoid is produced. The basic form of the modulation trapezoid is illustrated in Figure 3.6. The trapezoid is easily interpreted and is shown in Figure 3.6 for $a < 1$. In drawing Figure 3.6 it is assumed that $\max[m_n(t)] = 1$ and that $\min[m_n(t)] = -1$, which is typically the case. Note that

$$A = 2A_c(1 + a) \quad (3.25)$$

and

$$B = 2A_c(1 - a) \quad (3.26)$$

Therefore,

$$A + B = 4A_c \quad (3.27)$$

and

$$A - B = 4A_c a \quad (3.28)$$

The modulation index is given by

$$\frac{A - B}{A + B} = \frac{4A_c a}{4A_c} = a \quad (3.29)$$

Figure 3.7 provides specific examples for $a = 0.3, 0.7, 1.0$, and 1.5 .

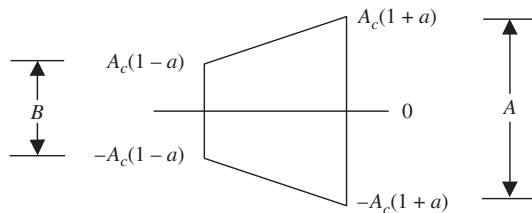


Figure 3.6

General form of the modulation trapezoid.

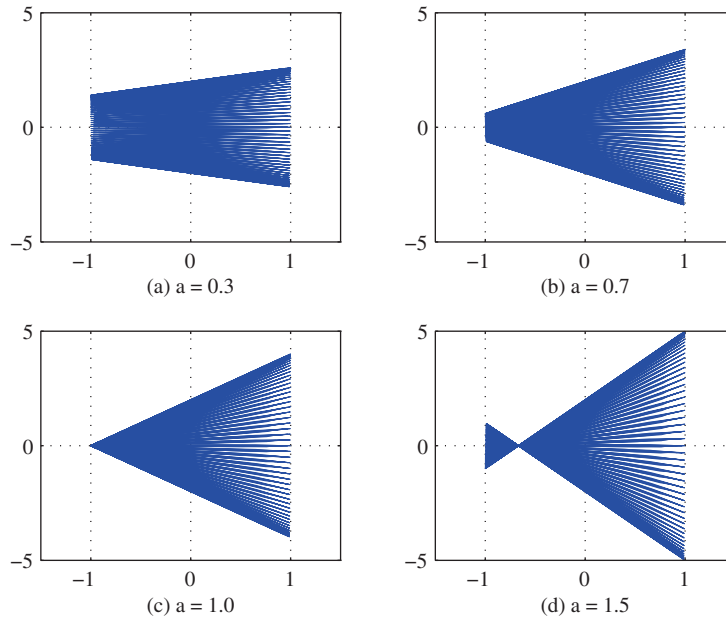


Figure 3.7
Modulation trapezoid for $a = 0.3, 0.7, 1.0,$ and 1.5 .

Note that the tops and bottoms of the envelope of the modulation trapezoid are straight lines. This provides a simple test for linearity. If the modulator/transmitter combination is not linear, the shape of the top and bottom edges of the trapezoid are no longer straight lines. Therefore, the modulation trapezoid is a test for linearity as illustrated in the following Computer Example 3.1

COMPUTER EXAMPLE 3.1

In this example we consider a modulation/transmitter combination with a third-order nonlinearity. Consider the following MATLAB program:

```
% Filename: c3ce1
a = 0.7;
fc = 2;
fm = 200.1;
t = 0:0.001:1;
m = cos(2*pi*fm*t);
c = cos(2*pi*fc*t);
xc = 2*(1+a*m).*c;
xc = xc+0.1*xc.*xc.*xc;
plot(m,xc)
axis([-1.2,1.2,-8,8])
grid
% End of script file.
```

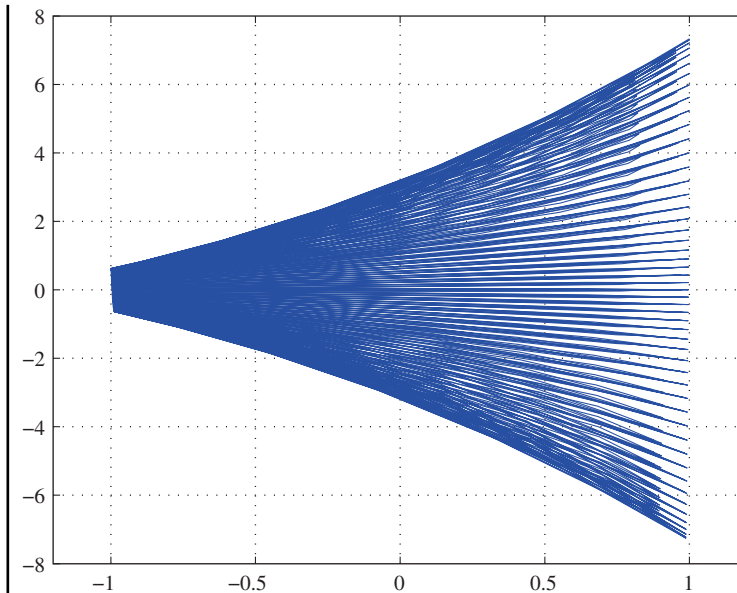


Figure 3.8
Modulation trapezoid for a modulator/transmitter having a third-order nonlinearity.

Executing this MATLAB program yields the modulation trapezoid illustrated in Figure 3.8. The effect of the nonlinearity is clear. ■

■ 3.3 SINGLE-SIDEBAND (SSB) MODULATION

In our development of DSB, we saw that the USB and LSB have even amplitude and odd phase symmetry about the carrier frequency. Thus, transmission of both sidebands is not necessary, since either sideband contains sufficient information to reconstruct the message signal $m(t)$. Elimination of one of the sidebands prior to transmission results in single-sideband (SSB), which reduces the bandwidth of the modulator output from $2W$ to W , where W is the bandwidth of $m(t)$. However, this bandwidth savings is accompanied by a considerable increase in complexity.

On the following pages, two different methods are used to derive the time-domain expression for the signal at the output of an SSB modulator. Although the two methods are equivalent, they do present different viewpoints. In the first method, the transfer function of the filter used to generate an SSB signal from a DSB signal is derived using the Hilbert transform. The second method derives the SSB signal directly from $m(t)$ using the results illustrated in Figure 2.29 and the frequency-translation theorem.

The generation of an SSB signal by sideband filtering is illustrated in Figure 3.9. First, a DSB signal, $x_{\text{DSB}}(t)$, is formed. Sideband filtering of the DSB signal then yields an upper-sideband or a lower-sideband SSB signal, depending on the filter passband selected.

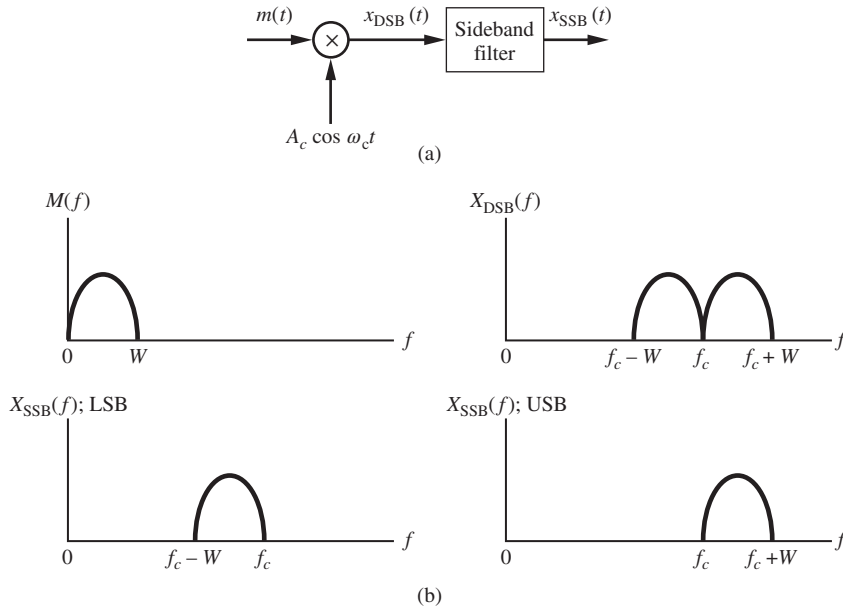


Figure 3.9 Generation of SSB by sideband filtering. (a) SSB modulator. (b) Spectra (single-sided).

The filtering process that yields lower-sideband SSB is illustrated in detail in Figure 3.10. A lower-sideband SSB signal can be generated by passing a DSB signal through an ideal filter that passes the LSB and rejects the USB. It follows from Figure 3.10(b) that the transfer function of this filter is

$$H_L(f) = \frac{1}{2} [\text{sgn}(f + f_c) - \text{sgn}(f - f_c)] \quad (3.30)$$

Since the Fourier transform of a DSB signal is

$$X_{\text{DSB}}(f) = \frac{1}{2} A_c M(f + f_c) + \frac{1}{2} A_c M(f - f_c) \quad (3.31)$$

the transform of the lower-sideband SSB signal is

$$\begin{aligned} X_c(f) &= \frac{1}{4} A_c [M(f + f_c) \text{sgn}(f + f_c) + M(f - f_c) \text{sgn}(f + f_c)] \\ &\quad - \frac{1}{4} A_c [M(f + f_c) \text{sgn}(f - f_c) + M(f - f_c) \text{sgn}(f - f_c)] \end{aligned} \quad (3.32)$$

which is

$$\begin{aligned} X_c(f) &= \frac{1}{4} A_c [M(f + f_c) + M(f - f_c)] \\ &\quad + \frac{1}{4} A_c [M(f + f_c) \text{sgn}(f + f_c) - M(f - f_c) \text{sgn}(f - f_c)] \end{aligned} \quad (3.33)$$

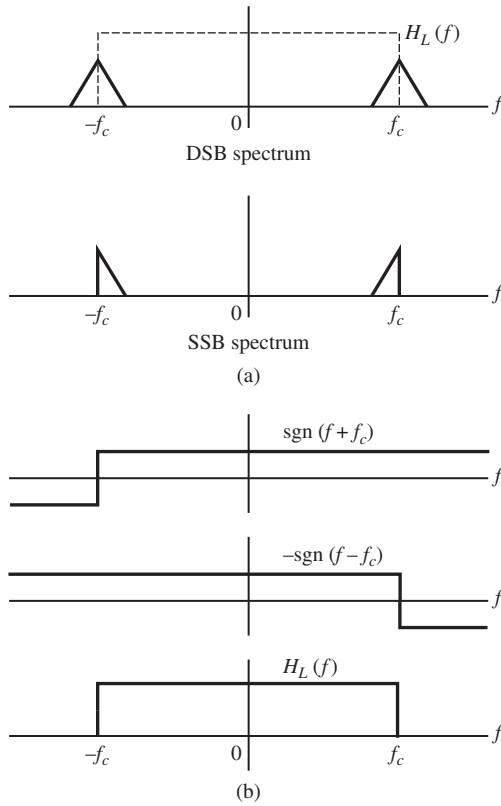


Figure 3.10
 Generation of lower-sideband SSB.
 (a) Sideband filtering process. (b) Generation of lower-sideband filter.

From our study of DSB, we know that

$$\frac{1}{2}A_c m(t) \cos(2\pi f_c t) \leftrightarrow \frac{1}{4}A_c [M(f + f_c) + M(f - f_c)] \quad (3.34)$$

and from our study of Hilbert transforms in Chapter 2, we recall that

$$\hat{m}(t) \leftrightarrow -j(\text{sgn } f)M(f)$$

By the frequency-translation theorem, we have

$$m(t)e^{\pm j2\pi f_c t} \leftrightarrow M(f \mp f_c) \quad (3.35)$$

Replacing $m(t)$ by $\hat{m}(t)$ in the previous equation yields

$$\hat{m}(t)e^{\pm j2\pi f_c t} \leftrightarrow -jM(f \mp f_c)\text{sgn}(f \mp f_c) \quad (3.36)$$

Thus,

$$\begin{aligned} & \mathfrak{F}^{-1} \left\{ \frac{1}{4}A_c [M(f + f_c)\text{sgn}(f + f_c) - M(f - f_c)\text{sgn}(f - f_c)] \right\} \\ &= -A_c \frac{1}{4j} \hat{m}(t)e^{-j2\pi f_c t} + A_c \frac{1}{4j} \hat{m}(t)e^{+j2\pi f_c t} = \frac{1}{2}A_c \hat{m}(t) \sin(2\pi f_c t) \end{aligned} \quad (3.37)$$

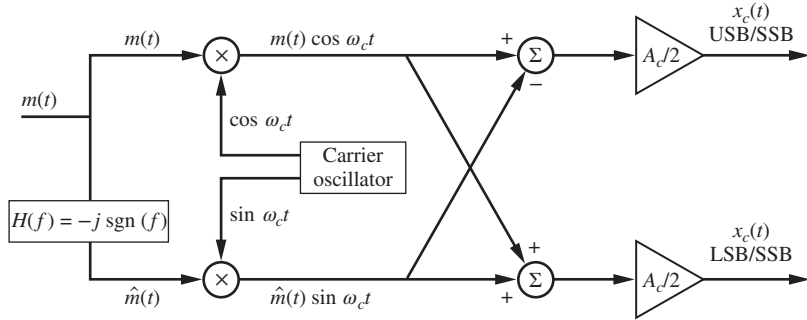


Figure 3.11
Phase-shift modulator.

Combining (3.34) and (3.37), we get the general form of a lower-sideband SSB signal:

$$x_c(t) = \frac{1}{2} A_c m(t) \cos(2\pi f_c t) + \frac{1}{2} A_c \hat{m}(t) \sin(2\pi f_c t) \quad (3.38)$$

A similar development can be carried out for upper-sideband SSB. The result is

$$x_c(t) = \frac{1}{2} A_c m(t) \cos(2\pi f_c t) - \frac{1}{2} A_c \hat{m}(t) \sin(2\pi f_c t) \quad (3.39)$$

which shows that LSB and USB modulators have the same defining equations except for the sign of the term representing the Hilbert transform of the modulation. Observation of the spectrum of an SSB signal illustrates that SSB systems do not have DC response.

The generation of SSB by the method of sideband filtering the output of DSB modulators requires the use of filters that are very nearly ideal if low-frequency information is contained in $m(t)$. Another method for generating an SSB signal, known as *phase-shift modulation*, is illustrated in Figure 3.11. This system is a term-by-term realization of (3.38) or (3.39). Like the ideal filters required for sideband filtering, the ideal wideband phase shifter, which performs the Hilbert transforming operation, is impossible to implement exactly. However, since the frequency at which the discontinuity occurs is $f = 0$ instead of $f = f_c$, ideal phase-shift devices can be closely approximated.

An alternative derivation of $x_c(t)$ for an SSB signal is based on the concept of the analytic signal. As shown in Figure 3.12(a), the positive-frequency portion of $M(f)$ is given by

$$M_p(f) = \frac{1}{2} \mathfrak{F}\{m(t) + j\hat{m}(t)\} \quad (3.40)$$

and the negative-frequency portion of $M(f)$ is given by

$$M_n(f) = \frac{1}{2} \mathfrak{F}\{m(t) - j\hat{m}(t)\} \quad (3.41)$$

By definition, an upper-sideband SSB signal is given in the frequency domain by

$$X_c(f) = \frac{1}{2} A_c M_p(f - f_c) + \frac{1}{2} A_c M_n(f + f_c) \quad (3.42)$$

Inverse Fourier-transforming yields

$$x_c(t) = \frac{1}{4}A_c[m(t) + j\hat{m}(t)]e^{j2\pi f_c t} + \frac{1}{4}A_c[m(t) - j\hat{m}(t)]e^{-j2\pi f_c t} \quad (3.43)$$

which is

$$\begin{aligned} x_c(t) &= \frac{1}{4}A_c m(t)[e^{j2\pi f_c t} + e^{-j2\pi f_c t}] + j\frac{1}{4}A_c \hat{m}(t)[e^{j2\pi f_c t} - e^{-j2\pi f_c t}] \\ &= \frac{1}{2}A_c m(t) \cos(2\pi f_c t) - \frac{1}{2}A_c \hat{m}(t) \sin(2\pi f_c t) \end{aligned} \quad (3.44)$$

The preceding expression is clearly equivalent to (3.39).

The lower-sideband SSB signal is derived in a similar manner. By definition, for a lower-sideband SSB signal,

$$X_c(f) = \frac{1}{2}A_c M_p(f + f_c) + \frac{1}{2}A_c M_n(f - f_c) \quad (3.45)$$

This becomes, after inverse Fourier-transforming,

$$x_c(t) = \frac{1}{4}A_c[m(t) + j\hat{m}(t)]e^{-j2\pi f_c t} + \frac{1}{4}A_c[m(t) - j\hat{m}(t)]e^{j2\pi f_c t} \quad (3.46)$$

which can be written as

$$\begin{aligned} x_c(t) &= \frac{1}{4}A_c m(t)[e^{j2\pi f_c t} + e^{-j2\pi f_c t}] - j\frac{1}{4}A_c \hat{m}(t)[e^{j2\pi f_c t} - e^{-j2\pi f_c t}] \\ &= \frac{1}{2}A_c m(t) \cos(2\pi f_c t) + \frac{1}{2}A_c \hat{m}(t) \sin(2\pi f_c t) \end{aligned}$$

This expression is clearly equivalent to (3.38). Figures 3.12(b) and (c) show the four signal spectra used in this development: $M_p(f + f_c)$, $M_p(f - f_c)$, $M_n(f + f_c)$, and $M_n(f - f_c)$.

There are several methods that can be employed to demodulate SSB. The simplest technique is to multiply $x_c(t)$ by a demodulation carrier and lowpass filter the result, as illustrated in Figure 3.1(a). We assume a demodulation carrier having a phase error $\theta(t)$ that yields

$$d(t) = \left[\frac{1}{2}A_c m(t) \cos(2\pi f_c t) \pm \frac{1}{2}A_c \hat{m}(t) \sin(2\pi f_c t) \right] \{4 \cos[2\pi f_c t + \theta(t)]\} \quad (3.47)$$

where the factor of 4 is chosen for mathematical convenience. The preceding expression can be written as

$$\begin{aligned} d(t) &= A_c m(t) \cos \theta(t) + A_c m(t) \cos[4\pi f_c t + \theta(t)] \\ &\mp A_c \hat{m}(t) \sin \theta(t) \pm A_c \hat{m}(t) \sin[4\pi f_c t + \theta(t)] \end{aligned} \quad (3.48)$$

Lowpass filtering and amplitude scaling yield

$$y_D(t) = m(t) \cos \theta(t) \mp \hat{m}(t) \sin \theta(t) \quad (3.49)$$

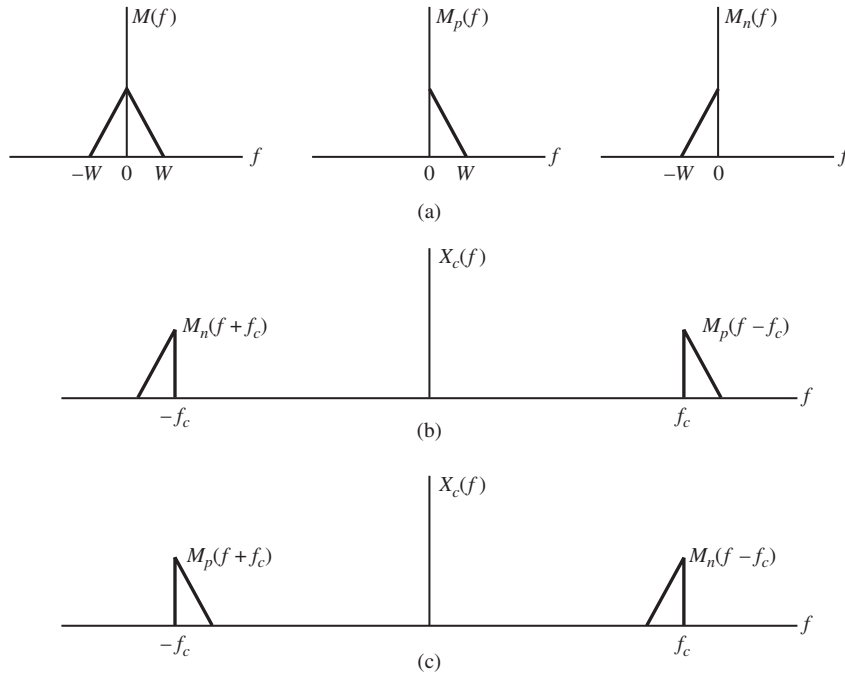


Figure 3.12 Alternative derivation of SSB signals. (a) $M(f)$, $M_p(f)$, and $M_n(f)$. (b) Upper-sideband SSB signal. (c) Lower-sideband SSB signal.

for the demodulated output. Observation of (3.49) illustrates that for $\theta(t)$ equal to zero, the demodulated output is the desired message signal. However, if $\theta(t)$ is nonzero, the output consists of the sum of two terms. The first term is a time-varying attenuation of the message signal and is the output present in a DSB system operating in a similar manner. The second term is a crosstalk term and can represent serious distortion if $\theta(t)$ is not small.

Another useful technique for demodulating an SSB signal is carrier reinsertion, which is illustrated in Figure 3.13. The output of a local oscillator is added to the received signal $x_r(t)$. This yields

$$e(t) = \left[\frac{1}{2} A_c m(t) + K \right] \cos(2\pi f_c t) \pm \frac{1}{2} A_c \hat{m}(t) \sin(2\pi f_c t) \quad (3.50)$$

which is the input to the envelope detector. The output of the envelope detector must next be computed. This is slightly more difficult for signals of the form of (3.50) than for signals of the form of (3.10) because both cosine and sine terms are present. In order to derive the

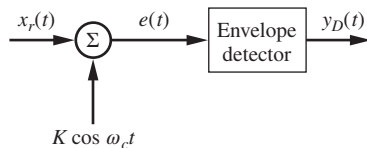


Figure 3.13 Demodulation using carrier reinsertion.

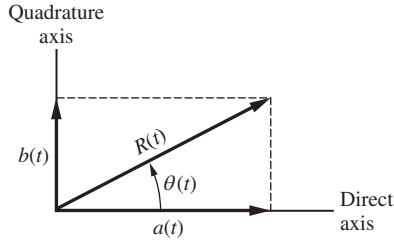


Figure 3.14
Direct-quadrature signal representation.

desired result, consider the signal

$$x(t) = a(t) \cos(2\pi f_c t) - b(t) \sin(2\pi f_c t) \quad (3.51)$$

which can be represented as illustrated in Figure 3.14. Figure 3.14 shows the amplitude of the direct component $a(t)$, the amplitude of the quadrature component $b(t)$, and the resultant $R(t)$. It follows from Figure 3.14 that

$$a(t) = R(t) \cos \theta(t) \quad \text{and} \quad b(t) = R(t) \sin \theta(t)$$

This yields

$$x(t) = R(t) [\cos \theta(t) \cos(2\pi f_c t) - \sin \theta(t) \sin(2\pi f_c t)] \quad (3.52)$$

which is

$$x(t) = R(t) \cos[2\pi f_c t + \theta(t)] \quad (3.53)$$

where

$$\theta(t) = \tan^{-1} \left(\frac{b(t)}{a(t)} \right) \quad (3.54)$$

The instantaneous amplitude $R(t)$, which is the envelope of the signal, is given by

$$R(t) = \sqrt{a^2(t) + b^2(t)} \quad (3.55)$$

and will be the output of an envelope detector with $x(t)$ on the input if $a(t)$ and $b(t)$ are slowly varying with respect to $\cos \omega_c t$.

A comparison of (3.50) and (3.55) illustrates that the envelope of an SSB signal, after carrier reinsertion, is given by

$$y_D(t) = \sqrt{\left[\frac{1}{2} A_c m(t) + K \right]^2 + \left[\frac{1}{2} A_c \hat{m}(t) \right]^2} \quad (3.56)$$

which is the demodulated output $y_D(t)$ in Figure 3.13. If K is chosen large enough such that

$$\left[\frac{1}{2} A_c m(t) + K \right]^2 \gg \left[\frac{1}{2} A_c \hat{m}(t) \right]^2$$

the output of the envelope detector becomes

$$y_D(t) \cong \frac{1}{2} A_c m(t) + K \quad (3.57)$$

from which the message signal can easily be extracted. The development shows that carrier reinsertion requires that the locally generated carrier must be phase coherent with the original modulation carrier. This is easily accomplished in speech-transmission systems. The frequency and phase of the demodulation carrier can be manually adjusted until intelligibility of the speech is obtained.

EXAMPLE 3.2

As we saw in the preceding analysis, the concept of single sideband is probably best understood by using frequency-domain analysis. However, the SSB time-domain waveforms are also interesting and are the subject of this example. Assume that the message signal is given by

$$m(t) = \cos(2\pi f_1 t) - 0.4 \cos(4\pi f_1 t) + 0.9 \cos(6\pi f_1 t) \quad (3.58)$$

The Hilbert transform of $m(t)$ is

$$\hat{m}(t) = \sin(2\pi f_1 t) - 0.4 \sin(4\pi f_1 t) + 0.9 \sin(6\pi f_1 t) \quad (3.59)$$

These two waveforms are shown in Figures 3.15(a) and (b).

As we have seen, the SSB signal is given by

$$x_c(t) = \frac{A_c}{2} [m(t) \cos(2\pi f_c t) \pm \hat{m}(t) \sin(2\pi f_c t)] \quad (3.60)$$

with the choice of sign depending upon the sideband to be used for transmission. Using (3.51) to (3.55), we can place $x_c(t)$ in the standard form of (3.1). This gives

$$x_c(t) = R(t) \cos[2\pi f_c t + \theta(t)] \quad (3.61)$$

where the envelope $R(t)$ is

$$R(t) = \frac{A_c}{2} \sqrt{m^2(t) + \hat{m}^2(t)} \quad (3.62)$$

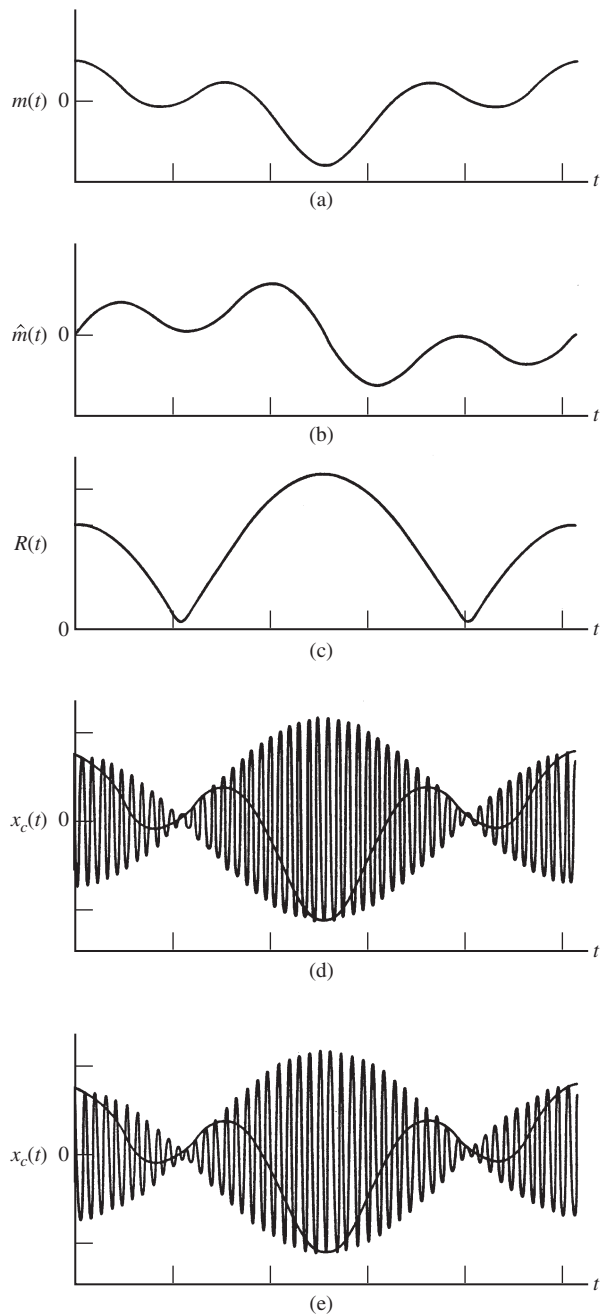
and $\theta(t)$, which is the phase deviation of $x_c(t)$, is given by

$$\theta(t) = \pm \tan^{-1} \left(\frac{\hat{m}(t)}{m(t)} \right) \quad (3.63)$$

The instantaneous frequency of $\theta(t)$ is therefore

$$\frac{d}{dt} [2\pi f_c t + \theta(t)] = 2\pi f_c \pm \frac{d}{dt} \left[\tan^{-1} \left(\frac{\hat{m}(t)}{m(t)} \right) \right] \quad (3.64)$$

From (3.62) we see that the envelope of the SSB signal is independent of the choice of the sideband. The instantaneous frequency, however, is a rather complicated function of the message signal and also depends upon the choice of sideband. We therefore see that the message signal $m(t)$ affects both the envelope and phase of the modulated carrier $x_c(t)$. In DSB and AM the message signal affected only the envelope of $x_c(t)$.

**Figure 3.15**

Time-domain signals for SSB system. (a) Message signal. (b) Hilbert transform of the message signal. (c) Envelope of the SSB signal. (d) Upper-sideband SSB signal with message signal. (e) Lower-sideband SSB signal with message signal.

The envelope of the SSB signal, $R(t)$, is shown in Figure 3.15(c). The upper-sideband SSB signal is illustrated in Figure 3.15(d) and the lower-sideband SSB signal is shown in Figure 3.15(e). It is easily seen that both the upper-sideband and lower-sideband SSB signals have the envelope shown in Figure 3.15(c). The message signal $m(t)$ is also shown in Figures 3.15(d) and (e).

3.4 VESTIGIAL-SIDEBAND (VSB) MODULATION

We have seen that DSB requires excessive bandwidth and that generation of SSB by sideband filtering can only be approximately realized. In addition SSB has poor low-frequency performance. *Vestigial-sideband (VSB) modulation* offers a compromise by allowing a small amount, or vestige, of the unwanted sideband to appear at the output of an SSB modulator, the design of the sideband filter is simplified, since the need for sharp cutoff at the carrier frequency is eliminated. In addition, a VSB system has improved low-frequency response compared to SSB and can even have DC response. A simple example will illustrate the technique.

EXAMPLE 3.3

For simplicity, let the message signal be the sum of two sinusoids:

$$m(t) = A \cos(2\pi f_1 t) + B \cos(2\pi f_2 t) \quad (3.65)$$

This message signal is then multiplied by a carrier, $\cos(2\pi f_c t)$, to form a DSB signal

$$\begin{aligned} e_{DSB}(t) = & \frac{1}{2} A \cos[2\pi(f_c - f_1)t] + \frac{1}{2} A \cos[2\pi(f_c + f_1)t] \\ & + \frac{1}{2} B \cos[2\pi(f_c - f_2)t] + \frac{1}{2} B \cos[2\pi(f_c + f_2)t] \end{aligned} \quad (3.66)$$

Figure 3.16(a) shows the single-sided spectrum of this signal. Prior to transmission a VSB filter is used to generate the VSB signal. Figure 3.16(b) shows the assumed amplitude response of the VSB filter. The skirt of the VSB filter must have the symmetry about the carrier frequency as shown. Figure 3.16(c) shows the single-sided spectrum of the VSB filter output.

Assume that the VSB filter has the following amplitude and phase responses:

$$H(f_c - f_2) = 0, \quad H(f_c - f_1) = \epsilon e^{-j\theta_a} \quad H(f_c + f_1) = (1 - \epsilon)e^{-j\theta_b}, \quad H(f_c + f_2) = 1e^{-j\theta_c} \quad (3.67)$$

The VSB filter input is the DSB signal that, in complex envelope form, can be expressed as

$$x_{DSB}(t) = \text{Re} \left[\left(\frac{A}{2} e^{-j2\pi f_1 t} + \frac{A}{2} e^{j2\pi f_1 t} + \frac{B}{2} e^{-j2\pi f_2 t} + \frac{B}{2} e^{j2\pi f_2 t} \right) e^{j2\pi f_c t} \right] \quad (3.68)$$

Using the amplitude and phase characteristics of the VSB filter yields the VSB signal

$$x_c(t) = \text{Re} \left\{ \left[\frac{A}{2} \epsilon e^{-j(2\pi f_1 t + \theta_a)} + \frac{A}{2} (1 - \epsilon) e^{j(2\pi f_1 t - \theta_b)} + \frac{B}{2} e^{j(2\pi f_2 t - \theta_c)} \right] e^{j2\pi f_c t} \right\} \quad (3.69)$$

Demodulation is accomplished by multiplying by $2e^{-j2\pi f_c t}$ and taking the real part. This gives

$$e(t) = A\epsilon \cos(2\pi f_1 t + \theta_a) + A(1 - \epsilon) \cos(2\pi f_1 t - \theta_b) + B \cos(2\pi f_2 t - \theta_c) \quad (3.70)$$

In order for the first two terms to combine as in (3.70), we must satisfy

$$\theta_a = -\theta_b \quad (3.71)$$

which shows that the phase response must have odd symmetry about f_c and, in addition, since $e(t)$ is real, the phase response of the VSB filter must also have odd phase response about $f = 0$. With $\theta_a = -\theta_b$ we have

$$e(t) = A \cos(2\pi f_1 t - \theta_b) + B \cos(2\pi f_2 t - \theta_c) \quad (3.72)$$

We still must determine the relationship between θ_c and θ_b .

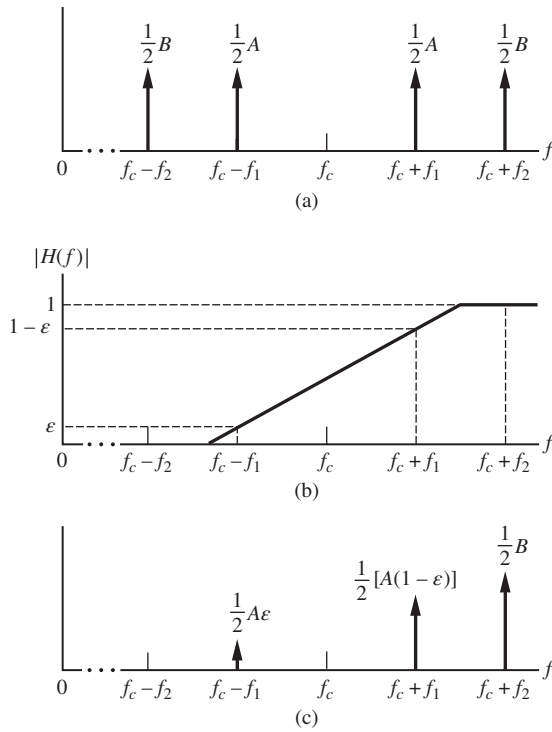


Figure 3.16
 Generation of vestigial-sideband modulation. (a) DSB magnitude spectrum (single sided). (b) VSB filter characteristic near f_c . (c) VSB magnitude spectrum.

As we saw in Chapter 2, in order for the demodulated signal $e(t)$ to be an undistorted (no amplitude or phase distortion) version of the original message signal $m(t)$, $e(t)$ must be an amplitude scaled and time-delayed version of $m(t)$. In other words

$$e(t) = Km(t - \tau) \tag{3.73}$$

Clearly the amplitude scaling $K = 1$. With time delay τ , $e(t)$ is

$$e(t) = A \cos[2\pi f_1(t - \tau)] + B \cos[2\pi f_2(t - \tau)] \tag{3.74}$$

Comparing (3.72) and (3.74) shows that

$$\theta_b = 2\pi f_1 \tau \tag{3.75}$$

and

$$\theta_c = 2\pi f_2 \tau \tag{3.76}$$

In order to have no phase distortion, the time delay must be the same for both components of $e(t)$. This gives

$$\theta_c = \frac{f_2}{f_1} \theta_b \tag{3.77}$$

We therefore see that the phase response of the VSB filter must be linear over the bandwidth of the input signal, which was to be expected from our discussion of distortionless systems in Chapter 2. ■

The slight increase in bandwidth required for VSB over that required for SSB is often more than offset by the resulting implementation simplifications. As a matter of fact, if a carrier component is added to a VSB signal, envelope detection can be used. The development

of this technique is similar to the development of envelope detection of SSB with carrier reinsertion and is relegated to the problems. The process, however, is demonstrated in the following example.

EXAMPLE 3.4

In this example we consider the time-domain waveforms corresponding to VSB modulation and consider demodulation using envelope detection or carrier reinsertion. We assume the same message signal

$$m(t) = \cos(2\pi f_1 t) - 0.4 \cos(4\pi f_1 t) + 0.9 \cos(6\pi f_1 t) \quad (3.78)$$

The message signal $m(t)$ is shown in Figure 3.17(a). The VSB signal can be expressed as

$$\begin{aligned} x_c(t) = A_c [& \epsilon_1 \cos[2\pi(f_c - f_1)t] + (1 - \epsilon_1) \cos[2\pi(f_c + f_1)t] \\ & - 0.4\epsilon_2 \cos[2\pi(f_c - 2f_1)t] - 0.4(1 - \epsilon_2) \cos[2\pi(f_c + 2f_1)t] \\ & + 0.9\epsilon_3 \cos[2\pi(f_c - 3f_1)t] + 0.9(1 - \epsilon_3) \cos[2\pi(f_c + 3f_1)t] \end{aligned} \quad (3.79)$$

The modulated carrier, along with the message signal, is shown in Figure 3.17(b) for $\epsilon_1 = 0.64$, $\epsilon_2 = 0.78$, and $\epsilon_3 = 0.92$. The result of carrier reinsertion and envelope detection is shown in Figure 3.17(c). The

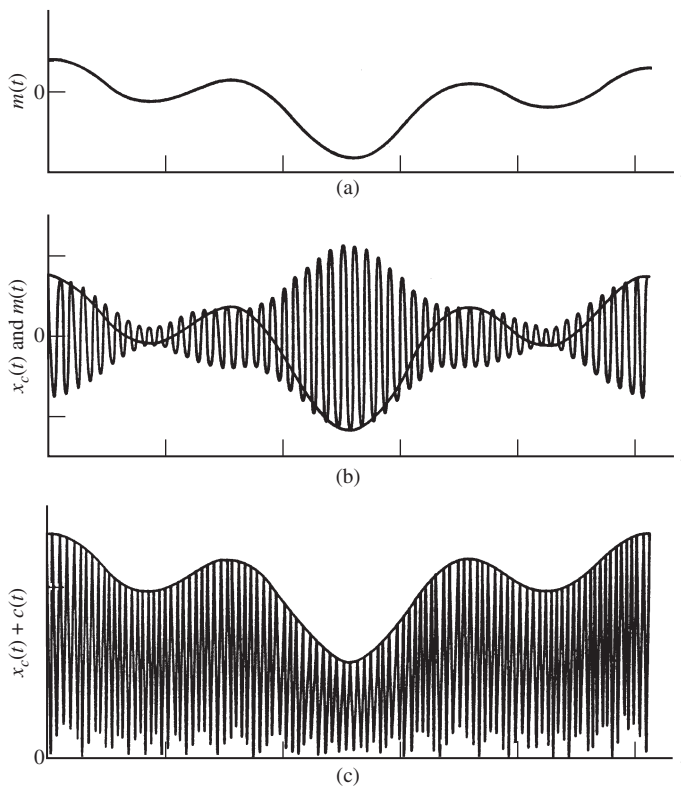


Figure 3.17 Time-domain signals for VSB system. (a) Message signal. (b) VSB signal and message signal. (c) Sum of VSB signal and carrier signal.

message signal, biased by the amplitude of the carrier component, is clearly shown and will be the output of an envelope detector. ■

3.5 FREQUENCY TRANSLATION AND MIXING

It is often desirable to translate a bandpass signal to a new center frequency. Frequency translation is used in the implementation of communications receivers as well as in a number of other applications. The process of frequency translation can be accomplished by multiplication of the bandpass signal by a periodic signal and is referred to as *mixing*. A block diagram of a mixer is given in Figure 3.18. As an example, the bandpass signal $m(t) \cos(2\pi f_1 t)$ can be translated from f_1 to a new carrier frequency f_2 by multiplying it by a local oscillator signal of the form $2 \cos[2\pi(f_1 \pm f_2)t]$. By using appropriate trigonometric identities, we can easily show that the result of the multiplication is

$$e(t) = m(t) \cos(2\pi f_2 t) + m(t) \cos(4\pi f_1 \pm 2\pi f_2)t \quad (3.80)$$

The undesired term is removed by filtering. The filter should have a bandwidth at least $2W$ for the assumed DSB modulation, where W is the bandwidth of $m(t)$.

A common problem with mixers results from the fact that two different input signals can be translated to the same frequency, f_2 . For example, inputs of the form $k(t) \cos[2\pi f_1 \pm 2f_2)t]$ are also translated to f_2 , since

$$\begin{aligned} 2k(t) \cos[2\pi(f_1 \pm 2f_2)t] \cos[2\pi(f_1 \pm f_2)t] &= k(t) \cos(2\pi f_2 t) \\ &+ k(t) \cos[2\pi(2f_1 \pm 3f_2)t] \end{aligned} \quad (3.81)$$

In (3.81), all three signs must be plus or all three signs must be minus. The input frequency $f_1 \pm 2f_2$, which results in an output at f_2 , is referred to as the *image frequency* of the desired frequency f_1 .

To illustrate that image frequencies must be considered in receiver design, consider the superheterodyne receiver shown in Figure 3.19. The carrier frequency of the signal to be demodulated is f_c , and the intermediate-frequency (IF) filter is a bandpass filter with center frequency f_{IF} , which is fixed. The superheterodyne receiver has good sensitivity (the ability to detect weak signals) and selectivity (the ability to separate closely spaced signals). This results because the IF filter, which provides most of the predetection filtering, need not be tunable. Thus, it can be a rather complex filter. Tuning of the receiver is accomplished by varying the

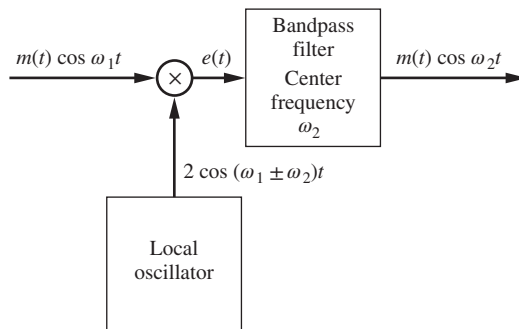


Figure 3.18
Mixer.

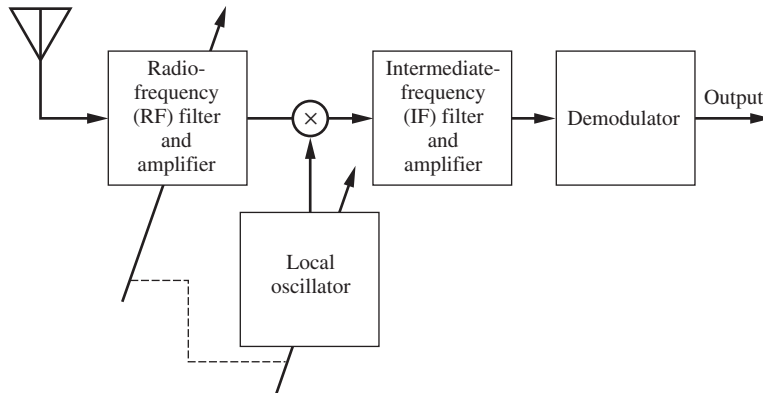


Figure 3.19
Superheterodyne receiver.

frequency of the local oscillator. The superheterodyne receiver of Figure 3.19 is the mixer of Figure 3.18 with $f_c = f_1$ and $f_{IF} = f_2$. The mixer translates the input frequency f_c to the IF frequency f_{IF} .

As shown previously, the image frequency $f_c \pm 2f_{IF}$, where the sign depends on the choice of local oscillator frequency, also will appear at the IF output. This means that if we are attempting to receive a signal having carrier frequency f_c , we can also receive a signal at $f_c + 2f_{IF}$ if the local oscillator frequency is $f_c + f_{IF}$ or a signal at $f_c - 2f_{IF}$ if the local oscillator frequency is $f_c - f_{IF}$. There is only one image frequency, and it is always separated from the desired frequency by $2f_{IF}$. Figure 3.20 shows the desired signal and image signal for

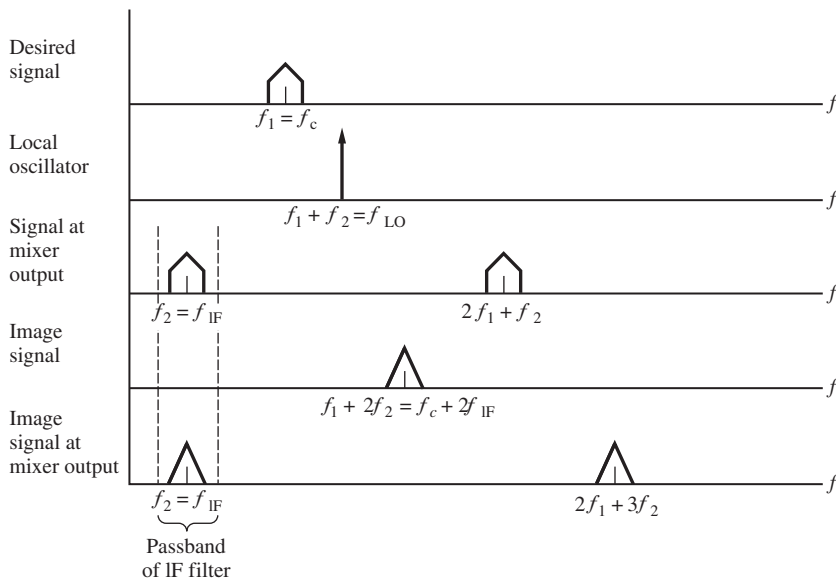


Figure 3.20
Illustration of image frequencies (high-side tuning).

Table 3.1 Low-Side and High-Side Tuning for AM Broadcast Band with $f_{IF} = 455$ kHz

	Lower frequency	Upper frequency	Tuning range of local oscillator
Standard AM broadcast band	540 kHz	1600 kHz	
Frequencies of local oscillator for low-side tuning	540 kHz – 455 kHz = 85 kHz	1600 kHz – 455 kHz = 1145 kHz	13.47 to 1
Frequencies of local oscillator for high-side tuning	540 kHz + 455 kHz = 995 kHz	1600 kHz + 455 kHz = 2055 kHz	2.07 to 1

a local oscillator having the frequency

$$f_{LO} = f_c + f_{IF} \quad (3.82)$$

The image frequency can be eliminated by the radio-frequency (RF) filter. A standard IF frequency for AM radio is 455 kHz. Thus, the image frequency is separated from the desired signal by almost 1 MHz. This shows that the RF filter need not be narrowband. Furthermore, since the AM broadcast band occupies the frequency range 540 kHz to 1.6 MHz, it is apparent that a tunable RF filter is not required, provided that stations at the high end of the band are not located geographically near stations at the low end of the band. Some inexpensive receivers take advantage of this fact. Additionally, if the RF filter is made tunable, it need be tunable only over a narrow range of frequencies.

One decision to be made when designing a superheterodyne receiver is whether the frequency of the local oscillator is to be below the frequency of the input carrier (*low-side tuning*) or above the frequency of the input carrier (*high-side tuning*). A simple example based on the standard AM broadcast band illustrates one major consideration. The standard AM

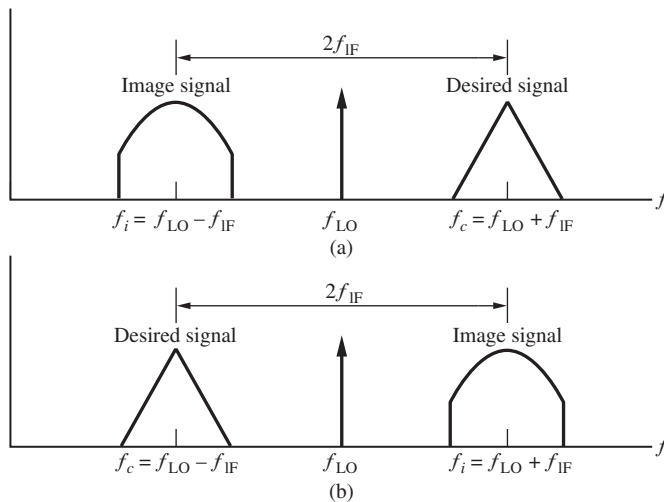


Figure 3.21
Relationship between f_c and f_i (a) low-side tuning and (b) high-side tuning.

broadcast band extends from 540 kHz to 1600 kHz. For this example, let us choose a common intermediate frequency, 455 kHz. As shown in Table 3.1, for low-side tuning, the frequency of the local oscillator must be variable from 85 to 1600 kHz, which represents a frequency range in excess of 13 to 1. If high-side tuning is used, the frequency of the local oscillator must be variable from 995 to 2055 kHz, which represents a frequency range slightly in excess of 2 to 1. Oscillators whose frequency must vary over a large ratio are much more difficult to implement than are those whose frequency varies over a small ratio.

The relationship between the desired signal to be demodulated and the image signal is summarized in Figure 3.21 for low-side and high-side tuning. The desired signal to be demodulated has a carrier frequency of f_c and the image signal has a carrier frequency of f_i .

3.6 INTERFERENCE IN LINEAR MODULATION

We now consider the effect of interference in communication systems. In real-world systems interference occurs from various sources, such as RF emissions from transmitters having carrier frequencies close to that of the carrier being demodulated. We also study interference because the analysis of systems in the presence of interference provides us with important insights into the behavior of systems operating in the presence of noise, which is the topic of Chapter 8. In this section we consider only interference in linear modulation. Interference in angle modulation will be treated in the next chapter.

As a simple case of linear modulation in the presence of interference, we consider the received signal having the spectrum (single sided) shown in Figure 3.22. The received signal consists of three components: a carrier component, a pair of sidebands representing a sinusoidal message signal, and an undesired interfering tone of frequency $f_c + f_i$. The input to the demodulator is therefore

$$x_c(t) = A_c \cos(2\pi f_c t) + A_i \cos[2\pi(f_c + f_i)t] + A_m \cos(2\pi f_m t) \cos(2\pi f_c t) \quad (3.83)$$

Multiplying $x_c(t)$ by $2 \cos(2\pi f_c t)$ and lowpass filtering (coherent demodulation) yields

$$y_D(t) = A_m \cos(2\pi f_m t) + A_i \cos(2\pi f_i t) \quad (3.84)$$

where we have assumed that the interference component is passed by the filter and that the DC term resulting from the carrier is blocked. From this simple example we see that the signal and interference are additive at the receiver output if the interference is additive at the receiver input. This result was obtained because the coherent demodulator operates as a linear demodulator.

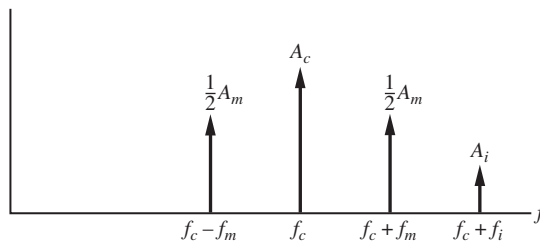


Figure 3.22
Assumed received-signal spectrum.

The effect of interference with envelope detection is quite different because of the non-linear nature of the envelope detector. The analysis with envelope detection is much more difficult than the coherent demodulation case. Some insight can be gained by writing $x_c(t)$ in a form that leads to the phasor diagram. In order to develop the phasor diagram, we write (3.83) in the form

$$x_r(t) = \text{Re} \left[\left(A_c + A_i e^{j2\pi f_i t} + \frac{1}{2} A_m e^{j2\pi f_m t} + \frac{1}{2} A_m e^{-j2\pi f_m t} \right) e^{j2\pi f_c t} \right] \quad (3.85)$$

The phasor diagram is constructed with respect to the carrier by taking the carrier frequency as equal to zero. In other words, we plot the phasor diagram corresponding to the complex envelope signal. The phasor diagrams are illustrated in Figure 3.23, both with and without interference. The output of an ideal envelope detector is $R(t)$ in both cases. The phasor diagrams illustrate that interference induces both an amplitude distortion and a phase deviation.

The effect of interference with envelope detection is determined by writing (3.83) as

$$\begin{aligned} x_r(t) = & A_c \cos(2\pi f_c t) + A_m \cos(2\pi f_m t) \cos(2\pi f_c t) \\ & + A_i [\cos(2\pi f_c t) \cos(2\pi f_i t) - \sin(2\pi f_c t) \sin(2\pi f_i t)] \end{aligned} \quad (3.86)$$

which is

$$x_r(t) = [A_c + A_m \cos(2\pi f_c t) + A_i \cos(2\pi f_i t)] \cos(2\pi f_c t) - A_i \sin(2\pi f_i t) \sin(2\pi f_c t) \quad (3.87)$$

If $A_c \gg A_i$, which is the usual case of interest, the last term in (3.87) is negligible compared to the first term and the output of the envelope detector is

$$y_D(t) \cong A_m \cos(2\pi f_m t) + A_i \cos(2\pi f_i t) \quad (3.88)$$

assuming that the DC term is blocked. Thus, for the small interference case, envelope detection and coherent demodulation are essentially equivalent.

If $A_c \ll A_i$, the assumption cannot be made that the last term of (3.87) is negligible, and the output is significantly different. To show this, (3.83) is rewritten as

$$\begin{aligned} x_r(t) = & A_c \cos[2\pi(f_c + f_i - f_i)t] + A_i \cos[2\pi(f_c + f_i)t] \\ & + A_m \cos(2\pi f_m t) \cos[2\pi(f_c + f_i - f_i)t] \end{aligned} \quad (3.89)$$

which, when we use appropriate trigonometric identities, becomes

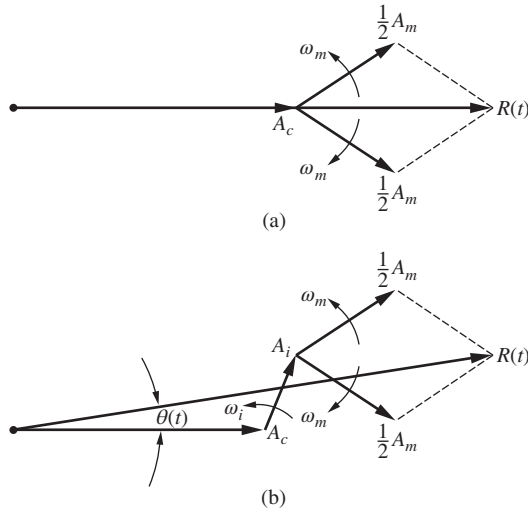
$$\begin{aligned} x_r(t) = & A_c \{ \cos[2\pi(f_c + f_i)t] \cos(2\pi f_i t) + \sin[2\pi(f_c + f_i)t] \sin(2\pi f_i t) \} \\ & + A_i \cos[2\pi(f_c + f_i)t] + A_m \cos(2\pi f_m t) \{ \cos[2\pi(f_c + f_i)t] \cos(2\pi f_i t) \\ & + \sin[2\pi(f_c + f_i)t] \sin(2\pi f_i t) \} \end{aligned} \quad (3.90)$$

Equation (3.90) can also be written as

$$\begin{aligned} x_r(t) = & [A_i + A_c \cos(2\pi f_i t) + A_m \cos(2\pi f_m t) \cos(2\pi f_i t)] \cos[2\pi(f_c + f_i)t] \\ & + [A_c \sin(2\pi f_i t) + A_m \cos(2\pi f_m t) \sin(2\pi f_i t)] \sin[2\pi(f_c + f_i)t] \end{aligned} \quad (3.91)$$

If $A_i \gg A_c$, the last term in (3.91) is negligible with respect to the first term. It follows that the envelope detector output is approximated by

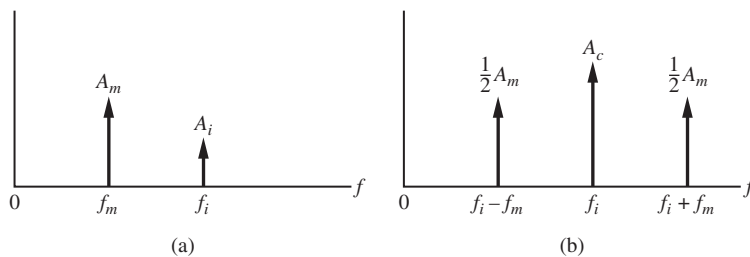
$$y_D(t) \cong A_c \cos(2\pi f_i t) + A_m \cos(2\pi f_m t) \cos(2\pi f_i t) \quad (3.92)$$

**Figure 3.23**

Phasor diagrams illustrating interference.
 (a) Phasor diagram without interference.
 (b) Phasor diagram with interference.

At this point, several observations are in order. In envelope detectors, the largest high-frequency component is treated as the carrier. If $A_c \gg A_i$, the effective demodulation carrier has a frequency f_c , whereas if $A_i \gg A_c$, the *effective* carrier frequency becomes the interference frequency $f_c + f_i$.

The spectra of the envelope detector output are illustrated in Figure 3.24 for $A_c \gg A_i$ and for $A_c \ll A_i$. For $A_c \gg A_i$ the interfering tone simply appears as a sinusoidal component, having frequency f_i at the output of the envelope detector. This illustrates that for $A_c \gg A_i$, the envelope detector performs as a linear demodulator. The situation is much different for $A_c \ll A_i$, as can be seen from (3.92) and Figure 3.24(b). For this case we see that the sinusoidal message signal, having frequency f_m , modulates the interference tone. The output of the envelope detector has a spectrum that reminds us of the spectrum of an AM signal with carrier frequency f_i and sideband components at $f_i + f_m$ and $f_i - f_m$. The message signal is effectively lost. This degradation of the desired signal is called the *threshold effect* and is a consequence of the nonlinear nature of the envelope detector. We shall study the threshold effect in detail in Chapter 8 when we investigate the effect of noise in analog systems.

**Figure 3.24**

Envelope detector output spectra. (a) $A_c \gg A_i$. (b) $A_c \ll A_i$.

3.7 PULSE AMPLITUDE MODULATION—PAM

In Chapter 2 (Section 2.8) we saw that continuous bandlimited signals can be represented by a sequence of discrete samples and that the continuous signal can be reconstructed with negligible error if the sampling rate is sufficiently high. Consideration of sampled signals leads us to the topic of pulse modulation. Pulse modulation can be either analog, in which some attribute of a pulse varies continuously in one-to-one correspondence with a sample value, or digital, in which some attribute of a pulse can take on a certain value from a set of allowable values. In this section we examine pulse amplitude modulation (PAM). In the following section we examine a couple of examples of digital pulse modulation.

As mentioned, *analog pulse modulation* results when some attribute of a pulse varies continuously in one-to-one correspondence with a sample value. Three attributes can be readily varied: amplitude, width, and position. These lead to pulse amplitude modulation (PAM), pulse-width modulation (PWM), and pulse-position modulation (PPM) as illustrated in Figure 3.25. Only PAM will be treated in this chapter. Pulse-width modulation and pulse-position modulation, which have the characteristics of angle modulation, are considered in the next chapter.

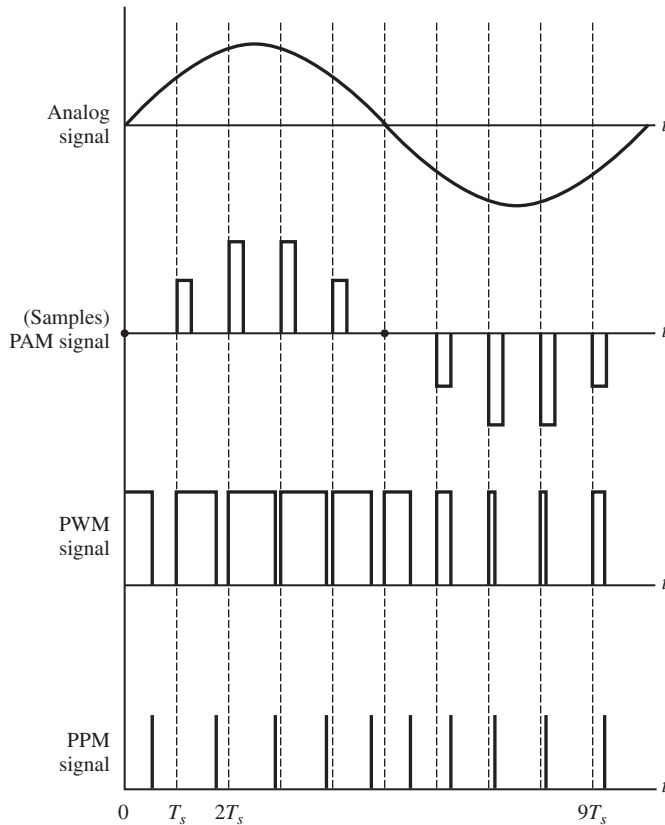


Figure 3.25
Illustration of PAM, PWM, and PPM.

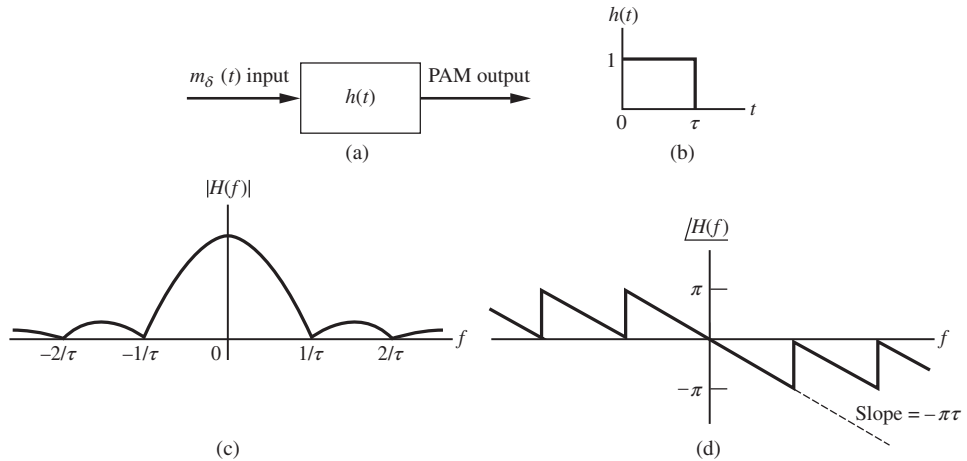


Figure 3.26 Generation of PAM. (a) Holding network. (b) Impulse response of holding network. (c) Amplitude response of holding network. (d) Phase response of holding network.

A PAM waveform consists of a sequence of flat-topped pulses designating sample values. The amplitude of each pulse corresponds to the value of the message signal $m(t)$ at the leading edge of the pulse. The essential difference between PAM and the sampling operation discussed in the previous chapter is that in PAM we allow the sampling pulse to have finite width. The finite-width pulse can be generated from the impulse-train sampling function by passing the impulse-train samples through a holding circuit as shown in Figure 3.26. The impulse response of the ideal holding circuit is given by

$$h(t) = \Pi\left(\frac{t - \frac{1}{2}\tau}{\tau}\right) \quad (3.93)$$

The holding circuit transforms the impulse function samples, given by

$$m_\delta(t) = m(nT_s)\delta(t - nT_s) \quad (3.94)$$

to the PAM waveform given by

$$m_c(t) = m(nT_s)\Pi\left[\frac{(t - nT_s) + \frac{1}{2}\tau}{\tau}\right] \quad (3.95)$$

as illustrated in Figure 3.26. The transfer function of the holding circuit is

$$H(f) = \tau \operatorname{sinc}(f\tau)e^{-j\pi f\tau} \quad (3.96)$$

Since the holding network does not have a constant amplitude response over the bandwidth of $m(t)$, amplitude distortion results. This amplitude distortion, which can be significant unless the pulse width τ is very small, can be removed by passing the samples, prior to reconstruction of $m(t)$, through a filter having an amplitude response equal to $1/|H(f)|$, over the bandwidth of $m(t)$. This process is referred to as *equalization* and will be treated in more detail later in this book. Since the phase response of the holding network is linear, the effect is a time delay and can usually be neglected.

3.8 DIGITAL PULSE MODULATION

We now briefly examine two types of digital pulse modulation: delta modulation and pulse-code modulation (PCM).

3.8.1 Delta Modulation

Delta modulation (DM) is a modulation technique in which the message signal is encoded into a sequence of binary symbols. These binary symbols are represented by the polarity of impulse functions at the modulator output. The electronic circuits to implement both the modulator and the demodulator are extremely simple. This simplicity makes DM an attractive technique for a number of applications.

A block diagram of a delta modulator is illustrated in Figure 3.27(a). The input to the pulse modulator portion of the circuit is

$$d(t) = m(t) - m_s(t) \quad (3.97)$$

where $m(t)$ is the message signal and $m_s(t)$ is a reference waveform. The signal $d(t)$ is hard-limited and multiplied by the pulse-generator output. This yields

$$x_c(t) = \Delta(t)\delta(t - nT_s) \quad (3.98)$$

where $\Delta(t)$ is a hard-limited version of $d(t)$. The preceding expression can be written as

$$x_c(t) = \Delta(nT_s)\delta(t - nT_s) \quad (3.99)$$

Thus, the output of the delta modulator is a series of impulses, each having positive or negative polarity depending on the sign of $d(t)$ at the sampling instants. In practical applications, the output of the pulse generator is not, of course, a sequence of impulse functions but rather a sequence of pulses that are narrow with respect to their periods. Impulse functions are assumed here because of the resulting mathematical simplicity. The reference signal $m_s(t)$ is generated by integrating $x_c(t)$. This yields at

$$m_s(t) = \Delta(nT_s) \int_0^t \delta(\alpha - nT_s) d\alpha \quad (3.100)$$

which is a staircase approximation of $m(t)$. The reference signal $m_s(t)$ is shown in Figure 3.27(b) for an assumed $m(t)$. The transmitted waveform $x_c(t)$ is illustrated in Figure 3.27(c).

Demodulation of DM is accomplished by integrating $x_c(t)$ to form the staircase approximation $m_s(t)$. This signal can then be lowpass filtered to suppress the discrete jumps in $m_s(t)$. Since a lowpass filter approximates an integrator, it is often possible to eliminate the integrator portion of the demodulator and to demodulate DM simply by lowpass filtering, as was done for PAM and PWM. A difficulty with DM is the problem of slope overload. Slope overload occurs when the message signal $m(t)$ has a slope greater than can be followed by the staircase approximation $m_s(t)$. This effect is illustrated in Figure 3.28(a), which shows a step change in $m(t)$ at time t_0 . Assuming that each pulse in $x_c(t)$ has weight δ_0 , the maximum slope that can be followed by $m_s(t)$ is δ_0/T_s , as shown. Figure 3.28(b) shows the resulting error signal due to a step change in $m(t)$ at t_0 . It can be seen that significant error exists for some time following the step change in $m(t)$. The duration of the error due to slope overload depends on the amplitude of the step, the impulse weights δ_0 , and the sampling period T_s .

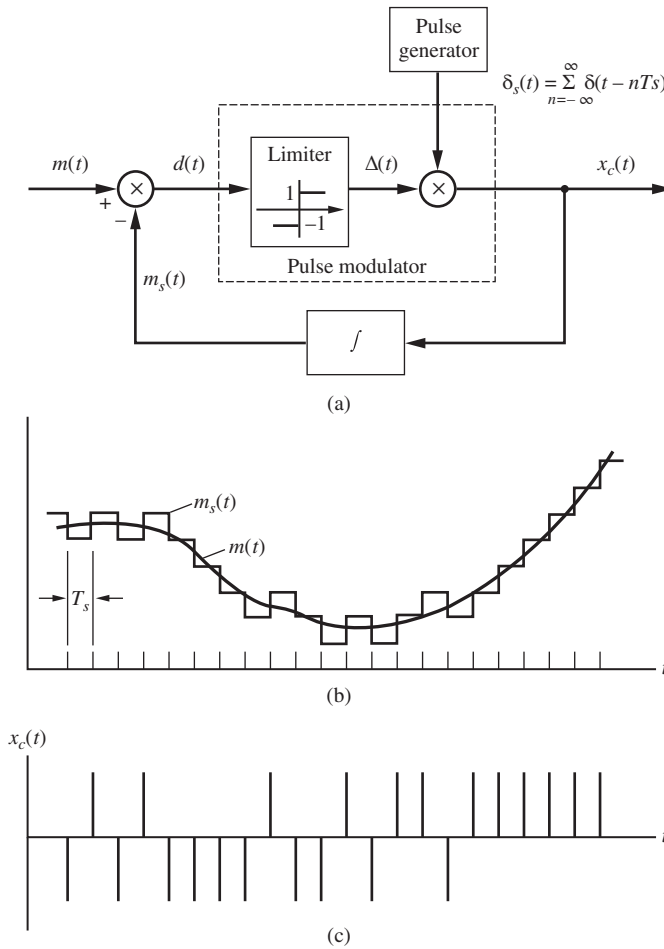


Figure 3.27 Delta modulation. (a) Delta modulator. (b) Modulation waveform and staircase approximation. (c) Modulator output.

A simple analysis can be carried out assuming that the message signal $m(t)$ is the sinusoidal signal

$$m(t) = A \sin(2\pi f_1 t) \tag{3.101}$$

The maximum slope that $m_s(t)$ can follow is

$$S_m = \frac{\delta_0}{T_s} \tag{3.102}$$

and the derivative of $m(t)$ is

$$\frac{d}{dt} m(t) = 2\pi A f_1 \cos(2\pi f_1 t) \tag{3.103}$$

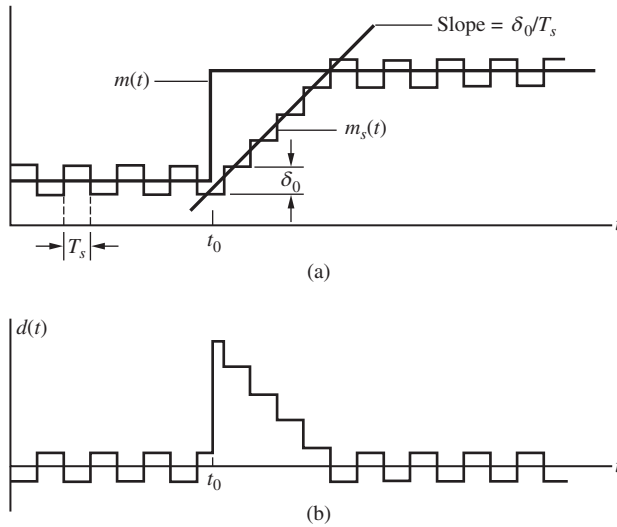
**Figure 3.28**

Illustration of slope overload. (a) Illustration of $m(t)$ and $m_s(t)$ for a step change in $m(t)$. (b) Error between $m(t)$ and $m_s(t)$ resulting from a step change in $m(t)$.

It follows that $m_s(t)$ can follow $m(t)$ without slope overload if

$$\frac{\delta_0}{T_s} \geq 2\pi A f_1 \quad (3.104)$$

3.8.2 Pulse-Code Modulation

The generation of PCM is a three-step process, as illustrated in Figure 3.29(a). The message signal $m(t)$ is first sampled, and the resulting sample values are then quantized. In PCM, the quantizing level of each sample is the transmitted quantity instead of the sample value. Typically, the quantization level is encoded into a binary sequence, as shown in Figure 3.29(b). The modulator output is a pulse representation of the binary sequence, which is shown in Figure 3.29(c). A binary “one” is represented as a pulse, and a binary “zero” is represented as the absence of a pulse. This absence of a pulse is indicated by a dashed line in Figure 3.29(c). The PCM waveform of Figure 3.29(c) shows that a PCM system requires synchronization so that the starting points of the digital words can be determined at the demodulator. PCM also requires a bandwidth sufficiently large to support transmission of the narrow pulses. Figure 3.29(c) is a highly idealized representation of a PCM signal. More practical signal representations, along with the bandwidth requirements for each, when we study line codes in Chapter 5.

To consider the bandwidth requirements of a PCM system, suppose that q quantization levels are used, satisfying

$$q = 2^n \quad (3.105)$$

where n , the word length, is an integer. For this case, $n = \log_2 q$ binary pulses must be transmitted for each sample of the message signal. If this signal has bandwidth W and the sampling rate is $2W$, then $2nW$ binary pulses must be transmitted per second. Thus, the

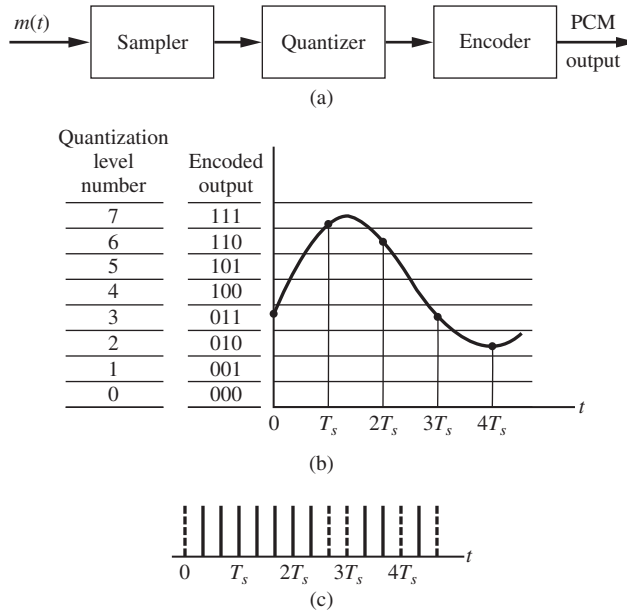


Figure 3.29 Generation of PCM. (a) PCM modulator. (b) Quantizer and coder. (c) Representation of coder output.

maximum width of each binary pulse is

$$(\Delta\tau)_{\max} = \frac{1}{2nW} \quad (3.106)$$

We saw in Chapter 2 that the bandwidth required for transmission of a pulse is inversely proportional to the pulse width, so that

$$B = 2knW \quad (3.107)$$

where B is the required bandwidth of the PCM system and k is a constant of proportionality. Note that we have assumed both a minimum sampling rate and a minimum value of bandwidth for transmitting a pulse. Equation (3.107) shows that the PCM signal bandwidth is proportional to the product of the message signal bandwidth W and the word length n .

If the major source of error in the system is quantizing error, it follows that a small error requirement dictates large word length resulting in large transmission bandwidth. Thus, in a PCM system, quantizing error can be exchanged for bandwidth. We shall see that this behavior is typical of many nonlinear systems operating in noisy environments. However, before noise effects can be analyzed, we must take a detour and develop the theory of probability and random processes. Knowledge of this area enables one to accurately model realistic and practical communication systems operating in everyday, nonidealized environments.

3.8.3 Time-Division Multiplexing

Time-division multiplexing (TDM) is best understood by considering Figure 3.30(a). The data sources are assumed to have been sampled at the Nyquist rate or higher. The commutator then

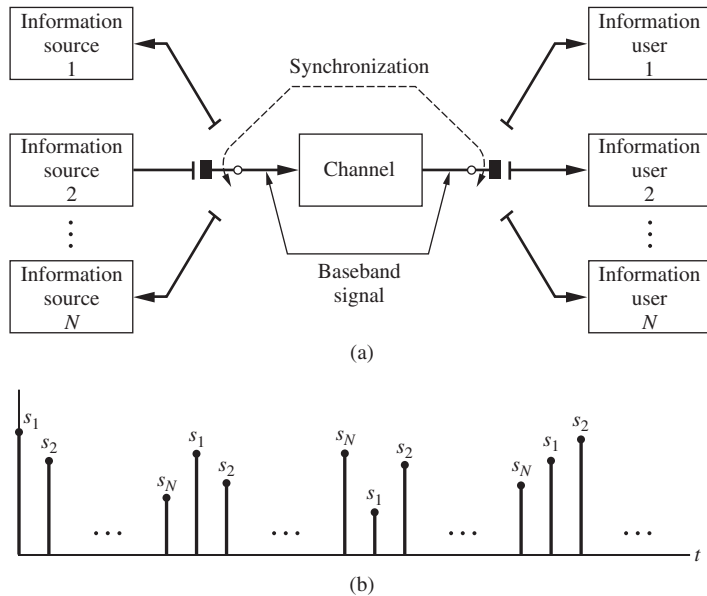


Figure 3.30
Time-division multiplexing. (a) TDM system. (b) Baseband signal.

interlaces the samples to form the baseband signal shown in Figure 3.30(b). At the channel output, the baseband signal is demultiplexed by using a second commutator as illustrated. Proper operation of this system obviously depends on proper synchronization between the two commutators.

If all message signals have equal bandwidth, then the samples are transmitted sequentially, as shown in Figure 3.30(b). If the sampled data signals have unequal bandwidths, more samples must be transmitted per unit time from the wideband channels. This is easily accomplished if the bandwidths are harmonically related. For example, assume that a TDM system has four channels of data. Also assume that the bandwidth of the first and second data sources, $s_1(t)$ and $s_2(t)$, is W Hz, the bandwidth of $s_3(t)$ is $2W$ Hz, and the bandwidth of $s_4(t)$ is $4W$ Hz. It is easy to show that a permissible sequence of baseband samples is a periodic sequence, one period of which is $\dots s_1 s_4 s_3 s_4 s_2 s_4 s_3 s_4 \dots$.

The minimum bandwidth of a TDM baseband is easy to determine using the sampling theorem. Assuming Nyquist rate sampling, the baseband contains $2W_i T$ samples from the i th channel in each T -s interval, where W is the bandwidth of the i th channel. Thus, the total number of baseband samples in a T -s interval is

$$n_s = \sum_{i=1}^N 2W_i T \quad (3.108)$$

Assuming that the baseband is a lowpass signal of bandwidth B , the required sampling rate is $2B$. In a T -s interval, we then have $2BT$ total samples. Thus,

$$n_s = 2BT = \sum_{i=1}^N 2W_i T \quad (3.109)$$

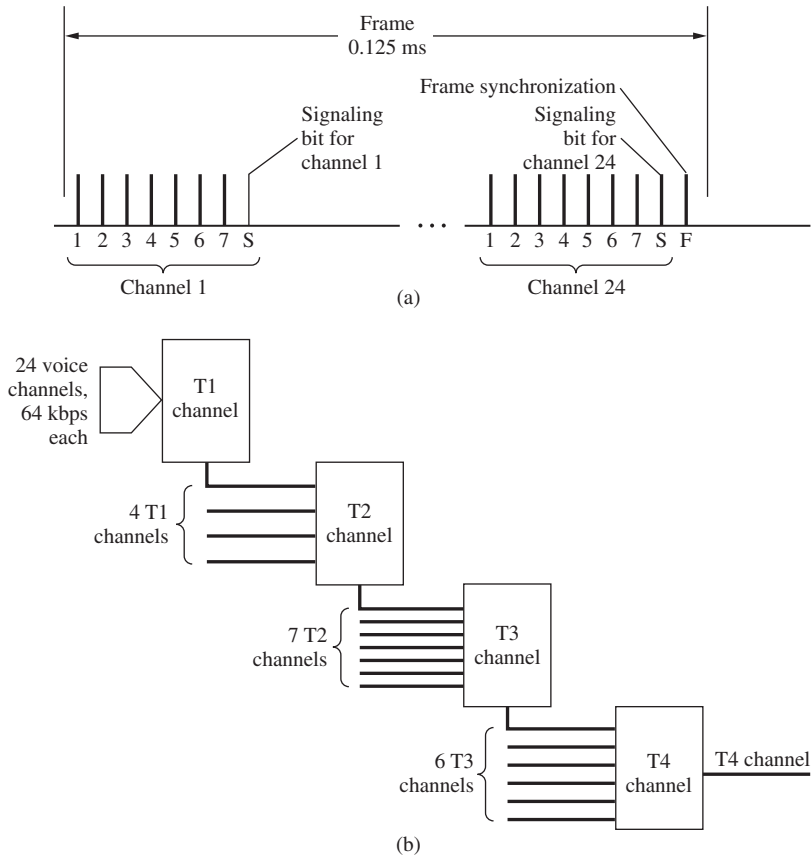


Figure 3.31 Digital multiplexing scheme for digital telephone. (a) T1 frame. (b) Digital multiplexing.

or

$$B = \sum_{i=1}^N W_i \quad (3.110)$$

which is the same as the minimum required bandwidth obtained for FDM.

3.8.4 An Example: The Digital Telephone System

As an example of a digital TDM system, we consider a multiplexing scheme common to many telephone systems. The sampling format is illustrated in Figure 3.31(a). A voice signal is sampled at 8000 samples per second, and each sample is quantized into seven binary digits. An additional binary digit, known as a *signaling bit*, is added to the basic seven bits that represent the sample value. The signaling bit is used in establishing calls and for synchronization. Thus, eight bits are transmitted for each sample value, yielding a bit rate of 64,000 bit/s (64 kbps). Twenty-four of these 64-kbps voice channels are grouped together to yield a T1 carrier. The T1 frame consists of $24(8) + 1 = 193$ bits. The extra bit is used for frame synchronization. The

frame duration is the reciprocal of the fundamental sampling frequency, or 0.125 ms. Since the frame rate is 8000 frames per second, with 193 bits per frame, the T1 data rate is 1.544 Mbps.

As shown in Figure 3.31(b), four T1 carriers can be multiplexed to yield a T2 carrier, which consists of 96 voice channels. Seven T2 carriers yield a T3 carrier, and six T3 carriers yield a T4 carrier. The bit rate of a T4 channel, consisting of 4032 voice channels with signaling bits and framing bits, is 274.176 Mbps. A T1 link is typically used for short transmission distances in areas of heavy usage. T4 and T5 channels are used for long transmission distances.

Further Reading

One can find basic treatments of modulation theory at about the same technical level of this text in a wide variety of books. Several selected examples are Carlson and Crilly (2009), Haykin and Moher (2009), Lathi and Ding (2009), and Couch (2013).

Summary

1. Modulation is the process by which a parameter of a carrier is varied in one-to-one correspondence with an information-bearing signal usually referred to as the *message*. Several uses of modulation are to achieve efficient transmission, to allocate channels, and for multiplexing.

2. If the carrier is continuous, the modulation is continuous-wave modulation. If the carrier is a sequence of pulses, the modulation is pulse modulation.

3. There are two basic types of continuous-wave modulation: linear modulation and angle modulation.

4. Assume that a general modulated carrier is given by

$$x_c(t) = A(t) \cos[2\pi f_c t + \phi(t)]$$

If $A(t)$ is proportional to the message signal, the result is linear modulation. If $\phi(t)$ is proportional to the message signal, the result is PM. If the time derivative of $\phi(t)$ is proportional to the message signal, the result is FM. Both PM and FM are examples of angle modulation. Angle modulation is a nonlinear process.

5. The simplest example of linear modulation is DSB. Double sideband is implemented as a simple product device, and coherent demodulation must be used, where coherent demodulation means that a local reference at the receiver that is of the same frequency and phase as the incoming carrier is used in demodulation.

6. An AM signal is formed by adding a carrier component to a DSB signal. This is a useful modulation technique because it allows simple envelope detection to be used for implementation very simple, and inexpensive, receivers.

7. The efficiency of a modulation process is defined as the percentage of total power that conveys information. For AM, this is given by

$$E = \frac{a^2 \langle m_n^2(t) \rangle}{1 + a^2 \langle m_n^2(t) \rangle} (100\%)$$

where the parameter a is the *modulation index* and $m_n(t)$ is $m(t)$ normalized so that the negative peak value is unity. If envelope demodulation is used, the index must be less than unity.

8. The modulation trapezoid provides a simple technique for monitoring the modulation index of an AM signal. It also provides a visual indication of the linearity of the modulator and transmitter.

9. An SSB signal is generated by transmitting only one of the sidebands in a DSB signal. Single-sideband signals are generated either by sideband filtering a DSB signal or by using a phase-shift modulator. Single-sideband signals can be written as

$$x_c(t) = \frac{1}{2} A_c m(t) \cos(2\pi f_c t) \pm \frac{1}{2} A_c \hat{m}(t) \sin(2\pi f_c t)$$

in which the plus sign is used for lower-sideband SSB and the minus sign is used for upper-sideband SSB. These signals can be demodulated either through the use of coherent demodulation or through the use of carrier reinsertion.

10. Vestigial sideband results when a vestige of one sideband appears on an otherwise SSB signal. Vestigial sideband is easier to generate than SSB. Either coherent demodulation or carrier reinsertion can be used for message recovery.

11. Frequency translation is accomplished by multiplying a signal by a carrier and filtering. These systems are known as mixers.

12. The concept of mixing has many applications including the implementation of superheterodyne receivers. Mixing results in *image frequencies*, which can be troublesome if not removed by filtering.

13. *Interference*, the presence of undesired signal components, can be a problem in demodulation. Interference at the input of a demodulator results in undesired components at the demodulator output. If the interference is large and if the demodulator is nonlinear, thresholding can occur. The result of this is a drastic loss of the signal component.

14. Pulse-amplitude modulation results when the amplitude of each carrier pulse is proportional to the value of the message signal at each sampling instant. Pulse-amplitude modulation is essentially a sample-and-hold operation. Demodulation of PAM is accomplished by lowpass filtering.

15. Digital pulse modulation results when the sample values of the message signal are quantized and encoded prior to transmission.

16. Delta modulation is an easily implemented form of digital pulse modulation. In DM, the message signal is encoded into a sequence of binary symbols. The binary symbols are represented by the polarity of impulse functions at the modulator output. Demodulation is ideally accomplished by integration, but lowpass filtering is often a simple and satisfactory substitute.

17. Pulse-code modulation results when the message signal is sampled and quantized, and each quantized sample value is encoded as a sequence of binary symbols. Pulse-code modulation differs from DM in that in PCM each quantized sample value is transmitted but in DM the transmitted quantity is the polarity of the change in the message signal from one sample to the next.

18. Multiplexing is a scheme allowing two or more message signals to be communicated simultaneously using a single system.

19. Frequency-division multiplexing results when simultaneous transmission is accomplished by translating message spectra, using modulation to *nonoverlapping* locations in a baseband spectrum. The baseband signal is then transmitted using any carrier modulation method.

20. Quadrature multiplexing results when two message signals are translated, using linear modulation with quadrature carriers, to the same spectral locations. Demodulation is accomplished coherently using quadrature demodulation carriers. A phase error in a demodulation carrier results in serious distortion of the demodulated signal. This distortion has two components: a time-varying attenuation of the desired output signal and crosstalk from the quadrature channel.

21. Time-division multiplexing results when samples from two or more data sources are interlaced, using commutation, to form a baseband signal. Demultiplexing is accomplished by using a second commutator, which must be synchronous with the multiplexing commutator.

Drill Problems

3.1 A DSB signal has the message signal

$$m(t) = 3 \cos(40\pi t) + 7 \sin(64\pi t)$$

The unmodulated carrier is given by

$$c(t) = 40 \cos(2000\pi t)$$

Determine the frequencies of the upper-sideband components, the frequencies of the lower-sideband components, and the total transmitted power.

3.2 Using the same message signal and unmodulated carrier as given in the previous problem, and assuming that the modulation technique is AM, determine the modulation index and the efficiency.

3.3 An AM system operates with $A_c = 100$ and $a = 0.8$. Sketch and fully dimension the modulation trapezoid.

3.4 Sketch and fully dimension the modulation trapezoid for AM with $a > 1$. Write the equation for determining the modulation index in terms of A and B .

3.5 Show that an AM signal can be demodulated using coherent demodulation by assuming a demodulation carrier of the form

$$2 \cos[2\pi f_c t + \theta(t)]$$

where $\theta(t)$ is the demodulation phase error.

3.6 A message signal is given by

$$m(t) = 3 \cos(40\pi t) + 7 \sin(64\pi t)$$

152 Chapter 3 • Linear Modulation Techniques

Also $A_c = 20$ V and $f_c = 300$ Hz. Determine the expression for the upper-sideband SSB signal and the lower-sideband SSB signal. Write these in a way that shows the amplitude and frequency of all transmitted components.

3.7 Equation (3.63) gives the amplitude and phase for the VSB signal components centered about $f = +f_c$. Give the amplitude and phase of the signal components centered about $f = -f_c$. Using these values show that the VSB signal is real.

3.8 An AM radio uses the standard IF frequency of 455 kHz and is tuned to receive a signal having a carrier frequency of 1020 kHz. Determine the frequency of the local oscillator for both low-side tuning and high-side tuning. Give the image frequencies for each.

3.9 The input to an AM receiver input consists of both modulated carrier (the message signal is a single tone) and interference terms. Assuming that $A_i = 100$ V, $A_m = 0.2$ V, $A_c = 1$ V, $f_m = 10$ Hz, $f_c = 300$ Hz, and $f_i = 320$ Hz, approximate the envelope

detector output by giving the amplitudes and frequencies of all components at the envelope detector output.

3.10 A PAM signal is formed by sampling an analog signal at 5 kHz. The duty cycle of the generated PAM pulses is to be 5%. Define the transfer function of the holding circuit by giving the value of τ in (3.92). Define the transfer function of the equalizing filter.

3.11 Rewrite (3.100) to show that relationship between δ_0/A and $T_s f_1$. A signal defined by

$$m(t) = A \cos(40\pi t)$$

is sampled at 1000 Hz to form a DM signal. Give the minimum value of δ_0/A to prevent slope overload.

3.12 A TDM signal consists of four signals having bandwidths of 1000, 2000, 4000, and 6000 Hz. What is the total bandwidth of the composite TDM signal. What is the lowest possible sampling frequency for the TDM signal?

Problems

Section 3.1

3.1 Assume that a DSB signal

$$x_c(t) = A_c m(t) \cos(2\pi f_c t + \phi_0)$$

is demodulated using the demodulation carrier $2 \cos[2\pi f_c t + \theta(t)]$. Determine, in general, the demodulated output $y_D(t)$. Let $A_c = 1$ and $\theta(t) = \theta_0$, where θ_0 is a constant, and determine the mean-square error between $m(t)$ and the demodulated output as a function of ϕ_0 and θ_0 . Now let $\theta_0 = 2\pi f_0 t$ and compute the mean-square error between $m(t)$ and the demodulated output.

3.2 A message signal is given by

$$m(t) = \sum_{k=1}^5 \frac{10}{k} \sin(2\pi k f_m t)$$

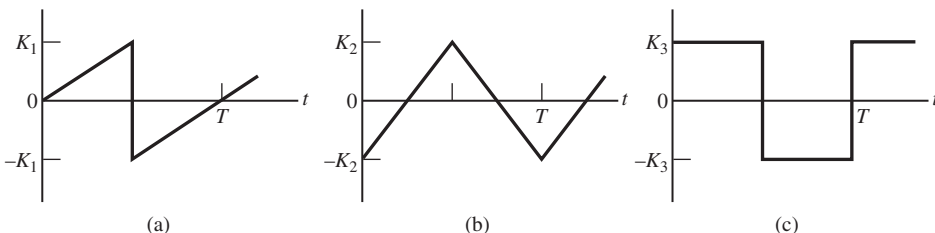


Figure 3.32

and the carrier is given by

$$c(t) = 100 \cos(200\pi t)$$

Write the transmitted signal as a Fourier series and determine the transmitted power.

Section 3.2

3.3 Design an envelope detector that uses a full-wave rectifier rather than the half-wave rectifier shown in Figure 3.3. Sketch the resulting waveforms, as was done in for a half-wave rectifier. What are the advantages of the full-wave rectifier?

3.4 Three message signals are periodic with period T , as shown in Figure 3.32. Each of the three message signals is applied to an AM modulator. For each message signal, determine the modulation efficiency for $a = 0.2$, $a = 0.3$, $a = 0.4$, $a = 0.7$, and $a = 1$.

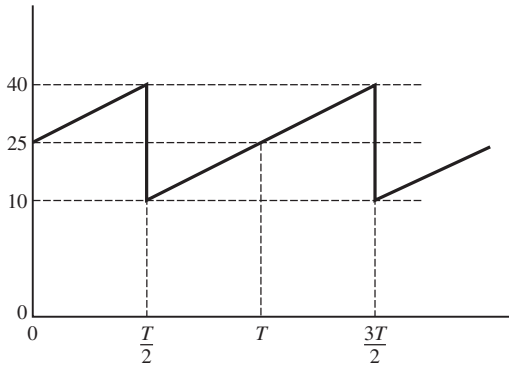


Figure 3.33

3.5 The positive portion of the envelope of the output of an AM modulator is shown in Figure 3.33. The message signal is a waveform having zero DC value. Determine the modulation index, the carrier power, the efficiency, and the power in the sidebands.

3.6 A message signal is a square wave with maximum and minimum values of 8 and -8 V, respectively. The modulation index $a = 0.7$ and the carrier amplitude $A_c = 100$ V. Determine the power in the sidebands and the efficiency. Sketch the modulation trapezoid.

3.7 In this problem we examine the efficiency of AM for the case in which the message signal does not have symmetrical maximum and minimum values. Two message signals are shown in Figure 3.34. Each is periodic with period T , and τ is chosen such that the DC value of $m(t)$ is zero. Calculate the efficiency for each $m(t)$ for $a = 0.7$ and $a = 1$.

3.8 An AM modulator operates with the message signal

$$m(t) = 9 \cos(20\pi t) - 8 \cos(60\pi t)$$

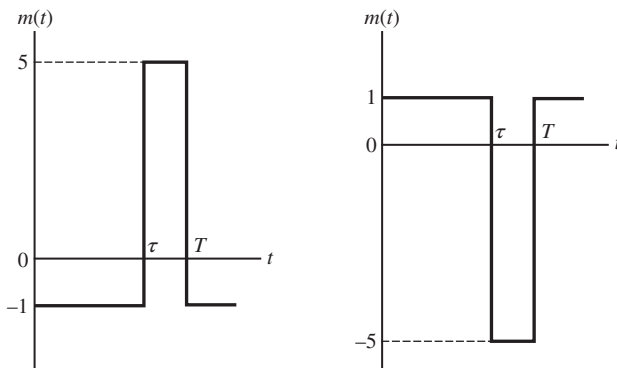


Figure 3.34

The unmodulated carrier is given by $110 \cos(200\pi t)$, and the system operates with an index of 0.8.

- (a) Write the equation for $m_n(t)$, the normalized signal with a minimum value of -1 .
- (b) Determine $\langle m_n^2(t) \rangle$, the power in $m_n(t)$.
- (c) Determine the efficiency of the modulator.
- (d) Sketch the double-sided spectrum of $x_c(t)$, the modulator output, giving the weights and frequencies of all components.

3.9 Rework Problem 3.8 for the message signal

$$m(t) = 9 \cos(20\pi t) + 8 \cos(60\pi t)$$

3.10 An AM modulator has output

$$x_c(t) = 40 \cos[2\pi(200)t] + 5 \cos[2\pi(180)t] + 5 \cos[2\pi(220)t]$$

Determine the modulation index and the efficiency.

3.11 An AM modulator has output

$$x_c(t) = A \cos[2\pi(200)t] + B \cos[2\pi(180)t] + B \cos[2\pi(220)t]$$

The carrier power is P_0 and the efficiency is E_{ff} . Derive an expression for E_{ff} in terms of P_0 , A , and B . Determine A , B , and the modulation index for $P_0 = 200$ W and $E_{ff} = 30\%$.

3.12 An AM modulator has output

$$x_c(t) = 25 \cos[2\pi(150)t] + 5 \cos[2\pi(160)t] + 5 \cos[2\pi(140)t]$$

Determine the modulation index and the efficiency.

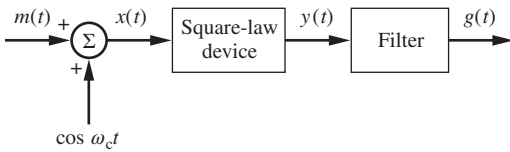


Figure 3.35

3.13 An AM modulator is operating with an index of 0.8. The modulating signal is

$$m(t) = 2 \cos(2\pi f_m t) + \cos(4\pi f_m t) + 2 \cos(10\pi f_m t)$$

- (a) Sketch the spectrum of the modulator output showing the weights of all impulse functions.
- (b) What is the efficiency of the modulation process?

3.14 Consider the system shown in Figure 3.35. Assume that the average value of $m(t)$ is zero and that the maximum value of $|m(t)|$ is M . Also assume that the square-law device is defined by $y(t) = 4x(t) + 2x^2(t)$.

- (a) Write the equation for $y(t)$.
- (b) Describe the filter that yields an AM signal for $g(t)$. Give the necessary filter type and the frequencies of interest.
- (c) What value of M yields a modulation index of 0.1?
- (d) What is an advantage of this method of modulation?

Section 3.3

3.15 Assume that a message signal is given by

$$m(t) = 4 \cos(2\pi f_m t) + \cos(4\pi f_m t)$$

Calculate an expression for

$$x_c(t) = \frac{1}{2} A_c m(t) \cos(2\pi f_c t) \pm \frac{1}{2} A_c \hat{m}(t) \sin(2\pi f_c t)$$

for $A_c = 10$. Show, by sketching the spectra, that the result is upper-sideband or lower-sideband SSB depending upon the choice of the algebraic sign.

3.16 Redraw Figure 3.10 to illustrate the generation of upper-sideband SSB. Give the equation defining the upper-sideband filter. Complete the analysis by deriving the expression for the output of an upper-sideband SSB modulator.

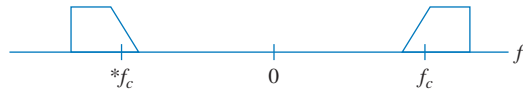


Figure 3.36

3.17 Squaring a DSB or AM signal generates a frequency component at twice the carrier frequency. Is this also true for SSB signals? Show that it is or is not.

Section 3.4

3.18 Prove analytically that carrier reinsertion with envelope detection can be used for demodulation of VSB.

3.19 Figure 3.36 shows the spectrum of a VSB signal. The amplitude and phase characteristics are the same as described in Example 3.3. Show that upon coherent demodulation, the output of the demodulator is real.

Section 3.5

3.20 Sketch Figure 3.20 for the case where $f_{LO} = f_c - f_{IF}$.

3.21 A mixer is used in a short-wave superheterodyne receiver. The receiver is designed to receive transmitted signals between 10 and 30 MHz. High-side tuning is to be used. Determine an acceptable IF frequency and the tuning range of the local oscillator. Strive to generate a design that yields the minimum tuning range.

3.22 A superheterodyne receiver uses an IF frequency of 455 kHz. The receiver is tuned to a transmitter having a carrier frequency of 1100 kHz. Give two permissible frequencies of the local oscillator and the image frequency for each. Repeat assuming that the IF frequency is 2500 kHz.

Section 3.6

3.23 A DSB signal is squared to generate a carrier component that may then be used for demodulation. (A technique for doing this, namely the phase-locked loop, will be studied in the next chapter.) Derive an expression that illustrates the impact of interference on this technique.

Section 3.7

3.24 A continuous-time signal is sampled and input to a holding circuit. The product of the holding time and the sampling frequency is τf_s . Plot the amplitude response of the required equalizer as a function of τf_s . What problem, or problems, arise if a large value of τ is used while the sampling frequency is held constant?

Section 3.8

3.25 A continuous data signal is quantized and transmitted using a PCM system. If each data sample at the receiving end of the system must be known to within $\pm 0.25\%$ of the peak-to-peak full-scale value, how many binary symbols must each transmitted digital word contain? Assume that the message signal is speech and has a bandwidth of 4 kHz. Estimate the bandwidth of the resulting PCM signal (choose k).

3.26 A delta modulator has the message signal

$$m(t) = 3 \sin 2\pi(10)t + 4 \sin 2\pi(20)t$$

Determine the minimum sampling frequency required to prevent slope overload, assuming that the impulse weights δ_0 are 0.05π .

3.27 Five messages bandlimited to $W, W, 2W, 4W$, and $4W$ Hz, respectively, are to be time-division multiplexed. Devise a commutator configuration such that each signal is periodically sampled at its own minimum rate and the samples are properly interlaced. What is the minimum transmission bandwidth required for this TDM signal?

3.28 Repeat the preceding problem assuming that the commutator is run at twice the minimum rate. What are the advantages and disadvantages of doing this?

3.29 Five messages bandlimited to $W, W, 2W, 5W$, and $7W$ Hz, respectively, are to be time-division multiplexed. Devise a sampling scheme requiring the minimum sampling frequency.

3.30 In an FDM communication system, the transmitted baseband signal is

$$x(t) = m_1(t) \cos(2\pi f_1 t) + m_2(t) \cos(2\pi f_2 t)$$

This system has a second-order nonlinearity between transmitter output and receiver input. Thus, the received baseband signal $y(t)$ can be expressed as

$$y(t) = a_1 x(t) + a_2 x^2(t)$$

Assuming that the two message signals, $m_1(t)$ and $m_2(t)$, have the spectra

$$M_1(f) = M_2(f) = \Pi\left(\frac{f}{W}\right)$$

sketch the spectrum of $y(t)$. Discuss the difficulties encountered in demodulating the received baseband signal. In many FDM systems, the subcarrier frequencies f_1 and f_2 are harmonically related. Describe any additional problems this presents.

Computer Exercises

3.1 In Example 3.1 we determined the minimum value of $m(t)$ using MATLAB. Write a MATLAB program that provides a complete solution for Example 3.1. Use the FFT for finding the amplitude and phase spectra of the transmitted signal $x_c(t)$.

3.2 The purpose of this exercise is to demonstrate the properties of SSB modulation. Develop a computer program to generate both upper-sideband and lower-sideband SSB signals and display both the time-domain signals and the amplitude spectra of these signals. Assume the message signal

$$m(t) = 2 \cos(2\pi f_m t) + \cos(4\pi f_m t)$$

Select both f_m and f_c so that both the time and frequency axes can be easily calibrated. Plot the envelope of the SSB signals, and show that both the upper-sideband and the lower-sideband SSB signals have the same envelope. Use the FFT algorithm to generate the amplitude spectrum for both the upper-sideband and the lower-sideband SSB signal.

3.3 Using the same message signal and value for f_m used in the preceding computer exercise, show that carrier rein-

sertion can be used to demodulate an SSB signal. Illustrate the effect of using a demodulation carrier with insufficient amplitude when using the carrier reinsertion technique.

3.4 In this computer exercise we investigate the properties of VSB modulation. Develop a computer program (using MATLAB) to generate and plot a VSB signal and the corresponding amplitude spectrum. Using the program, show that VSB can be demodulated using carrier reinsertion.

3.5 Using MATLAB simulate delta modulation. Generate a signal, using a sum of sinusoids, so that the bandwidth is known. Sample at an appropriate sampling frequency (no slope overload). Show the staircase approximation. Now reduce the sampling frequency so that slope overload occurs. Once again, show the staircase approximation.

3.6 Using a sum of sinusoids as the sampling frequency, sample and generate a PAM signal. Experiment with various values of τf_s . Show that the message signal is recovered by lowpass filtering. A third-order Butterworth filter is suggested.

CHAPTER 4

ANGLE MODULATION AND
MULTIPLEXING

In the previous chapter, we considered analog linear modulation. We now consider angle modulation. To generate angle modulation, the amplitude of the modulated carrier is held constant and either the phase or the time derivative of the phase of the carrier is varied linearly with the message signal $m(t)$. These lead to phase modulation (PM) or frequency modulation (FM), respectively.

The most efficient technique for demodulating angle modulated signals is the phase-locked loop (PLL). The PLL is ubiquitous in modern communication systems. Both analog systems and, as we will see later, digital systems make extensive use of PLLs. Because of the importance of the PLL, we give considerable emphasis to it in this chapter.

Also in this chapter we consider pulse modulation techniques related to angle modulation, PWM and PPM. The motivation for doing this is, with the exception of pulse-amplitude modulation, many of the characteristics of pulse modulation are similar to the characteristics of angle modulation.

4.1 PHASE AND FREQUENCY MODULATION DEFINED

Our starting point is the general signal model first used in the previous chapter, which is

$$x_c(t) = A_c \cos[2\pi f_c t + \phi(t)] \quad (4.1)$$

For angle modulation, the amplitude $A(t)$ is held constant at A_c and the message signal is communicated by the phase. The instantaneous phase of $x_c(t)$ is defined as

$$\theta_i(t) = 2\pi f_c t + \phi(t) \quad (4.2)$$

and the instantaneous frequency, in hertz, is defined as

$$f_i(t) = \frac{1}{2\pi} \frac{d\theta_i}{dt} = f_c + \frac{1}{2\pi} \frac{d\phi}{dt} \quad (4.3)$$

The functions $\phi(t)$ and $d\phi/dt$ are known as the *phase deviation* and *frequency deviation* (in radians per second), respectively.

Phase modulation implies that the phase deviation of the carrier is proportional to the message signal. Thus, for phase modulation,

$$\phi(t) = k_p m(t) \quad (4.4)$$

where k_p is the *deviation constant* in radians per unit of $m(t)$. Similarly, FM implies that the frequency deviation of the carrier is proportional to the modulating signal. This yields

$$\frac{d\phi}{dt} = k_f m(t) \quad (4.5)$$

The phase deviation of a frequency-modulated carrier is given by

$$\phi(t) = k_f \int_{t_0}^t m(\alpha) d\alpha + \phi_0 \quad (4.6)$$

in which ϕ_0 is the phase deviation at $t = t_0$. It follows from (4.5) that k_f is the frequency-deviation constant, expressed in radians per second per unit of $m(t)$. Since it is often more convenient to measure frequency deviation in Hz, we define

$$k_f = 2\pi f_d \quad (4.7)$$

where f_d is known as the *frequency-deviation constant* of the modulator and is expressed in Hz per unit of $m(t)$.

With these definitions, the phase modulator output is

$$x_c(t) = A_c \cos[2\pi f_c t + k_p m(t)] \quad (4.8)$$

and the frequency modulator output is

$$x_c(t) = A_c \cos \left[2\pi f_c t + 2\pi f_d \int_{t_0}^t m(\alpha) d\alpha \right] \quad (4.9)$$

The lower limit of the integral is typically not specified, since to do so would require the inclusion of an initial condition as shown in (4.6).

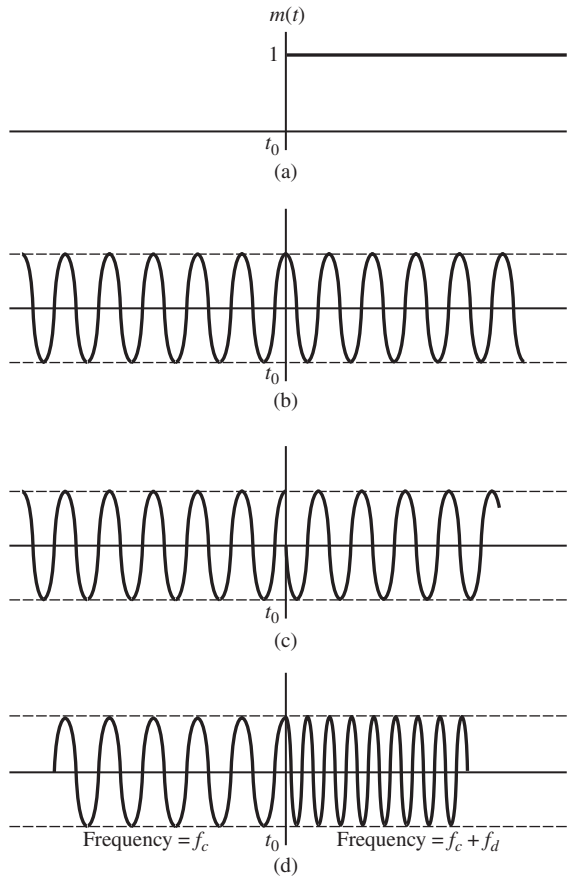
Figures 4.1 and 4.2 illustrate the outputs of PM and FM modulators. With a unit-step message signal, the instantaneous frequency of the PM modulator output is f_c for both $t < t_0$ and $t > t_0$. The phase of the unmodulated carrier is advanced by $k_p = \pi/2$ radians for $t > t_0$ giving rise to a signal that is discontinuous at $t = t_0$. The frequency of the output of the FM modulator is f_c for $t < t_0$, and the frequency is $f_c + f_d$ for $t > t_0$. The modulator output phase is, however, continuous at $t = t_0$.

With a sinusoidal message signal, the phase deviation of the PM modulator output is proportional to $m(t)$. The frequency deviation is proportional to the derivative of the phase deviation. Thus, the instantaneous frequency of the output of the PM modulator is maximum when the *slope* of $m(t)$ is maximum and minimum when the *slope* of $m(t)$ is minimum. The frequency deviation of the FM modulator output is proportional to $m(t)$. Thus, the instantaneous frequency of the FM modulator output is maximum when $m(t)$ is maximum and minimum when $m(t)$ is minimum. It should be noted that if $m(t)$ were not shown along with the modulator outputs, it would not be possible to distinguish the PM and FM modulator outputs. In the following sections we will devote considerable attention to the case in which $m(t)$ is sinusoidal.

4.1.1 Narrowband Angle Modulation

We start with a discussion of narrowband angle modulation because of the close relationship of narrowband angle modulation to AM, which we studied in the preceding chapter. To begin, we write an angle-modulated carrier in exponential form by writing (4.1) as

$$x_c(t) = \operatorname{Re}(A_c e^{j\phi(t)} e^{j2\pi f_c t}) \quad (4.10)$$

**Figure 4.1**

Comparison of PM and FM modulator outputs for a unit-step input.

(a) Message signal. (b) Unmodulated carrier. (c) Phase modulator output ($k_p = \frac{1}{2}\pi$). (d) Frequency modulator output.

where $\text{Re}(\cdot)$ implies that the real part of the argument is to be taken. Expanding $e^{j\phi(t)}$ in a power series yields

$$x_c(t) = \text{Re} \left\{ A_c \left[1 + j\phi(t) - \frac{\phi^2(t)}{2!} - \dots \right] e^{j2\pi f_c t} \right\} \quad (4.11)$$

If the peak phase deviation is small, so that the maximum value of $|\phi(t)|$ is much less than unity, the modulated carrier can be approximated as

$$x_c(t) \cong \text{Re}[A_c e^{j2\pi f_c t} + A_c \phi(t) j e^{j2\pi f_c t}]$$

Taking the real part yields

$$x_c(t) \cong A_c \cos(2\pi f_c t) - A_c \phi(t) \sin(2\pi f_c t) \quad (4.12)$$

The form of (4.12) is reminiscent of AM. The modulator output contains a carrier component and a term in which a function of $m(t)$ multiplies a 90° phase-shifted carrier. The first term yields a carrier component. The second term generates a pair of sidebands. Thus, if $\phi(t)$ has a bandwidth W , the bandwidth of a narrowband angle modulator output is $2W$. The important difference between AM and angle modulation is that the sidebands are produced by multiplication of the message-bearing signal, $\phi(t)$, with a carrier that is in phase

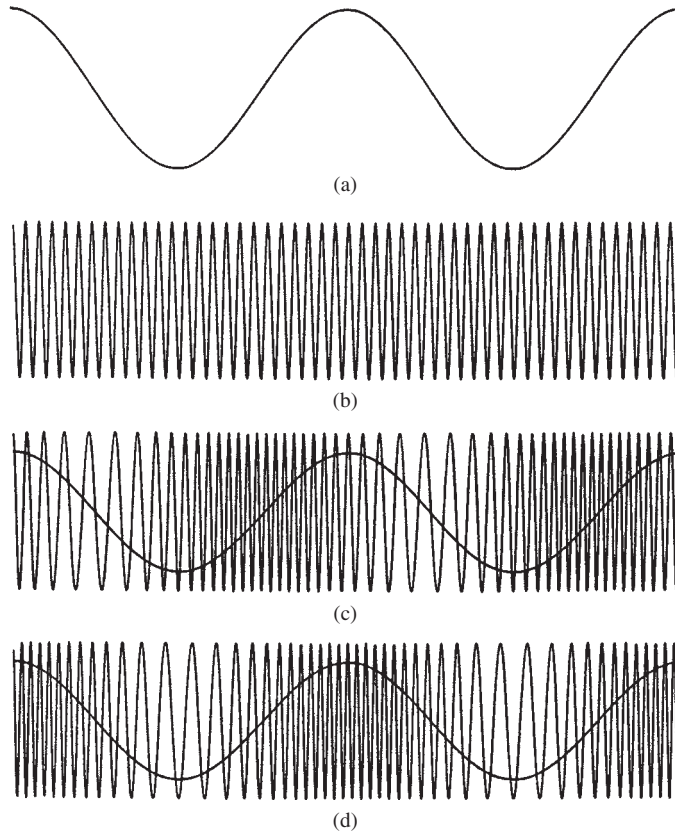


Figure 4.2

Angle modulation with sinusoidal message signal. (a) Message signal. (b) Unmodulated carrier. (c) Output of phase modulator with $m(t)$. (d) Output of frequency modulator with $m(t)$.

quadrature with the carrier component, whereas for AM they are not. This will be illustrated in Example 4.1.

The generation of narrowband angle modulation is easily accomplished using the method shown in Figure 4.3. The switch allows for the generation of either narrowband FM or narrow-

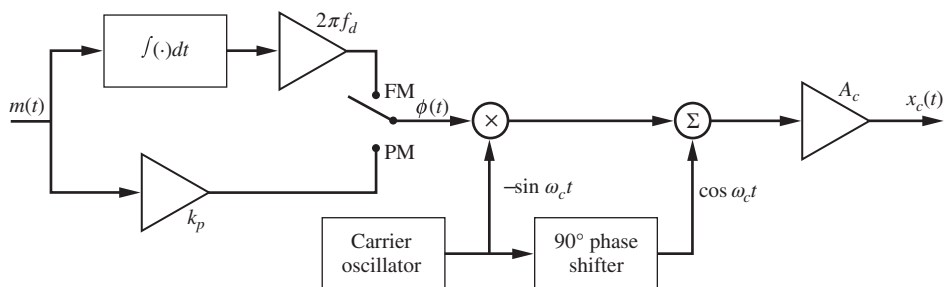


Figure 4.3

Generation of narrowband angle modulation.

band PM. We will show later that narrowband angle modulation is useful for the generation of angle-modulated signals that are not necessarily narrowband. This is accomplished through a process called *narrowband-to-wideband conversion*.

EXAMPLE 4.1

Consider an FM system with message signal

$$m(t) = A \cos(2\pi f_m t) \quad (4.13)$$

From (4.6), with t_0 and $\phi(t_0)$ equal to zero,

$$\phi(t) = k_f \int_0^t A \cos(2\pi f_m \alpha) d\alpha = \frac{Ak_f}{2\pi f_m} \sin(2\pi f_m t) = \frac{Af_d}{f_m} \sin(2\pi f_m t) \quad (4.14)$$

so that

$$x_c(t) = A_c \cos \left[2\pi f_c t + \frac{Af_d}{f_m} \sin(2\pi f_m t) \right] \quad (4.15)$$

If $Af_d/f_m \ll 1$, the modulator output can be approximated as

$$x_c(t) = A_c \left[\cos(2\pi f_c t) - \frac{Af_d}{f_m} \sin(2\pi f_c t) \sin(2\pi f_m t) \right] \quad (4.16)$$

which is

$$x_c(t) = A_c \cos(2\pi f_c t) + \frac{A_c Af_d}{2 f_m} \{ \cos[2\pi(f_c + f_m)t] - \cos[2\pi(f_c - f_m)t] \} \quad (4.17)$$

Thus, $x_c(t)$ can be written as

$$x_c(t) = A_c \operatorname{Re} \left\{ \left[1 + \frac{Af_d}{2f_m} (e^{j2\pi f_m t} - e^{-j2\pi f_m t}) \right] e^{j2\pi f_c t} \right\} \quad (4.18)$$

It is interesting to compare this result with the equivalent result for an AM signal. Since sinusoidal modulation is assumed, the AM signal can be written as

$$x_c(t) = A_c [1 + a \cos(2\pi f_m t)] \cos(2\pi f_c t) \quad (4.19)$$

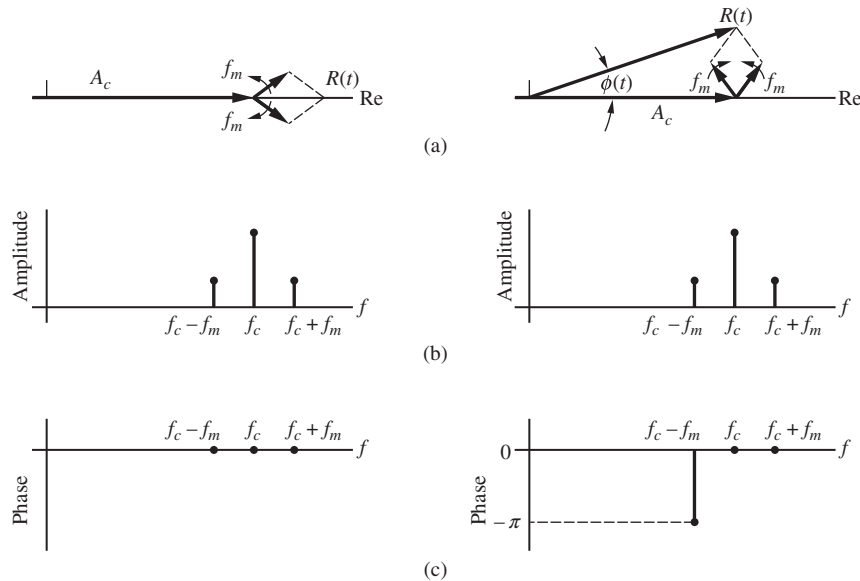
where $a = Af_d/f_m$ is the modulation index. Combining the two cosine terms yields

$$x_c(t) = A_c \cos(2\pi f_c t) + \frac{A_c a}{2} [\cos 2\pi(f_c + f_m)t + \cos 2\pi(f_c - f_m)t] \quad (4.20)$$

This can be written in exponential form as

$$x_c(t) = A_c \operatorname{Re} \left\{ \left[1 + \frac{a}{2} (e^{j2\pi f_m t} + e^{-j2\pi f_m t}) \right] e^{j2\pi f_c t} \right\} \quad (4.21)$$

Comparing (4.18) and (4.21) illustrates the similarity between the two signals. The first, and most important, difference is the sign of the term at frequency $f_c - f_m$, which represents the lower sideband. The other difference is that the index a in the AM signal is replaced by Af_d/f_m in the narrowband FM signal. We will see in the following section that Af_d/f_m determines the modulation index for an FM signal. Thus, these two parameters are in a sense equivalent since each defines the modulation index.

**Figure 4.4**

Comparison of AM and narrowband angle modulation. (a) Phasor diagrams. (b) Single-sided amplitude spectra. (c) Single-sided phase spectra.

Additional insight is gained by sketching the phasor diagrams and the amplitude and phase spectra for both signals. These are given in Figure 4.4. The phasor diagrams are drawn using the carrier phase as a reference. The difference between AM and narrowband angle modulation with a sinusoidal message signal lies in the fact that the phasor resulting from the LSB and USB phasors adds to the carrier for AM but is in phase quadrature with the carrier for angle modulation. This difference results from the minus sign in the LSB component and is also clearly seen in the phase spectra of the two signals. The amplitude spectra are equivalent.

4.1.2 Spectrum of an Angle-Modulated Signal

The derivation of the spectrum of an angle-modulated signal is typically a very difficult task. However, if the message signal is sinusoidal, the instantaneous phase deviation of the modulated carrier is sinusoidal for both FM and PM, and the spectrum can be obtained with ease. This is the case we will consider. Even though we are restricting our attention to a very special case, the results provide much insight into the frequency-domain behavior of angle modulation. In order to compute the spectrum of an angle-modulated signal with a sinusoidal message signal, we assume that

$$\phi(t) = \beta \sin(2\pi f_m t) \quad (4.22)$$

The parameter β is known as the *modulation index* and is the maximum phase deviation for both FM and PM. The signal

$$x_c(t) = A_c \cos[2\pi f_c t + \beta \sin(2\pi f_m t)] \quad (4.23)$$

can be expressed as

$$x_c(t) = \operatorname{Re} \left[A_c e^{j\beta \sin(2\pi f_m t)} e^{j2\pi f_c t} \right] \quad (4.24)$$

This expression has the form

$$x_c(t) = \operatorname{Re} [\tilde{x}_c(t) e^{j2\pi f_c t}] \quad (4.25)$$

where

$$\tilde{x}_c(t) = A_c e^{j\beta \sin(2\pi f_m t)} \quad (4.26)$$

is the complex envelope of the modulated carrier signal. The complex envelope is periodic with frequency f_m and can therefore be expanded in a Fourier series. The Fourier coefficients are given by

$$f_m \int_{-1/2f_m}^{1/2f_m} e^{j\beta \sin(2\pi f_m t)} e^{-j2\pi n f_m t} dt = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-[jn\pi - \beta \sin(x)]} dx \quad (4.27)$$

This integral cannot be evaluated in closed form. However, this integral arises in a variety of studies and, therefore, has been well tabulated. The integral is a function of n and β and is known as the *Bessel function* of the first kind of order n and argument β . It is denoted $J_n(\beta)$ and is tabulated for several values of n and β in Table 4.1. The significance of the underlining of various values in the table will be explained later.

With the aid of Bessel functions, we have

$$e^{j\beta \sin(2\pi f_m t)} = \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} \quad (4.28)$$

which allows the modulated carrier to be written as

$$x_c(t) = \operatorname{Re} \left[\left(A_c \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} \right) e^{j2\pi f_c t} \right] \quad (4.29)$$

Taking the real part yields

$$x_c(t) = A_c \sum_{n=-\infty}^{\infty} J_n(\beta) \cos[2\pi(f_c + n f_m)t] \quad (4.30)$$

from which the spectrum of $x_c(t)$ can be determined by inspection. The spectrum has components at the carrier frequency and has an infinite number of sidebands separated from the carrier frequency by integer multiples of the modulation frequency f_m . The amplitude of each spectral component can be determined from a table of values of the Bessel function. Such tables typically give $J_n(\beta)$ only for positive values of n . However, from the definition of $J_n(\beta)$ it can be determined that

$$J_{-n}(\beta) = J_n(\beta), \quad n \text{ even} \quad (4.31)$$

and

$$J_{-n}(\beta) = -J_n(\beta), \quad n \text{ odd} \quad (4.32)$$

These relationships allow us to plot the spectrum of (4.30), which is shown in Figure 4.5. The single-sided spectrum is shown for convenience.

A useful relationship between values of $J_n(\beta)$ for various values of n is the recursion formula

$$J_{n+1}(\beta) = \frac{2n}{\beta} J_n(\beta) + J_{n-1}(\beta) \quad (4.33)$$

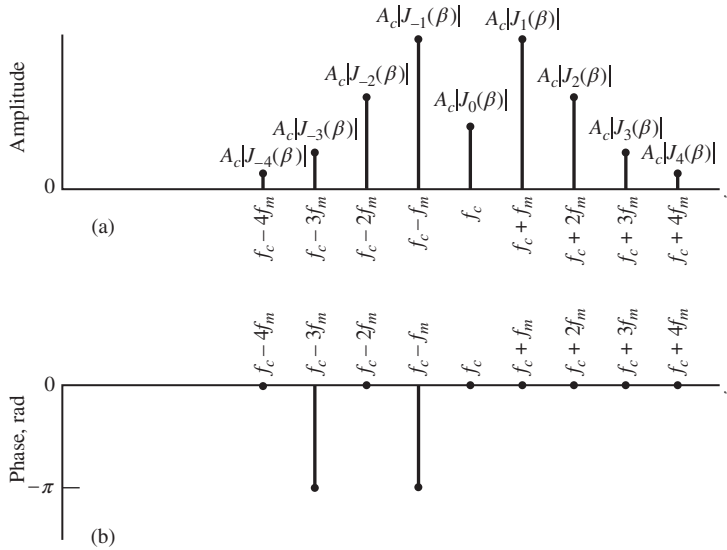


Figure 4.5
Spectra of an angle-modulated signal. (a) Single-sided amplitude spectrum. (b) Single-sided phase spectrum.

Thus, $J_{n+1}(\beta)$ can be determined from knowledge of $J_n(\beta)$ and $J_{n-1}(\beta)$. This enables us to compute a table of values of the Bessel function, as shown in Table 4.1, for any value of n from $J_0(\beta)$ and $J_1(\beta)$.

Figure 4.6 illustrates the behavior of the Fourier–Bessel coefficients $J_n(\beta)$, for $n = 0, 1, 2, 4,$ and 6 with $0 \leq \beta \leq 9$. Several interesting observations can be made. First, for $\beta \ll 1$,

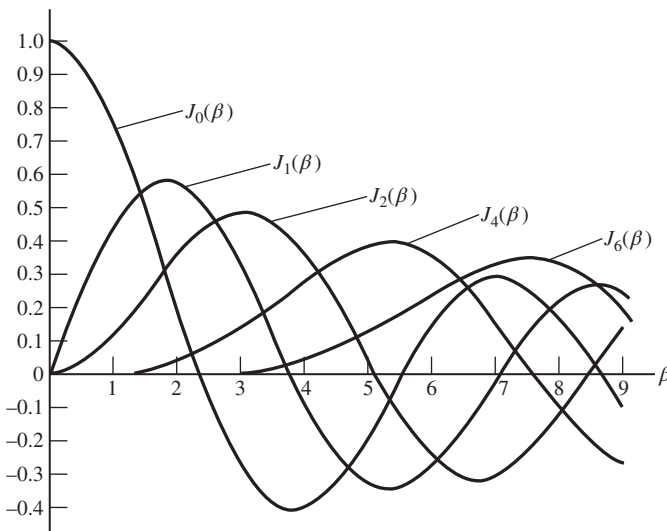


Figure 4.6
 $J_n(\beta)$ as a function of β .

Table 4.2 Values of β for which $J_n(\beta) = 0$ for $0 \leq \beta \leq 9$

n		β_{n0}	β_{n1}	β_{n2}
0	$J_0(\beta) = 0$	2.4048	5.5201	8.6537
1	$J_1(\beta) = 0$	0.0000	3.8317	7.0156
2	$J_2(\beta) = 0$	0.0000	5.1356	8.4172
4	$J_4(\beta) = 0$	0.0000	7.5883	–
6	$J_6(\beta) = 0$	0.0000	–	–

it is clear that $J_0(\beta)$ predominates, giving rise to narrowband angle modulation. It also can be seen that $J_n(\beta)$ oscillates for increasing β but that the amplitude of oscillation decreases with increasing β . Also of interest is the fact that the maximum value of $J_n(\beta)$ decreases with increasing n .

As Figure 4.6 shows, $J_n(\beta)$ is equal to zero at several values of β . Denoting these values of β by β_{nk} , where $k = 0, 1, 2$, we have the results in Table 4.2. As an example, $J_0(\beta)$ is zero for β equal to 2.4048, 5.5201, and 8.6537. Of course, there are an infinite number of points at which $J_n(\beta)$ is zero for any n , but consistent with Figure 4.6, only the values in the range $0 \leq \beta \leq 9$ are shown in Table 4.2. It follows that since $J_0(\beta)$ is zero at β equal to 2.4048, 5.5201, and 8.6537, the spectrum of the modulator output will not contain a component at the carrier frequency for these values of the modulation index. These points are referred to as *carrier nulls*. In a similar manner, the components at $f = f_c \pm f_m$ are zero if $J_1(\beta)$ is zero. The values of the modulation index giving rise to this condition are 0, 3.8317, and 7.0156. It should be obvious why only $J_0(\beta)$ is nonzero at $\beta = 0$. If the modulation index is zero, then either $m(t)$ is zero or the deviation constant f_d is zero. In either case, the modulator output is the unmodulated carrier, which has frequency components only at the carrier frequency. In computing the spectrum of the modulator output, our starting point was the assumption that

$$\phi(t) = \beta \sin(2\pi f_m t) \quad (4.34)$$

Note that in deriving the spectrum of the angle-modulated signal defined by (4.30), the modulator type (FM or PM) was not specified. The assumed $\phi(t)$, defined by (4.34), could represent either the phase deviation of a PM modulator with $m(t) = A \sin(\omega_m t)$ and an index $\beta = k_p A$, or an FM modulator with $m(t) = A \cos(2\pi f_m t)$ with index

$$\beta = \frac{f_d A}{f_m} \quad (4.35)$$

Equation (4.35) shows that the modulation index for FM is a function of the modulation frequency. This is not the case for PM. The behavior of the spectrum of an FM signal is illustrated in Figure 4.7, as f_m is decreased while holding $A f_d$ constant. For large values of f_m , the signal is narrowband FM, since only two sidebands are significant. For small values of f_m , many sidebands have significant value. Figure 4.7 is derived in the following computer example.

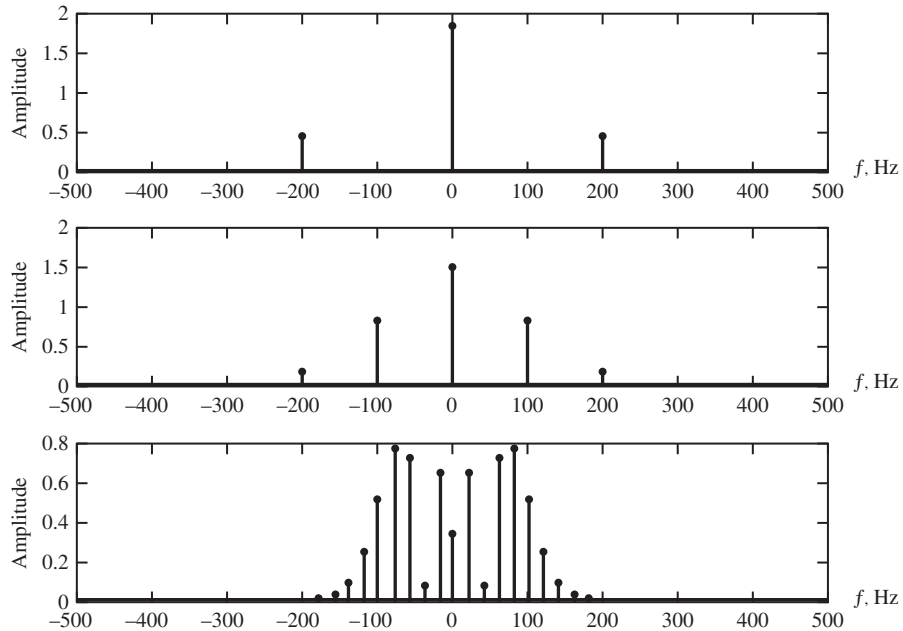


Figure 4.7

Amplitude spectrum of a complex envelope signal for increasing β and decreasing f_m .

COMPUTER EXAMPLE 4.1

In this computer example we determine the spectrum of the complex envelope signal given by (4.26). In the next computer example we will determine and plot the two-sided spectrum, which is determined from the complex envelope by writing the real bandpass signal as

$$x_c(t) = \frac{1}{2}\tilde{x}(t)e^{j2\pi f_c t} + \frac{1}{2}\tilde{x}^*(t)e^{-j2\pi f_c t} \quad (4.36)$$

Note once more that knowledge of the complex envelope signal and the carrier frequency fully determine the bandpass signal. In this example the spectrum of the complex envelope signal is determined for three different values of the modulation index. The MATLAB program, which uses the FFT for determination of the spectrum, follows.

```
%file c4ce1.m
fs=1000;
delt=1/fs;
t=0:delt:1-delt;
npts=length(t);
fm=[200 100 20];
fd=100;
for k=1:3
    beta=fd/fm(k);
    cxce=exp(i*beta*sin(2*pi*fm(k)*t));
    as=(1/npts)*abs(fft(cxce));
    evenf=[as(fs/2:fs)as(1:fs/2-1)];
    fn=-fs/2:fs/2-1;
```

```

subplot(3,1,k); stem(fn,2*evenf, '. ')
ylabel('Amplitude')
end
%End of script file.

```

Note that the modulation index is set by varying the frequency of the sinusoidal message signal f_m with the peak deviation held constant at 100 Hz. Since f_m takes on the values of 200, 100, and 20, the corresponding values of the modulation index are 0.5, 1, and 5, respectively. The corresponding spectra of the complex envelope signal are illustrated as a function of frequency in Figure 4.7. ■

COMPUTER EXAMPLE 4.2

We now consider the calculation of the two-sided amplitude spectrum of an FM (or PM) signal using the FFT algorithm. As can be seen from the MATLAB code, a modulation index of 3 is assumed. Note the manner in which the amplitude spectrum is divided into positive frequency and negative frequency segments (line nine in the following program). The student should verify that the various spectral components fall at the correct frequencies and that the amplitudes are consistent with the Bessel function values given in Table 4.1. The output of the MATLAB program is illustrated in Figure 4.8.

```

%File: c4ce2.m
fs=1000; %sampling frequency
delt=1/fs; %sampling increment
t=0:delt:1-delt; %time vector
npts=length(t); %number of points
fn=(0:npts)-(fs/2); %frequency vector for plot
m=3*cos(2*pi*25*t); %modulation
xc=sin(2*pi*200*t+m); %modulated carrier
asxc=(1/npts)*abs(fft(xc)); %amplitude spectrum

```

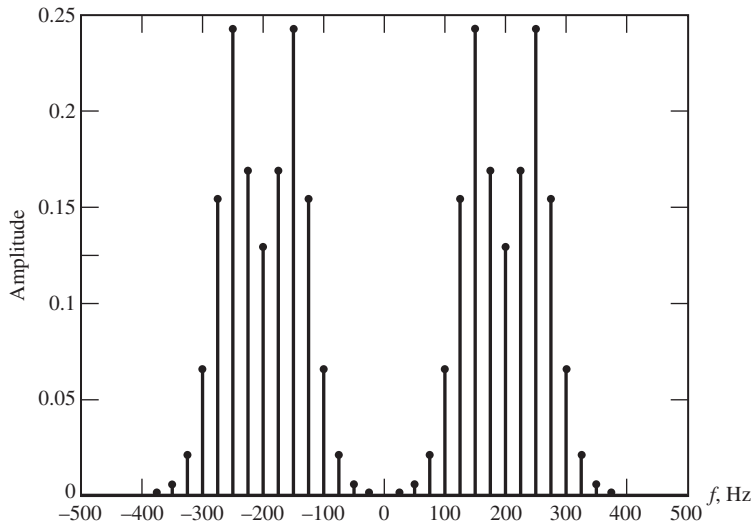


Figure 4.8

Two-sided amplitude spectrum computed using the FFT algorithm.

```

evenf=[asxc((npts/2):npts)asxc(1:npts/2)]; %even amplitude spectrum
stem(fn,evenf, '. ');
xlabel('Frequency-Hz')
ylabel('Amplitude')
%End of script.file.

```

4.1.3 Power in an Angle-Modulated Signal

The power in an angle-modulated signal is easily computed from (4.1). Squaring (4.1) and taking the time-average value yields

$$\langle x_c^2(t) \rangle = A_c^2 \langle \cos^2[2\pi f_c t + \phi(t)] \rangle \quad (4.37)$$

which can be written as

$$\langle x_c^2(t) \rangle = \frac{1}{2} A_c^2 + \frac{1}{2} A_c^2 \langle \cos\{2[2\pi f_c t + \phi(t)]\} \rangle \quad (4.38)$$

If the carrier frequency is large so that $x_c(t)$ has negligible frequency content in the region of DC, the second term in (4.38) is negligible and

$$\langle x_c^2(t) \rangle = \frac{1}{2} A_c^2 \quad (4.39)$$

Thus, the power contained in the output of an angle modulator is independent of the message signal. Given that, for this example, $x_c(t)$ is a sinusoid, although of varying frequency, the result expressed by (4.39) was to be expected. Constant transmitter power, independent of the message signal, is one important difference between angle modulation and linear modulation.

4.1.4 Bandwidth of Angle-Modulated Signals

Strictly speaking, the bandwidth of an angle-modulated signal is infinite, since angle modulation of a carrier results in the generation of an infinite number of sidebands. However, it can be seen from the series expansion of $J_n(\beta)$ (Appendix F, Table F.3) that for large n

$$J_n(\beta) \approx \frac{\beta^n}{2^n n!} \quad (4.40)$$

Thus, for fixed β ,

$$\lim_{n \rightarrow \infty} J_n(\beta) = 0 \quad (4.41)$$

This behavior can also be seen from the values of $J_n(\beta)$ given in Table 4.1. Since the values of $J_n(\beta)$ become negligible for sufficiently large n , the bandwidth of an angle-modulated signal can be defined by considering only those terms that contain significant power. The power ratio

P_r is defined as the ratio of the power contained in the carrier ($n = 0$) component and the k components on each side of the carrier to the total power in $x_c(t)$. Thus,

$$P_r = \frac{\frac{1}{2}A_c^2 \sum_{n=-k}^k J_n^2(\beta)}{\frac{1}{2}A_c^2} = \sum_{n=-k}^k J_n^2(\beta) \quad (4.42)$$

or simply

$$P_r = J_0^2(\beta) + 2 \sum_{n=1}^k J_n^2(\beta) \quad (4.43)$$

Bandwidth for a particular application is often determined by defining an acceptable power ratio, solving for the required value of k using a table of Bessel functions, and then recognizing that the resulting bandwidth is

$$B = 2kf_m \quad (4.44)$$

The acceptable value of the power ratio is dictated by the particular application of the system. Two power ratios are depicted in Table 4.1: $P_r \geq 0.7$ and $P_r \geq 0.98$. The value of n corresponding to k for $P_r \geq 0.7$ is indicated by a single underscore, and the value of n corresponding to k for $P_r \geq 0.98$ is indicated by a double underscore. For $P_r \geq 0.98$ it is noted that n is equal to the integer part of $1 + \beta$, so that

$$B \cong 2(\beta + 1)f_m \quad (4.45)$$

which will take on greater significance when Carson's rule is discussed in the following paragraph.

The preceding expression assumes sinusoidal modulation, since the modulation index β is defined only for sinusoidal modulation. For arbitrary $m(t)$, a generally accepted expression for bandwidth results if the deviation ratio D is defined as

$$D = \frac{\text{peak frequency deviation}}{\text{bandwidth of } m(t)} \quad (4.46)$$

which is

$$D = \frac{f_d}{W}(\max |m(t)|) \quad (4.47)$$

The deviation ratio plays the same role for nonsinusoidal modulation as the modulation index plays for sinusoidal systems. Replacing β by D and replacing f_m by W in (4.45), we obtain

$$B = 2(D + 1)W \quad (4.48)$$

This expression for bandwidth is generally referred to as *Carson's rule*. If $D \ll 1$, the bandwidth is approximately $2W$, and the signal is known as a *narrowband angle-modulated signal*. Conversely, if $D \gg 1$, the bandwidth is approximately $2DW = 2f_d(\max |m(t)|)$, which is twice the peak frequency deviation. Such a signal is known as a *wideband angle-modulated signal*.

EXAMPLE 4.2

In this example we consider an FM modulator with output

$$x_c(t) = 100 \cos[2\pi(1000)t + \phi(t)] \quad (4.49)$$

The modulator operates with $f_d = 8$ and has the input message signal

$$m(t) = 5 \cos 2\pi(8)t \quad (4.50)$$

The modulator is followed by a bandpass filter with a center frequency of 1000 Hz and a bandwidth of 56 Hz, as shown in Figure 4.9(a). Our problem is to determine the power at the filter output.

The peak deviation is $5f_d$ or 40 Hz, and $f_m = 8$ Hz. Thus, the modulation index is $40/5 = 8$. This yields the single-sided amplitude spectrum shown in Figure 4.9(b). Figure 4.9(c) shows the passband of the bandpass filter. The filter passes the component at the carrier frequency and three components on each side of the carrier. Thus, the power ratio is

$$P_r = J_0^2(5) + 2[J_1^2(5) + J_2^2(5) + J_3^2(5)] \quad (4.51)$$

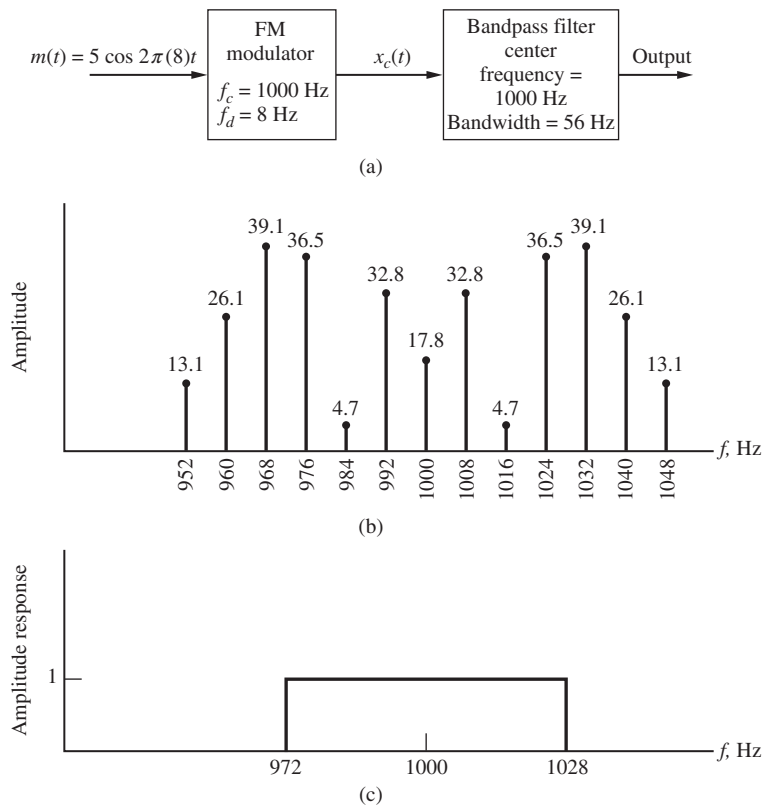


Figure 4.9

System and spectra for Example 4.2. (a) FM system. (b) Single-sided spectrum of modulator output. (c) Amplitude response of bandpass filter.

which is

$$P_r = (0.178)^2 + 2 [(0.328)^2 + (0.047)^2 + (0.365)^2] \quad (4.52)$$

This yields

$$P_r = 0.518 \quad (4.53)$$

The power at the output of the modulator is

$$\overline{x_c^2} = \frac{1}{2} A_c^2 = \frac{1}{2} (100)^2 = 5000 \text{ W} \quad (4.54)$$

The power at the filter output is the power of the modulator output multiplied by the power ratio. Thus, the power at the filter output is

$$P_r \overline{x_c^2} = 2589 \text{ W} \quad (4.55)$$

EXAMPLE 4.3

In the development of the spectrum of an angle-modulated signal, it was assumed that the message signal was a single sinusoid. We now consider a somewhat more general problem in which the message signal is the sum of two sinusoids. Let the message signal be

$$m(t) = A \cos(2\pi f_1 t) + B \cos(2\pi f_2 t) \quad (4.56)$$

For FM modulation the phase deviation is therefore given by

$$\phi(t) = \beta_1 \sin(2\pi f_1 t) + \beta_2 \sin(2\pi f_2 t) \quad (4.57)$$

where $\beta_1 = A f_d / f_1 > 1$ and $\beta_2 = B f_d / f_2$. The modulator output for this case becomes

$$x_c(t) = A_c \cos[2\pi f_c t + \beta_1 \sin(2\pi f_1 t) + \beta_2 \sin(2\pi f_2 t)] \quad (4.58)$$

which can be expressed as

$$x_c(t) = A_c \operatorname{Re} \left\{ e^{j\beta_1 \sin(2\pi f_1 t)} e^{j\beta_2 \sin(2\pi f_2 t)} e^{j2\pi f_c t} \right\} \quad (4.59)$$

Using the Fourier series

$$e^{j\beta_1 \sin(2\pi f_1 t)} = \sum_{n=-\infty}^{\infty} J_n(\beta_1) e^{j2\pi n f_1 t} \quad (4.60)$$

and

$$e^{j\beta_2 \sin(2\pi f_2 t)} = \sum_{m=-\infty}^{\infty} J_m(\beta_2) e^{j2\pi m f_2 t} \quad (4.61)$$

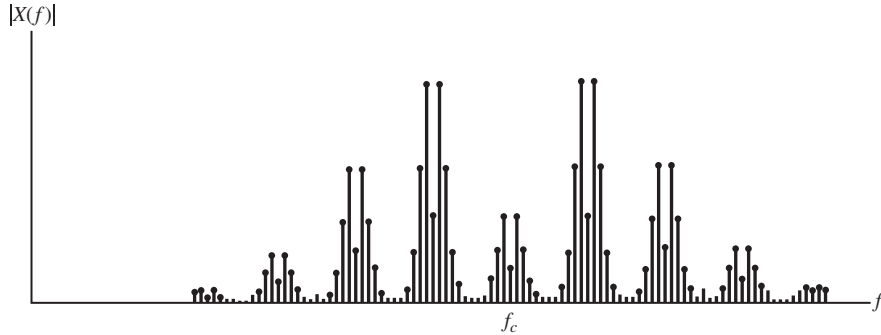


Figure 4.10
Amplitude spectrum for (4.63) with $\beta_1 = \beta_2$ and $f_2 = 12f_1$.

the modulator output can be written

$$x_c(t) = A_c \operatorname{Re} \left\{ \left[\sum_{n=-\infty}^{\infty} J_n(\beta_1) e^{j2\pi f_1 t} \sum_{m=-\infty}^{\infty} J_m(\beta_2) e^{j2\pi f_2 t} \right] e^{j2\pi f_c t} \right\} \quad (4.62)$$

Taking the real part gives

$$x_c(t) = A_c \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} J_n(\beta_1) J_m(\beta_2) \cos[2\pi(f_c + nf_1 + mf_2)t] \quad (4.63)$$

Examination of the signal $x_c(t)$ shows that it not only contains frequency components at $f_c + nf_1$ and $f_c + mf_2$, but also contains frequency components at $f_c + nf_1 + mf_2$ for all combinations of n and m . Therefore, the spectrum of the modulator output due to a message signal consisting of the sum of two sinusoids contains additional components over the spectrum formed by the superposition of the two spectra resulting from the individual message components. This example therefore illustrates the nonlinear nature of angle modulation. The spectrum resulting from a message signal consisting of the sum of two sinusoids is shown in Figure 4.10 for the case in which $\beta_1 = \beta_2$ and $f_2 = 12f_1$. ■

COMPUTER EXAMPLE 4.3

In this computer example we consider a MATLAB program for computing the amplitude spectrum of an FM (or PM) signal having a message signal consisting of a pair of sinusoids. The single-sided amplitude spectrum is calculated. (Note the multiplication by 2 in the definitions of `ampspec1` and `ampspec2` in the following computer program.) The single-sided spectrum is determined by using only the positive portion of the spectrum represented by the first $N/2$ points generated by the FFT program. In the following program N is represented by the variable `npts`.

Two plots are generated for the output. Figure 4.11(a) illustrates the spectrum with a single sinusoid for the message signal. The frequency of this sinusoidal component (50 Hz) is evident. Figure 4.11(b) illustrates the amplitude spectrum of the modulator output when a second component, having a frequency of 5 Hz, is added to the message signal. For this exercise the modulation index associated with each component of the message signal was carefully chosen to ensure that the spectra were essentially constrained to lie within the bandwidth defined by the carrier frequency (250 Hz).

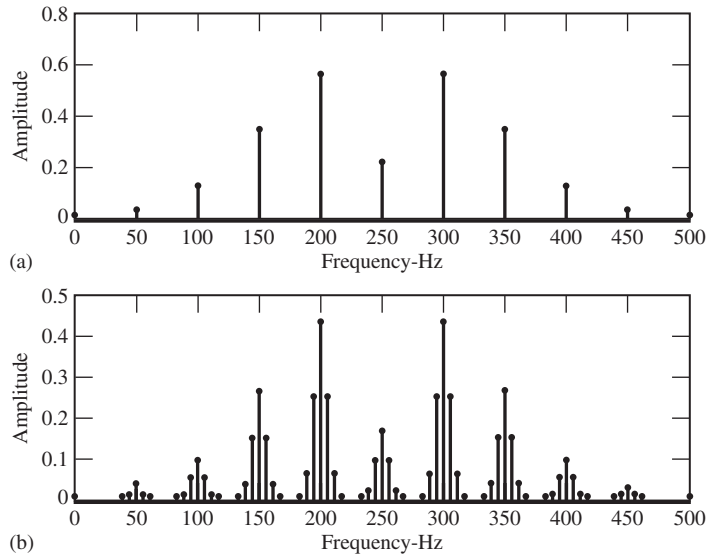


Figure 4.11

Frequency modulation spectra. (a) Single-tone modulating signal. (b) Two-tone modulating signal.

```

%File: c4ce3.m
fs=1000; %sampling frequency
delt=1/fs; %sampling increment
t=0:delt:1-delt; %time vector
npts=length(t); %number of points
fn=(0:(npts/2))*(fs/npts); %frequency vector for plot
m1=2*cos(2*pi*50*t); %modulation signal 1
m2=2*cos(2*pi*50*t)+1*cos(2*pi*5*t); %modulation signal 2
xc1=sin(2*pi*250*t+m1); %modulated carrier 1
xc2=sin(2*pi*250*t+m2); %modulated carrier 2
asxc1=(2/npts)*abs(fft(xc1)); %amplitude spectrum 1
asxc2=(2/npts)*abs(fft(xc2)); %amplitude spectrum 2
ampspec1=asxc1(1:(npts/2)+1); %positive frequency portion 1
ampspec2=asxc2(1:(npts/2)+1); %positive frequency portion 2
subplot(211)
stem(fn,ampspec1,'.k');
xlabel('Frequency-Hz')
ylabel('Amplitude')
subplot(212)
stem(fn,ampspec2,'.k');
xlabel('Frequency-Hz')
ylabel('Amplitude')
subplot(111)
%End of script file.

```

4.1.5 Narrowband-to-Wideband Conversion

One technique for generating wideband FM is illustrated in Figure 4.12. The carrier frequency of the narrowband frequency modulator is f_{c1} , and the peak frequency deviation is f_{d1} . The frequency multiplier multiplies the argument of the input sinusoid by n . In other words, if the

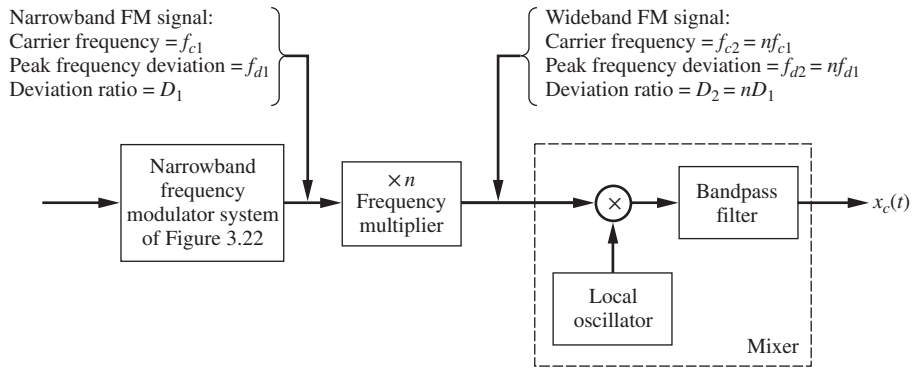


Figure 4.12
Frequency modulation utilizing narrowband-to-wideband conversion.

input of a frequency multiplier is

$$x(t) = A_c \cos[2\pi f_0 t + \phi(t)] \quad (4.64)$$

the output of the frequency multiplier is

$$y(t) = A_c \cos[2\pi n f_0 t + n\phi(t)] \quad (4.65)$$

Assuming that the output of the local oscillator is

$$e_{LO}(t) = 2 \cos(2\pi f_{LO} t) \quad (4.66)$$

results in

$$\begin{aligned} e(t) = & A_c \cos[2\pi(n f_0 + f_{LO})t + n\phi(t)] \\ & + A_c \cos[2\pi(n f_0 - f_{LO})t + n\phi(t)] \end{aligned} \quad (4.67)$$

for the multiplier output. This signal is then filtered, using a bandpass filter having center frequency f_c , given by

$$f_c = n f_0 + f_{LO} \quad \text{or} \quad f_c = n f_0 - f_{LO}$$

This yields the output

$$x_c(t) = A_c \cos[2\pi f_c t + n\phi(t)] \quad (4.68)$$

The bandwidth of the bandpass filter is chosen in order to pass the desired term in (4.67). One can use Carson's rule to determine the bandwidth of the bandpass filter if the transmitted signal is to contain 98% of the power in $x_c(t)$.

The central idea in narrowband-to-wideband conversion is that the frequency multiplier changes both the carrier frequency and the deviation ratio by a factor of n , whereas the mixer changes the effective carrier frequency but does not affect the deviation ratio. This technique of implementing wideband frequency modulation is known as *indirect frequency modulation*.

EXAMPLE 4.4

A narrowband-to-wideband converter is implemented as shown in Figure 4.12. The output of the narrowband frequency modulator is given by (4.64) with $f_0 = 100,000$ Hz. The peak frequency deviation of $\phi(t)$ is 50 Hz and the bandwidth of $\phi(t)$ is 500 Hz. The wideband output $x_c(t)$ is to have a carrier frequency of 85 MHz and a deviation ratio of 5. In this example we determine the frequency multiplier factor, n , two possible local oscillator frequencies, and the center frequency and the bandwidth of the bandpass filter.

The deviation ratio at the output of the narrowband FM modulator is

$$D_1 = \frac{f_{d1}}{W} = \frac{50}{500} = 0.1 \quad (4.69)$$

The frequency multiplier factor is therefore

$$n = \frac{D_2}{D_1} = \frac{5}{0.1} = 50 \quad (4.70)$$

Thus, the carrier frequency at the output of the narrowband FM modulator is

$$nf_0 = 50(100,000) = 5 \text{ MHz} \quad (4.71)$$

The two permissible frequencies for the local oscillator are

$$85 + 5 = 90 \text{ MHz} \quad (4.72)$$

and

$$85 - 5 = 80 \text{ MHz} \quad (4.73)$$

The center frequency of the bandpass filter must be equal to the desired carrier frequency of the wideband output. Thus, the center frequency of the bandpass filter is 85 MHz. The bandwidth of the bandpass filter is established using Carson's rule. From (4.48) we have

$$B = 2(D + 1)W = 2(5 + 1)(500) \quad (4.74)$$

Thus,

$$B = 6000 \text{ Hz} \quad (4.75)$$

■

4.2 DEMODULATION OF ANGLE-MODULATED SIGNALS

The demodulation of an FM signal requires a circuit that yields an output proportional to the frequency deviation of the input. Such circuits are known as *frequency discriminators*.¹ If the input to an ideal discriminator is the angle-modulated signal

$$x_r(t) = A_c \cos[2\pi f_c t + \phi(t)] \quad (4.76)$$

the output of the ideal discriminator is

$$y_D(t) = \frac{1}{2\pi} K_D \frac{d\phi}{dt} \quad (4.77)$$

¹The terms *frequency demodulator* and *frequency discriminator* are equivalent.

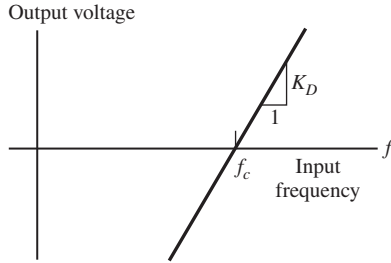


Figure 4.13
Ideal discriminator.

For FM, $\phi(t)$ is given by

$$\phi(t) = 2\pi f_d \int^t m(\alpha) d\alpha \quad (4.78)$$

so that (4.77) becomes

$$y_D(t) = K_D f_d m(t) \quad (4.79)$$

The constant K_D is known as the *discriminator constant* and has units of volts per Hz. Since an ideal discriminator yields an output signal proportional to the frequency deviation of a carrier, it has a linear frequency-to-voltage transfer function, which passes through zero at $f = f_c$. This is illustrated in Figure 4.13.

The system characterized by Figure 4.13 can also be used to demodulate PM signals. Since $\phi(t)$ is proportional to $m(t)$ for PM, $y_D(t)$ given by (4.77) is proportional to the time derivative of $m(t)$ for PM inputs. Integration of the discriminator output yields a signal proportional to $m(t)$. Thus, a demodulator for PM can be implemented as an FM discriminator followed by an integrator. We define the output of a PM discriminator as

$$y_D(t) = K_D k_p m(t) \quad (4.80)$$

It will be clear from the context whether $y_D(t)$ and K_D refer to an FM or a PM system.

An approximation to the characteristic illustrated in Figure 4.13 can be obtained by the use of a differentiator followed by an envelope detector, as shown in Figure 4.14. If the input to the differentiator is

$$x_r(t) = A_c \cos[2\pi f_c t + \phi(t)] \quad (4.81)$$

the output of the differentiator is

$$e(t) = -A_c \left[2\pi f_c + \frac{d\phi}{dt} \right] \sin[2\pi f_c t + \phi(t)] \quad (4.82)$$

This is exactly the same form as an AM signal, except for the phase deviation $\phi(t)$. Thus, after differentiation, envelope detection can be used to recover the message signal. The envelope of $e(t)$ is

$$y(t) = A_c \left(2\pi f_c + \frac{d\phi}{dt} \right) \quad (4.83)$$

and is always positive if

$$f_c > -\frac{1}{2\pi} \frac{d\phi}{dt} \quad \text{for all } t$$

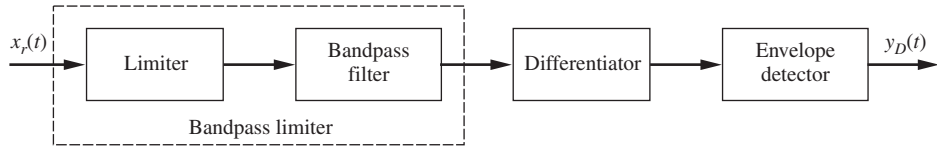


Figure 4.14
FM discriminator implementation.

which is usually satisfied since f_c is typically significantly greater than the bandwidth of the message signal. Thus, the output of the envelope detector is

$$y_D(t) = A_c \frac{d\phi}{dt} = 2\pi A_c f_d m(t) \quad (4.84)$$

assuming that the DC term, $2\pi A_c f_c$, is removed. Comparing (4.84) and (4.79) shows that the discriminator constant for this discriminator is

$$K_D = 2\pi A_c \quad (4.85)$$

We will see later that interference and channel noise perturb the amplitude A_c of $x_r(t)$. In order to ensure that the amplitude at the input to the differentiator is constant, a *limiter* is placed before the differentiator. The output of the limiter is a signal of square-wave type, which is $K \text{sgn}[x_r(t)]$. A bandpass filter having center frequency f_c is then placed after the limiter to convert the signal back to the sinusoidal form required by the differentiator to yield the response defined by (4.82). The cascade combination of a limiter and a bandpass filter is known as a *bandpass limiter*. The complete discriminator is illustrated in Figure 4.14.

The process of differentiation can often be realized using a time-delay implementation, as shown in Figure 4.15. The signal $e(t)$, which is the input to the envelope detector, is given by

$$e(t) = x_r(t) - x_r(t - \tau) \quad (4.86)$$

which can be written

$$\frac{e(t)}{\tau} = \frac{x_r(t) - x_r(t - \tau)}{\tau} \quad (4.87)$$

Since, by definition,

$$\lim_{\tau \rightarrow 0} \frac{e(t)}{\tau} = \lim_{\tau \rightarrow 0} \frac{x_r(t) - x_r(t - \tau)}{\tau} = \frac{dx_r(t)}{dt} \quad (4.88)$$

it follows that for small τ ,

$$e(t) \cong \tau \frac{dx_r(t)}{dt} \quad (4.89)$$

This is, except for the constant factor τ , identical to the envelope detector input shown in Figure 4.15 and defined by (4.82). The resulting discriminator constant K_D is $2\pi A_c \tau$. There are many other techniques that can be used to implement a discriminator. Later in this chapter we will examine the phase-locked loop, which is an especially attractive, and common, implementation.

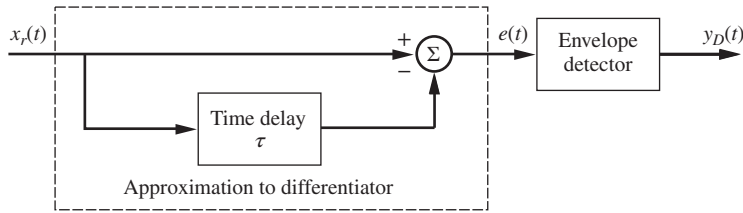


Figure 4.15
Discriminator implementation using a time delay and envelope detection.

EXAMPLE 4.5

Consider the simple RC network shown in Figure 4.16(a). The transfer function is

$$H(f) = \frac{R}{R + 1/j2\pi fC} = \frac{j2\pi fRC}{1 + j2\pi fRC} \quad (4.90)$$

The amplitude response is shown in Figure 4.16(b). If all frequencies present in the input are low, so that

$$f \ll \frac{1}{2\pi RC}$$

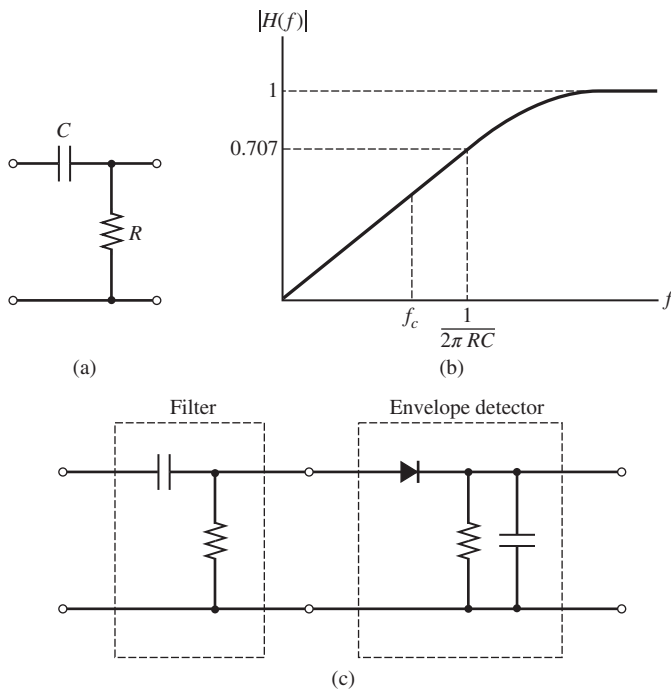


Figure 4.16
Implementation of a simple frequency discriminator based on a high-pass filter. (a) RC network. (b) Transfer function. (c) Discriminator.

the transfer function can be approximated by

$$H(f) = j2\pi fRC \quad (4.91)$$

Thus, for small f , the RC network has the linear amplitude–frequency characteristic required of an ideal discriminator. Equation (4.91) illustrates that for small f , the RC filter acts as a differentiator with gain RC . Thus, the RC network can be used in place of the differentiator in Figure 4.14 to yield a discriminator with

$$K_D = 2\pi A_c RC \quad (4.92)$$

This example again illustrates the essential components of a frequency discriminator, a circuit that has an amplitude response linear with frequency and an envelope detector. However, a highpass filter does not in general yield a practical implementation. This can be seen from the expression for K_D . Clearly the 3-dB frequency of the filter, $1/2\pi RC$, must exceed the carrier frequency f_c . In commercial FM broadcasting, the carrier frequency at the discriminator input, i.e., the IF frequency, is on the order of 10 MHz. As a result, the discriminator constant K_D is very small indeed.

A solution to the problem of a very small K_D is to use a bandpass filter, as illustrated in Figure 4.17. However, as shown in Figure 4.17(a), the region of linear operation is often unacceptably small. In addition, use of a bandpass filter results in a DC bias on the discriminator output. This DC bias could of course be removed by a blocking capacitor, but the blocking capacitor would negate an inherent advantage of FM—namely, that FM has DC response. One can solve these problems by using two filters with staggered center frequencies f_1 and f_2 , as shown in Figure 4.17(b). The magnitudes of the envelope detector outputs following the two filters are proportional to $|H_1(f)|$ and $|H_2(f)|$. Subtracting these two outputs yields the overall characteristic

$$H(f) = |H_1(f)| - |H_2(f)| \quad (4.93)$$

as shown in Figure 4.17(c). The combination, known as a *balanced discriminator*, is linear over a wider frequency range than would be the case for either filter used alone, and it is clearly possible to make $H(f_c) = 0$.

In Figure 4.17(d), a center-tapped transformer supplies the input signal $x_c(t)$ to the inputs of the two bandpass filters. The center frequencies of the two bandpass filters are given by

$$f_i = \frac{1}{2\pi\sqrt{L_i C_i}} \quad (4.94)$$

for $i = 1, 2$. The envelope detectors are formed by the diodes and the resistor–capacitor combinations $R_e C_e$. The output of the upper envelope detector is proportional to $|H_1(f)|$, and the output of the lower envelope detector is proportional to $|H_2(f)|$. The output of the upper envelope detector is the positive portion of its input envelope, and the output of the lower envelope detector is the negative portion of its input envelope. Thus, $y_D(t)$ is proportional to $|H_1(f)| - |H_2(f)|$. The term *balanced discriminator* is used because the response to the undeviated carrier is balanced so that the net response is zero.

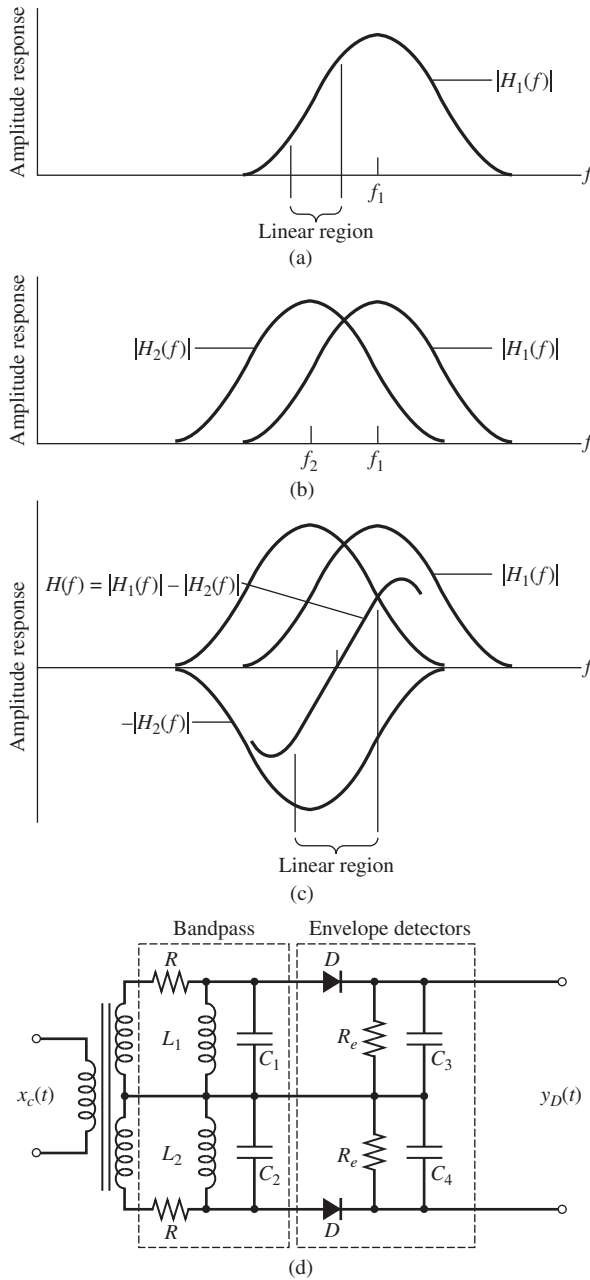


Figure 4.17

Derivation of a balanced discriminator. (a) Bandpass filter. (b) Stagger-tuned bandpass filters. (c) Amplitude response of a balanced discriminator. (d) Typical implementation of a balanced discriminator.

4.3 FEEDBACK DEMODULATORS: THE PHASE-LOCKED LOOP

We have previously studied the technique of FM to AM conversion for demodulating an angle-modulated signal. We shall see in Chapter 8 that improved performance in the presence of noise can be gained by utilizing a feedback demodulator. The subject of this section is the phase-locked loop (PLL), which is a basic form of the feedback demodulator. Phase-locked loops are widely used in today's communication systems, not only for demodulation of angle-modulated signals but also for carrier and symbol synchronization, for frequency synthesis, and as the basic building block for a variety of digital demodulators. Phase-locked loops are flexible in that they can be used in a wide variety of applications, are easily implemented, and give superior performance to many other techniques. It is therefore not surprising that they are ubiquitous in modern communications systems. Therefore, a detailed look at the PLL is justified.

4.3.1 Phase-Locked Loops for FM and PM Demodulation

A block diagram of a PLL is shown in Figure 4.18. The basic PLL contains four basic elements. These are

1. Phase detector
2. Loop filter
3. Loop amplifier (assume $\mu = 1$)
4. Voltage-controlled oscillator (VCO).

In order to understand the operation of the PLL, assume that the input signal is given by

$$x_r(t) = A_c \cos[2\pi f_c t + \phi(t)] \quad (4.95)$$

and that the VCO output signal is given by

$$e_0(t) = A_v \sin[2\pi f_c t + \theta(t)] \quad (4.96)$$

(Note that these are in phase quadrature.) There are many different types of phase detectors, all having different operating properties. For our application, we assume that the phase detector is a multiplier followed by a lowpass filter to remove the second harmonic of the carrier. We also assume that an inverter is present to remove the minus sign resulting from the multiplication. With these assumptions, the output of the phase detector becomes

$$e_d(t) = \frac{1}{2} A_c A_v K_d \sin[\phi(t) - \theta(t)] = \frac{1}{2} A_c A_v K_d \sin[\psi(t)] \quad (4.97)$$

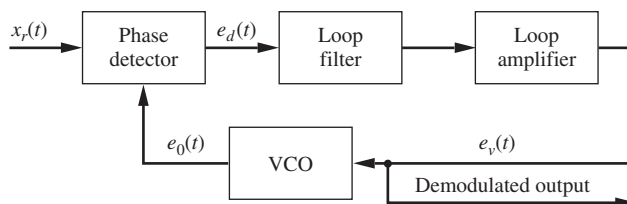


Figure 4.18
Phase-locked loop for demodulation of FM.

where K_d is the phase detector constant and $\psi(t) = \phi(t) - \theta(t)$ is the phase error. Note that for small phase error the two inputs to the multiplier are approximately orthogonal so that the result of the multiplication is an odd function of the phase error $\phi(t) - \theta(t)$. This is a necessary requirement so that the phase detector can distinguish between positive and negative phase errors. This illustrates why the PLL input and VCO output must be in phase quadrature.

The output of the phase detector is filtered, amplified, and applied to the VCO. A VCO is essentially a frequency modulator in which the frequency deviation of the output, $d\theta/dt$, is proportional to the VCO input signal. In other words,

$$\frac{d\theta}{dt} = K_v e_v(t) \text{ rad/s} \quad (4.98)$$

which yields

$$\theta(t) = K_v \int^t e_v(\alpha) d\alpha \quad (4.99)$$

The parameter K_v is known as the *VCO constant* and is measured in radians per second per unit of input.

From the block diagram of the PLL it is clear that

$$E_v(s) = F(s)E_d(s) \quad (4.100)$$

where $F(s)$ is the transfer function of the loop filter. In the time domain the preceding expression is

$$e_v(\alpha) = \int^t e_d(\lambda) f(\alpha - \lambda) d\lambda \quad (4.101)$$

which follows by simply recognizing that multiplication in the frequency domain is convolution in the time domain. Substitution of (4.97) into (4.101) and this result into (4.99) gives

$$\theta(t) = K_t \int^t \int^{\alpha} \sin[\phi(\lambda) - \theta(\lambda)] f(\alpha - \lambda) d\lambda d\alpha \quad (4.102)$$

where K_t is the total loop gain defined by

$$K_t = \frac{1}{2} A_v A_c K_d K_v \quad (4.103)$$

Equation (4.102) is the general expression relating the VCO phase $\theta(t)$ to the input phase $\phi(t)$. The system designer must select the loop filter transfer function $F(s)$, thereby defining the filter impulse response $f(t)$, and the loop gain K_t . We see from (4.103) that the loop gain is a function of the input signal amplitude A_v . Thus, PLL design requires knowledge of the input signal level, which is often unknown and time varying. This dependency on the input signal level is typically removed by placing a hard limiter at the loop input. If a limiter is used, the loop gain K_t is selected by appropriately choosing A_v , K_d , and K_v , which are all parameters of the PLL. The individual values of these parameters are arbitrary so long as their product gives the desired loop gain. However, hardware considerations typically place constraints on these parameters.

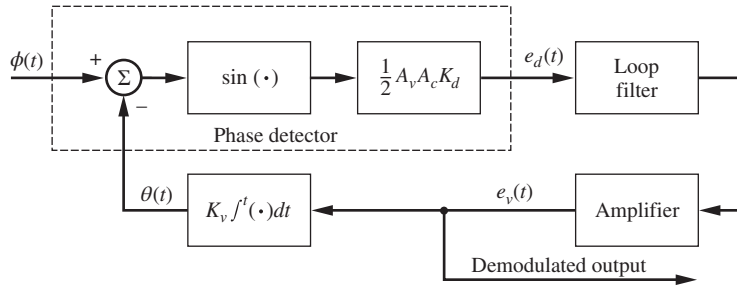


Figure 4.19
Nonlinear PLL model.

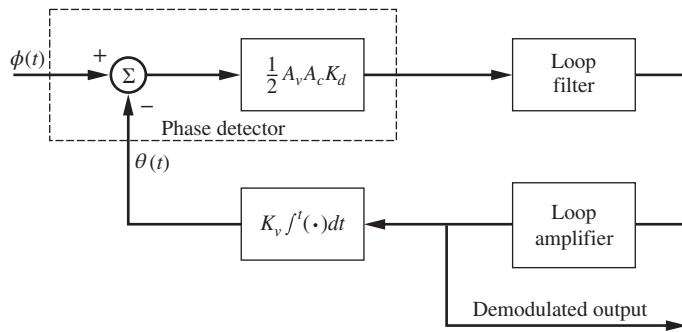


Figure 4.20
Linear PLL model.

Equation (4.102) defines the nonlinear model of the PLL, having a sinusoidal nonlinearity.² This model is illustrated in Figure 4.19. Since (4.102) is nonlinear, analysis of the PLL using (4.102) is difficult and often involves a number of approximations. In practice, we typically have interest in PLL operation in either the tracking mode or in the acquisition mode. In the acquisition mode the PLL is attempting to acquire a signal by synchronizing the frequency and phase of the VCO with the input signal. In the acquisition mode of operation, the phase errors are typically large, and the nonlinear model is required for analysis.

In the tracking mode, however, the phase error $\phi(t) - \theta(t)$ is typically small the linear model for PLL design and analysis in the tracking mode can be used. For small phase errors the sinusoidal nonlinearity may be neglected and the PLL becomes a linear feedback system. Equation (4.102) simplifies to the linear model defined by

$$\theta(t) = K_t \int^t \int^\alpha [\phi(\lambda) - \theta(\lambda)] f(\alpha - \lambda) d\lambda d\alpha \quad (4.104)$$

The linear model that results is illustrated in Figure 4.20. Both the nonlinear and linear models involve $\theta(t)$ and $\phi(t)$ rather than $x_r(t)$ and $e_0(t)$. However, note that if we know f_c , knowledge of $\theta(t)$ and $\phi(t)$ fully determine $x_r(t)$ and $e_0(t)$, as can be seen from (4.95) and (4.96). If the

²Many nonlinearities are possible and used for various purposes.

Table 4.3 Loop Filter Transfer Functions

PLL order	Loop filter transfer function, $F(s)$
1	1
2	$1 + \frac{a}{s} = (s + a)/s$
3	$1 + \frac{a}{s} + \frac{b}{s^2} = (s^2 + as + b)/s^2$

PLL is in phase lock, $\theta(t) \cong \phi(t)$, and it follows that, assuming FM,

$$\frac{d\theta(t)}{dt} \cong \frac{d\phi(t)}{dt} = 2\pi f_d m(t) \quad (4.105)$$

and the VCO frequency deviation is a good estimate of the input frequency deviation, which is proportional to the message signal. Since the VCO frequency deviation is proportional to the VCO input $e_v(t)$, it follows that the input is proportional to $m(t)$ if (4.105) is satisfied. Thus, the VCO input, $e_v(t)$, is the demodulated output for FM systems.

The form of the loop filter transfer function $F(s)$ has a profound effect on both the tracking and acquisition behavior of the PLL. In the work to follow we will have interest in first-order, second-order, and third-order PLLs. The loop filter transfer functions for these three cases are given in Table 4.3. Note that the order of the PLL exceeds the order of the loop filter by one. The extra integration results from the VCO as we will see in the next section. We now consider the PLL in both the tracking and acquisition mode. Tracking mode operation is considered first since the model is linear and, therefore, more straightforward.

4.3.2 Phase-Locked Loop Operation in the Tracking Mode: The Linear Model

As we have seen, in the tracking mode the phase error is small, and linear analysis can be used to define PLL operation. Considerable insight into PLL operation can be gained by investigating the steady-state errors for first-order, second-order, and third-order PLLs with a variety of input signals.

The Loop Transfer Function and Steady-State Errors

The frequency-domain equivalent of Figure 4.20 is illustrated in Figure 4.21. It follows from Figure 4.21 and (4.104) that

$$\Theta(s) = K_t[\Phi(s) - \Theta(s)] \frac{F(s)}{s} \quad (4.106)$$

from which the transfer function relating the VCO phase to the input phase is

$$H(s) = \frac{\Theta(s)}{\Phi(s)} = \frac{K_t F(s)}{s + K_t F(s)} \quad (4.107)$$

immediately follows. The Laplace transform of the phase error is

$$\Psi(s) = \Phi(s) - \Theta(s) \quad (4.108)$$

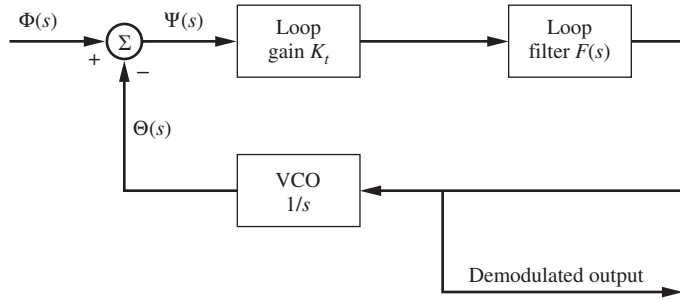


Figure 4.21
Linear PLL model in the frequency domain.

Therefore, we can write the transfer function relating the phase error to the input phase as

$$G(s) = \frac{\Psi(s)}{\Phi(s)} = \frac{\Phi(s) - \Theta(s)}{\Phi(s)} = 1 - H(s) \quad (4.109)$$

so that

$$G(s) = \frac{s}{s + K_t F(s)} \quad (4.110)$$

The steady-state error can be determined through the final value theorem from Laplace transform theory. The final value theorem states that the $\lim_{t \rightarrow \infty} a(t)$ is given by $\lim_{s \rightarrow 0} sA(s)$, where $a(t)$ and $A(s)$ are a Laplace transform pair.

In order to determine the steady-state errors for various loop orders, we assume that the phase deviation has the general form

$$\phi(t) = \pi R t^2 + 2\pi f_{\Delta} t + \theta_0, \quad t > 0 \quad (4.111)$$

The corresponding frequency deviation is

$$\frac{1}{2\pi} \frac{d\phi}{dt} = R t + f_{\Delta}, \quad t > 0 \quad (4.112)$$

We see that the frequency deviation is the sum of a frequency ramp, R Hz/s, and a frequency step f_{Δ} . The Laplace transform of $\phi(t)$ is

$$\Phi(s) = \frac{2\pi R}{s^3} + \frac{2\pi f_{\Delta}}{s^2} + \frac{\theta_0}{s} \quad (4.113)$$

Thus, the steady-state phase error is given by

$$\psi_{ss} = \lim_{s \rightarrow 0} s \left[\frac{2\pi R}{s^3} + \frac{2\pi f_{\Delta}}{s^2} + \frac{\theta_0}{s} \right] G(s) \quad (4.114)$$

where $G(s)$ is given by (4.110).

In order to generalize, consider the third-order filter transfer function defined in Table 4.4:

$$F(s) = \frac{1}{s^2} (s^2 + as + b) \quad (4.115)$$

If $a = 0$ and $b = 0$, $F(s) = 1$, which is the loop filter transfer function for a first-order PLL. If $a \neq 0$, and $b = 0$, $F(s) = (s + a)/s$, which defines the loop filter for second-order PLL. With

Table 4.4 Steady-state Errors

PLL order	$\theta_0 \neq 0$ $f_{\Delta} = 0$ $R = 0$	$\theta_0 \neq 0$ $f_{\Delta} \neq 0$ $R = 0$	$\theta_0 \neq 0$ $f_{\Delta} \neq 0$ $R \neq 0$
1 ($a = 0, b = 0$)	0	$2\pi f_{\Delta}/K_t$	∞
2 ($a \neq 0, b = 0$)	0	0	$2\pi R/K_t$
3 ($a \neq 0, b \neq 0$)	0	0	0

$a \neq 0$ and $b \neq 0$ we have a third-order PLL. We can therefore use $F(s)$, as defined by (4.115) with a and b taking on appropriate values, to analyze first-order, second-order, and third-order PLLs.

Substituting (4.115) into (4.110) yields

$$G(s) = \frac{s^3}{s^3 + K_t s^2 + K_t a s + K_t b} \quad (4.116)$$

Using the expression for $G(s)$ in (4.114) gives the steady-state phase error expression

$$\psi_{ss} = \lim_{s \rightarrow 0} \frac{s(\theta_0 s^2 + 2\pi f_{\Delta} s + 2\pi R)}{s^3 + K_t s^2 + K_t a s + K_t b} \quad (4.117)$$

We now consider the steady-state phase errors for first-order, second-order, and third-order PLLs. For various input signal conditions, defined by θ_0 , f_{Δ} , and R and the loop filter parameters a and b , the steady-state errors given in Table 4.4 can be determined. Note that a first-order PLL can track a phase step with a zero steady-state error. A second-order PLL can track a frequency step with zero steady-state error, and a third-order PLL can track a frequency ramp with zero steady-state error.

Note that for the cases given in Table 4.4 for which the steady-state error is nonzero and finite, the steady-state error can be made as small as desired by increasing the loop gain K_t . However, increasing the loop gain increases the loop bandwidth. When we consider the effects of noise in Chapter 8, we will see that increasing the loop bandwidth makes the PLL performance more sensitive to the presence of noise. We therefore see a trade-off between steady-state error and loop performance in the presence of noise.

EXAMPLE 4.6

We now consider a first-order PLL, which from (4.110) and (4.115), with $a = 0$ and $b = 0$, has the transfer function

$$H(s) = \frac{\Theta(s)}{\Phi(s)} = \frac{K_t}{s + K_t} \quad (4.118)$$

The loop impulse response is therefore

$$h(t) = K_t e^{-K_t t} u(t) \quad (4.119)$$

The limit of $h(t)$ as the loop gain K_t tends to infinity satisfies all properties of the delta function. Therefore,

$$\lim_{K_t \rightarrow \infty} K_t e^{-K_t t} u(t) = \delta(t) \quad (4.120)$$

which illustrates that for large loop gain $\theta(t) \approx \phi(t)$. This also illustrates, as we previously discussed, that the PLL serves as a demodulator for angle-modulated signals. Used as an FM demodulator, the VCO input is the demodulated output since the VCO input signal is proportional to the frequency deviation of the PLL input signal. For PM the VCO input is simply integrated to form the demodulated output, since phase deviation is the integral of frequency deviation. ■

EXAMPLE 4.7

As an extension of the preceding example, assume that the input to an FM modulator is $m(t) = Au(t)$. The resulting modulated carrier

$$x_c(t) = A_c \cos \left[2\pi f_c t + k_f A \int^t u(\alpha) d\alpha \right] \quad (4.121)$$

is to be demodulated using a first-order PLL. The demodulated output is to be determined.

This problem will be solved using linear analysis and the Laplace transform. The loop transfer function (4.118) is

$$\frac{\Theta(s)}{\Phi(s)} = \frac{K_t}{s + K_t} \quad (4.122)$$

The phase deviation of the PLL input $\phi(t)$ is

$$\phi(t) = Ak_f \int^t u(\alpha) d\alpha \quad (4.123)$$

The Laplace transform of $\phi(t)$ is

$$\Phi(s) = \frac{Ak_f}{s^2} \quad (4.124)$$

which gives

$$\Theta(s) = \frac{AK_f}{s^2} \frac{K_t}{s + K_t} \quad (4.125)$$

The Laplace transform of the defining equation of the VCO, (4.99), yields

$$E_v(s) = \frac{s}{K_v} \Theta(s) \quad (4.126)$$

so that

$$E_v(s) = \frac{AK_f}{K_v} \frac{K_t}{s(s + K_t)} \quad (4.127)$$

Partial fraction expansion gives

$$E_v(s) = \frac{AK_f}{K_v} \left(\frac{1}{s} - \frac{1}{s + K_t} \right) \quad (4.128)$$

Thus, the demodulated output is given by

$$e_v(t) = \frac{AK_f}{K_v} (1 - e^{-K_t t}) u(t) \quad (4.129)$$

Note that for $t \gg 1/K_t$ and $K_f = K_v$ we have, as desired, $e_v(t) = Au(t)$ as the demodulated output. The transient time is set by the total loop gain K_t , and K_f/K_v is simply an amplitude scaling of the demodulated output signal. ■

As previously mentioned, very large values of loop gain cannot be used in practical applications without difficulty. However, the use of appropriate loop filters allows good performance to be achieved with reasonable values of loop gain and bandwidth. These filters make the analysis more complicated than our simple example, as we shall soon see.

Even though the first-order PLL can be used for demodulation of angle-modulated signals and for synchronization, the first-order PLL has a number of drawbacks that limit its use for most applications. Among these drawbacks are the limited lock range and the nonzero steady-state phase error to a step-frequency input. Both these problems can be solved by using a second-order PLL, which is obtained by using a loop filter of the form

$$F(s) = \frac{s+a}{s} = 1 + \frac{a}{s} \quad (4.130)$$

This choice of loop filter results in what is generally referred to as a *perfect second-order PLL*. Note that the loop filter defined by (4.130) can be implemented using a single integrator, as will be demonstrated in a Computer Example 4.4 to follow.

The Second-Order PLL: Loop Natural Frequency and Damping Factor

With $F(s)$ given by (4.130), the transfer function (4.107) becomes

$$H(s) = \frac{\Theta(s)}{\Phi(s)} = \frac{K_t(s+a)}{s^2 + K_t s + K_t a} \quad (4.131)$$

We also can write the relationship between the phase error $\Psi(s)$ and the input phase $\Phi(s)$. From Figure 4.21 or (4.110), we have

$$G(s) = \frac{\Psi(s)}{\Phi(s)} = \frac{s^2}{s^2 + K_t s + K_t a} \quad (4.132)$$

Since the performance of a linear second-order system is typically parameterized in terms of the natural frequency and damping factor, we now place the transfer function in the standard form for a second-order system. The result is

$$\frac{\Psi(s)}{\Phi(s)} = \frac{s^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (4.133)$$

in which ζ is the damping factor and ω_n is the natural frequency. It follows from the preceding expression that the natural frequency is

$$\omega_n = \sqrt{K_t a} \quad (4.134)$$

and that the damping factor is

$$\zeta = \frac{1}{2} \sqrt{\frac{K_t}{a}} \quad (4.135)$$

A typical value of the damping factor is $1/\sqrt{2} = 0.707$. Note that this choice of damping factor gives a second-order Butterworth response.

In simulating a second-order PLL, one usually specifies the loop natural frequency and the damping factor and determines loop performance as a function of these two fundamental parameters. The PLL simulation model, however, is a function of the physical parameters K_t

and a . Equations (4.134) and (4.135) allow K_t and a to be written in terms of ω_n and ζ . The results are

$$a = \frac{\omega_n}{2\zeta} = \frac{\pi f_n}{\zeta} \quad (4.136)$$

and

$$K_t = 4\pi\zeta f_n \quad (4.137)$$

where $2\pi f_n = \omega_n$. These last two expressions will be used to develop the simulation program for the second-order PLL that is given in Computer Example 4.4.

EXAMPLE 4.8

We now work a simple second-order example. Assume that the input signal to the PLL experiences a small step change in frequency. (The step in frequency must be small to ensure that the linear model is applicable. We will consider the result of large step changes in PLL input frequency when we consider operation in the acquisition mode.) Since instantaneous phase is the integral of instantaneous frequency and integration is equivalent to division by s , the input phase due to a step in frequency of magnitude Δf is

$$\Phi(s) = \frac{2\pi\Delta f}{s^2} \quad (4.138)$$

From (4.133) we see that the Laplace transform of the phase error $\psi(t)$ is

$$\Psi(s) = \frac{\Delta\omega}{s^2 + 2\zeta\omega_n s + \omega_n^2} \quad (4.139)$$

Inverse transforming and replacing ω_n by $2\pi f_n$ yields, for $\zeta < 1$,

$$\psi(t) = \frac{\Delta f}{f_n\sqrt{1-\zeta^2}} e^{-2\pi\zeta f_n t} [\sin(2\pi f_n\sqrt{1-\zeta^2}t)]u(t) \quad (4.140)$$

and we see that $\psi(t) \rightarrow 0$ as $t \rightarrow \infty$. Note that the steady-state phase error is zero, which is consistent with the values shown in Table 4.4. ■

4.3.3 Phase-Locked Loop Operation in the Acquisition Mode

In the acquisition mode we must determine that the PLL actually achieves phase lock and the time required for the PLL to achieve phase lock. In order to show that the phase error signal tends to drive the PLL into lock, we will simplify the analysis by assuming a first-order PLL for which the loop filter transfer function $F(s) = 1$ or $f(t) = \delta(t)$. Simulation will be used for higher-order loops. Using the general nonlinear model defined by (4.102) with $h(t) = \delta(t)$ and applying the sifting property of the delta function yields

$$\theta(t) = K_t \int^t \sin[\phi(\alpha) - \theta(\alpha)]d\alpha \quad (4.141)$$

Taking the derivative of $\theta(t)$ gives

$$\frac{d\theta}{dt} = K_t \sin[\phi(t) - \theta(t)] \quad (4.142)$$

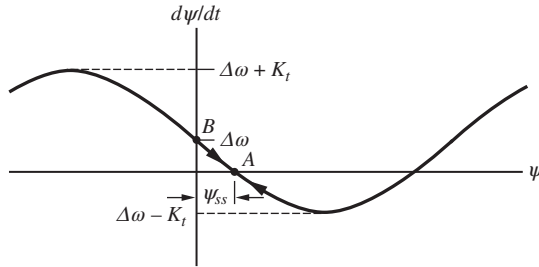


Figure 4.22
Phase-plane plot for sinusoidal nonlinearity.

Assume that the input to the FM modulator is a unit step so that the frequency deviation $d\phi/dt$ is a unit step of magnitude $2\pi\Delta f = \Delta\omega$. Let the phase error $\phi(t) - \theta(t)$ be denoted $\psi(t)$. This yields

$$\frac{d\theta}{dt} = \frac{d\phi}{dt} - \frac{d\psi}{dt} = \Delta\omega - \frac{d\psi}{dt} = K_t \sin \psi(t), \quad t \geq 0 \quad (4.143)$$

or

$$\frac{d\psi}{dt} + K_t \sin \psi(t) = \Delta\omega \quad (4.144)$$

This equation is shown in Figure 4.22. It relates the frequency error and the phase error and is known as a phase plane.

The phase plane tells us much about the operation of a nonlinear system. The PLL must operate with a phase error $\psi(t)$ and a frequency error $d\psi/dt$ that are consistent with (4.144). To demonstrate that the PLL achieves lock, assume that the PLL is operating with zero phase and frequency error prior to the application of the frequency step. When the step in frequency is applied, the frequency error becomes $\Delta\omega$. This establishes the initial operating point, point B in Figure 4.22, assuming $\Delta\omega > 0$. In order to determine the trajectory of the operating point, we need only recognize that since dt , a time increment, is always a positive quantity, $d\psi$ must be positive if $d\psi/dt$ is positive. Thus, in the upper half plane ψ increases. In other words, the operating point moves from left-to-right in the upper half plane. In the same manner, the operating point moves from right-to-left in the lower half plane, the region for which $d\psi/dt$ is less than zero. Thus, the operating point must move from point B to point A . When the operating point attempts to move from point A by a small amount, it is forced back to point A . Thus, point A is a stable operating point and is the steady-state operating point of the system. The steady-state phase error is ψ_{ss} , and the steady-state frequency error is zero as shown.

The preceding analysis illustrates that the loop locks only if there is an intersection of the operating curve with the $d\psi/dt = 0$ axis. Thus, if the loop is to lock, $\Delta\omega$ must be less than K_t . For this reason, K_t is known as the *lock range* for the first-order PLL.

The phase-plane plot for a first-order PLL with a frequency-step input is illustrated in Figure 4.23. The loop gain is $2\pi(50)$, and four values for the frequency step are shown: $\Delta f = 12, 24, 48, \text{ and } 55$ Hz. The steady-state phase errors are indicated by $A, B, \text{ and } C$ for frequency-step values of 12, 24, and 48 Hz, respectively. For $\Delta f = 55$, the loop does not lock but forever oscillates.

A mathematical development of the phase-plane plot of a second-order PLL is well beyond the level of our treatment here. However, the phase-plane plot is easily obtained, using

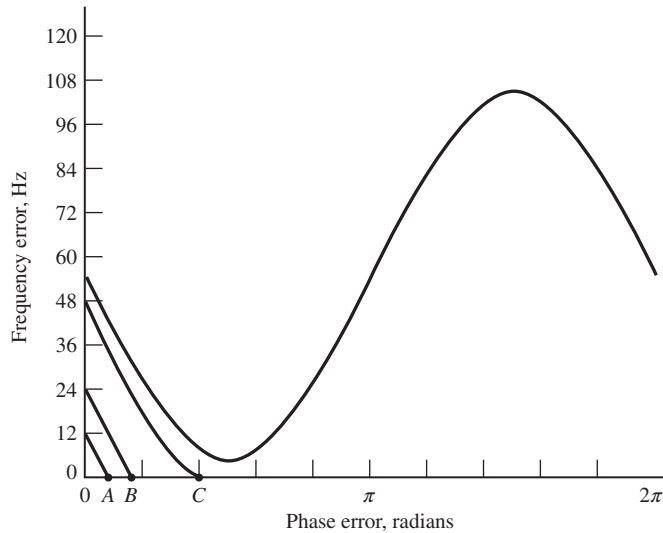


Figure 4.23
Phase-plane plot for first-order PLL for several step function frequency errors.

computer simulation. For illustrative purposes, assume a second-order PLL having a damping factor ζ of 0.707 and a natural frequency f_n of 10 Hz. For these parameters, the loop gain K_t is 88.9, and the filter parameter a is 44.4. The input to the PLL is assumed to be a step change in frequency at time $t = t_0$. Four values were used for the step change in frequency $\Delta\omega = 2\pi(\Delta f)$. These were $\Delta f = 20, 35, 40,$ and 45 Hz.

The results are illustrated in Figure 4.24. Note that for $\Delta f = 20$ Hz, the operating point returns to a steady-state value for which the frequency and phase error are both zero, as should be the case from Table 4.4. For $\Delta f = 35$ Hz, the phase plane is somewhat more complicated. The steady-state frequency error is zero, but the steady-state phase error is

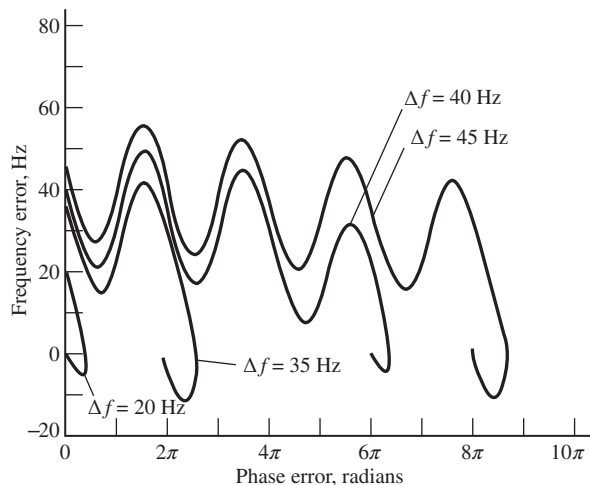


Figure 4.24
Phase-plane plot for second-order PLL for several step function frequency errors.

2π rad. We say that the PLL has slipped one cycle. Note that the steady-state error is zero mod(2π). The cycle-slipping phenomenon accounts for the nonzero steady-state phase error. The responses for $\Delta f = 40$ and 45 Hz illustrate that three and four cycles are slipped, respectively. The instantaneous VCO frequency is shown in Figure 4.24 for these four cases. The cycle-slipping behavior is clearly shown. The second-order PLL does indeed have an infinite lock range, and cycle slipping occurs until the phase error is within π rad of the steady-state value.

COMPUTER EXAMPLE 4.4

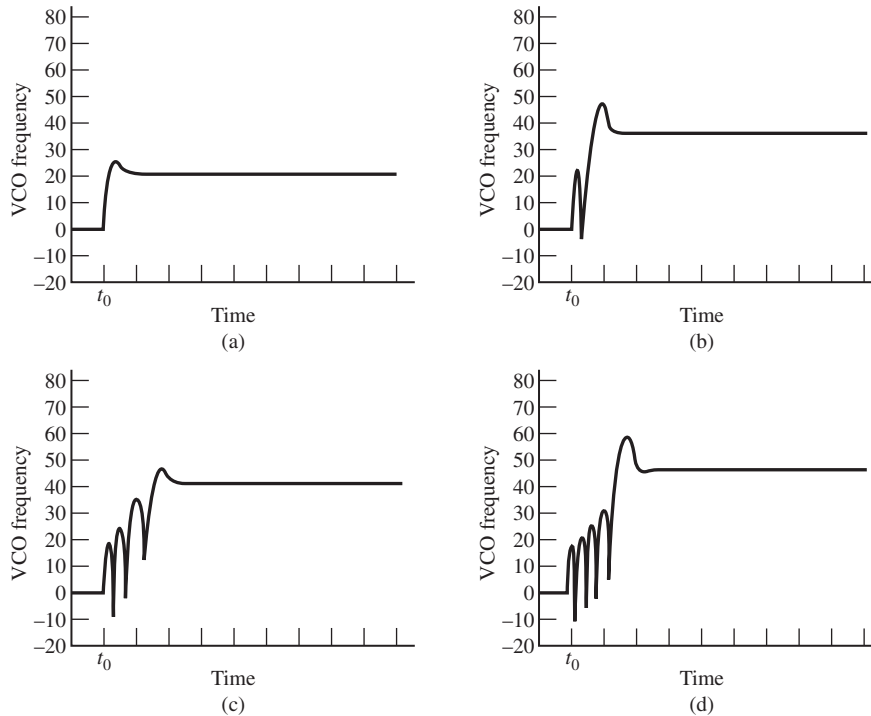
A simulation program is easily developed for the PLL. We simply replace the continuous-time integrators by appropriate discrete-time integrators. Many different discrete-time integrators exist, all of which are approximations to the continuous-time integrators. Here we consider only the trapezoidal approximation. Two integration routines are required; one for the loop filter and one for the VCO. The trapezoidal approximation is

$$y[n] = y[n-1] + (T/2)[x[n] + x[n-1]]$$

where $y[n]$ represents the current output of the integrator, $y[n-1]$ represents the previous integrator output, $x[n]$ represents the current integrator input, $x[n-1]$ represents the previous integrator input, and T represents the simulation step size, which is the reciprocal of the sampling frequency. The values of $y[n-1]$ and $x[n-1]$ must be initialized prior to entering the simulation loop. Initializing the integrator inputs and outputs usually result in a transient response. The parameter `nsettle`, which in the simulation program to follow, is set equal to 10% of the simulation run length, allows any initial transients to decay to negligible values prior to applying the loop input. The following simulation program is divided into three parts. The preprocessor defines the system parameters, the system input, and the parameters necessary for execution of the simulation, such as the sampling frequency. The simulation loop actually performs the simulation. Finally, the postprocessor allows for the data generated by the simulation to be displayed in a manner convenient for interpretation by the simulation user. Note that the postprocessor used here is interactive in that a menu is displayed and the simulation user can execute postprocessor commands without typing them. The simulation program given here assumes a frequency step on the loop input and can therefore be used to generate Figures 4.24 and 4.25.

```
%File: c4ce4.m
%beginning of preprocessor
clear all %be safe
fdel = input('Enter frequency step size in Hz > ');
n = input('Enter the loop natural frequency in Hz > ');
zeta = input('Enter zeta (loop damping factor) > ');
npts = 2000; %default number of simulation points
fs = 2000; %default sampling frequency
T = 1/fs;
t = (0:(npts-1))/fs; %time vector
nsettle = fix(npts/10) %set nsettle time as 0.1*npts
Kt = 4*pi*zeta*fn; %loop gain
a = pi*fn/zeta; %loop filter parameter
filt.inlast = 0; filt.out.last=0;
vco.inlast = 0; vco.out.last=0;
%end of preprocessor

%beginning of simulation loop
for i=1:npts
    if i < nsettle
```

**Figure 4.25**

Voltage-controlled frequency for four values of the input frequency step. (a) VCO frequency for $\Delta f = 20$ Hz. (b) VCO frequency for $\Delta f = 35$ Hz. (c) VCO frequency for $\Delta f = 40$ Hz. (d) VCO frequency for $\Delta f = 45$ Hz.

```

        fin(i) = 0;
        phin = 0;
    else
        fin(i) = fdel;
        phin = 2*pi*fdel*T*(i-nsettle);
    end
    s1=phin - vco.out;
    s2=sin(s1); %sinusoidal phase detector
    s3=Kt*s2;
    filt.in = a*s3;
    filt.out = filt.out_last + (T/2)*(filt.in + filt.in_last);
    filt.in_last = filt.in;
    filt.out_last = filt.out;
    vco.in = s3 + filt.out;
    vco.out = vco.out_last + (T/2)*(vco.in + vco.in_last);
    vco.in_last = vco.in;
    vco.out_last = vco.out;
    phierror(i)=s1;
    fvco(i)=vco.in/(2*pi);
    freqerror(i) = fin(i)-fvco(i);
end
%end of simulation loop

%beginning of postprocessor

```

```

kk = 0;
while kk == 0
    k = menu('Phase Lock Loop Postprocessor',...
            'Input Frequency and VCO Frequency',...
            'Phase Plane Plot',...
            'Exit Program');
    if k == 1
        plot(t,fin,t,fvco)
        title('Input Frequency and VCO Frequency')
        xlabel('Time - Seconds')
        ylabel('Frequency - Hertz')
        pause
    elseif k == 2
        plot(phierror/2/pi,freqerror)
        title('Phase Plane')
        xlabel('Phase Error / pi')
        ylabel('Frequency Error - Hz')
        pause
    elseif k == 3
        kk = 1;
    end
end
%end of postprocessor

```

4.3.4 Costas PLLs

We have seen that systems utilizing feedback can be used to demodulate angle-modulated carriers. A feedback system also can be used to generate the coherent demodulation carrier necessary for the demodulation of DSB signals. One system that accomplishes this is the Costas PLL illustrated in Figure 4.26. The input to the loop is the assumed DSB signal

$$x_r(t) = m(t) \cos(2\pi f_c t) \quad (4.145)$$

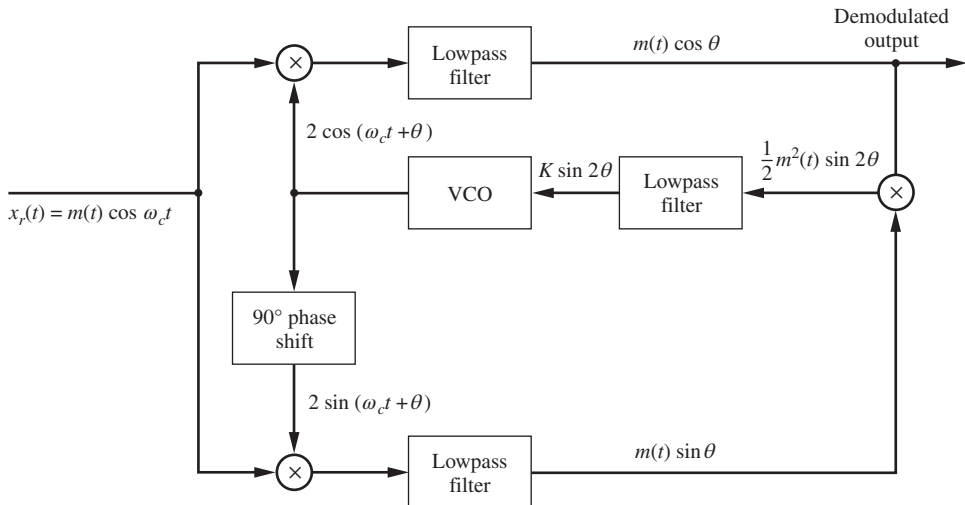


Figure 4.26
Costas phase-locked loop.

The signals at the various points within the loop are easily derived from the assumed input and VCO output and are included in Figure 4.26. The lowpass filter preceding the VCO is assumed to have sufficiently small bandwidth so that the output is approximately $K \sin(2\theta)$, essentially the DC value of the filter input. This signal drives the VCO such that θ is reduced. For sufficiently small θ , the output of the top lowpass filter is the demodulated output, and the output of the lower filter is negligible. We will later see in that the Costas PLL is useful in the implementation of digital receivers.

4.3.5 Frequency Multiplication and Frequency Division

Phase-locked loops also allow for simple implementation of frequency multipliers and dividers. There are two basic schemes. In the first scheme, harmonics of the input are generated, and the VCO tracks one of these harmonics. This scheme is most useful for implementing frequency multipliers. The second scheme is to generate harmonics of the VCO output and to phase lock one of these frequency components to the input. This scheme can be used to implement either frequency multipliers or frequency dividers.

Figure 4.27 illustrates the first technique. The limiter is a nonlinear device and therefore generates harmonics of the input frequency. If the input is sinusoidal, the output of the limiter is a square wave; therefore, odd harmonics are present. In the example illustrated, the VCO quiescent frequency [VCO output frequency f_c with $e_v(t)$ equal to zero] is set equal to $5f_0$. The result is that the VCO phase locks to the fifth harmonic of the input. Thus, the system shown multiplies the input frequency by 5.

Figure 4.28 illustrates frequency division by a factor of 2. The VCO quiescent frequency is $f_0/2$ Hz, but the VCO output waveform is a narrow pulse that has the spectrum shown. The

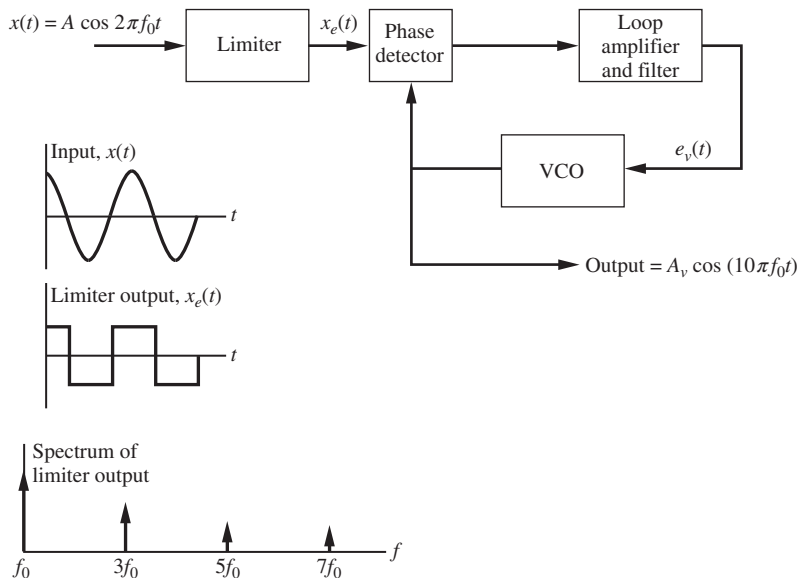


Figure 4.27
Phase-locked loop implementation of a frequency multiplier.

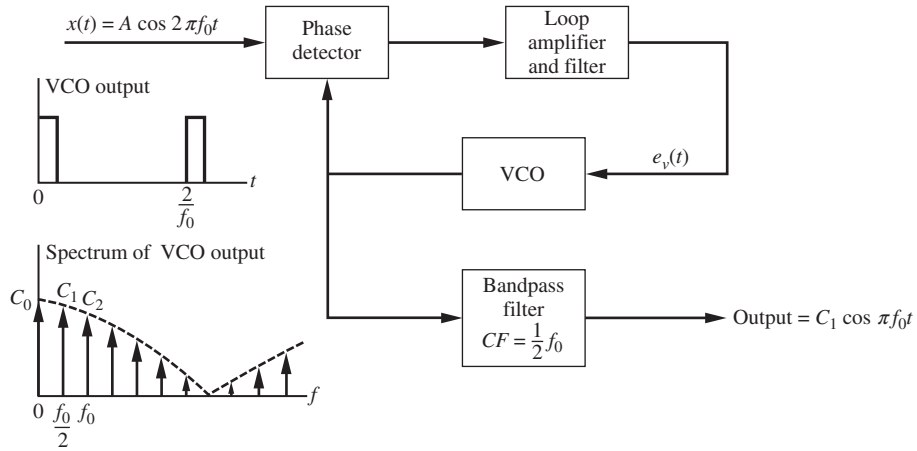


Figure 4.28
Phase-locked loop implementation of a frequency divider.

component at frequency f_0 phase locks to the input. A bandpass filter can be used to select the component desired from the VCO output spectrum. For the example shown, the center frequency of the bandpass filter should be $f_0/2$. The bandwidth of the bandpass filter must be less than the spacing between the components in the VCO output spectrum; in this case, this spacing is $f_0/2$. It is worth noting that the system shown in Figure 4.28 could also be used to multiply the input frequency by 5 by setting the center frequency of the bandpass filter to $5f_0$. Thus, this system could also serve as a $\times 5$ frequency multiplier, like the first example. Many variations of these basic techniques are possible.

4.4 INTERFERENCE IN ANGLE MODULATION

We now consider the effect of interference in angle modulation. We will see that the effect of interference in angle modulation is quite different from what was observed in linear modulation. Furthermore, we will see that the effect of interference in an FM system can be reduced by placing a lowpass filter at the discriminator output. We will consider this problem in considerable detail since the results will provide significant insight into the behavior of FM discriminators operating in the presence of noise, a subject to be treated in Chapter 8.

Assume that the input to a PM or FM ideal discriminator is an unmodulated carrier plus an interfering tone at frequency $f_c + f_i$. Thus, the input to the discriminator is assumed to have the form

$$x_i(t) = A_c \cos(2\pi f_c t) + A_i \cos[2\pi(f_c + f_i)t] \quad (4.146)$$

which can be written as

$$x_i(t) = A_c \cos(2\pi f_i t) + A_i \cos(2\pi f_i t) \cos(2\pi f_c t) - A_i \sin(2\pi f_i t) \sin(2\pi f_c t) \quad (4.147)$$

Writing the preceding expression in magnitude and phase form gives

$$x_r(t) = R(t) \cos[2\pi f_c t + \psi(t)] \quad (4.148)$$

in which the amplitude $R(t)$ is given by

$$R(t) = \sqrt{[A_c + A_i \cos(2\pi f_i t)]^2 + [A_i \sin(2\pi f_i t)]^2} \quad (4.149)$$

and the phase deviation $\psi(t)$ is given by

$$\psi(t) = \tan^{-1} \left(\frac{A_i \sin(2\pi f_i t)}{A_c + A_i \cos(2\pi f_i t)} \right) \quad (4.150)$$

If $A_c \gg A_i$, Equations (4.149) and (4.150) can be approximated

$$R(t) = A_c + A_i \cos(2\pi f_i t) \quad (4.151)$$

and

$$\psi(t) = \frac{A_i}{A_c} \sin(2\pi f_i t) \quad (4.152)$$

Thus, (4.148) is

$$x_r(t) = A_c \left[1 + \frac{A_i}{A_c} \cos(2\pi f_i t) \right] \cos \left[2\pi f_c t + \frac{A_i}{A_c} \sin(2\pi f_i t) \right] \quad (4.153)$$

The instantaneous phase deviation $\psi(t)$ is given by

$$\psi(t) = \frac{A_i}{A_c} \sin(2\pi f_i t) \quad (4.154)$$

Thus, the output of an ideal PM discriminator is

$$y_D(t) = K_D \frac{A_i}{A_c} \sin(2\pi f_i t) \quad (4.155)$$

and the output of an ideal FM discriminator is

$$y_D(t) = \frac{1}{2\pi} K_D \frac{d}{dt} \frac{A_i}{A_c} \sin(2\pi f_i t) \quad (4.156)$$

or

$$y_D(t) = K_D \frac{A_i}{A_c} f_i \cos(2\pi f_i t) \quad (4.157)$$

As with linear modulation, the discriminator output is a sinusoid of frequency f_i . The amplitude of the discriminator output, however, is proportional to the frequency f_i for the FM case. It can be seen that for small f_i , the interfering tone has less effect on the FM system than on the PM system and that the opposite is true for large values of f_i . Values of $f_i > W$, the bandwidth of $m(t)$, are of little interest, since they can be removed by a lowpass filter following the discriminator.

If the condition $A_i \ll A_c$ does not hold, the discriminator is not operating above threshold and the analysis becomes much more difficult. Some insight into this case can be obtained from the phasor diagram, which is obtained by writing (4.146) in the form

$$x_r(t) = \text{Re}[(A_c + A_i e^{j2\pi f_i t}) e^{j2\pi f_c t}] \quad (4.158)$$

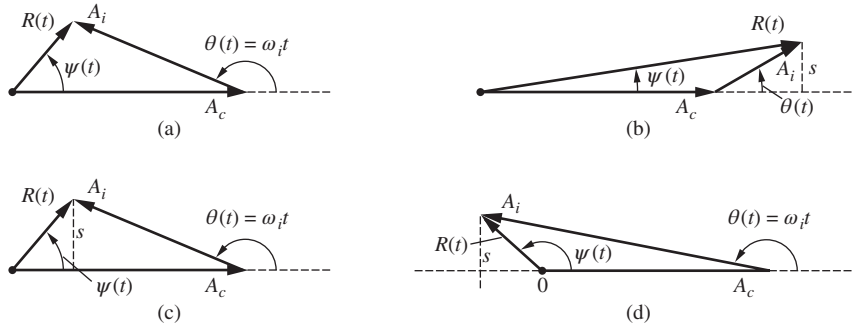


Figure 4.29 Phasor diagram for carrier plus single-tone interference. (a) Phasor diagram for general $\theta(t)$. (b) Phasor diagram for $\theta(t) \approx 0$. (c) Phasor diagram for $\theta(t) \approx \pi$ and $A_i \leq A_c$. (d) Phasor diagram for $\theta(t) \approx \pi$ and $A_i \geq A_c$.

The term in parentheses defines a phasor, which is the complex envelope signal. The phasor diagram is shown in Figure 4.29(a). The carrier phase is taken as the reference and the interference phase is

$$\theta(t) = 2\pi f_i t \tag{4.159}$$

Approximations to the phase of the resultant $\psi(t)$ can be determined using the phasor diagram.

From Figure 4.29(b) we see that the magnitude of the discriminator output will be small when $\theta(t)$ is near zero. This results because for $\theta(t)$ near zero, a given change in $\theta(t)$ will result in a much smaller change in $\psi(t)$. Using the relationship between arc length s , angle θ , and radius r , which is $s = \theta r$, we obtain

$$s = \theta(t)A_i \approx (A_c + A_i)\psi(t), \quad \theta(t) \approx 0 \tag{4.160}$$

Solving for $\psi(t)$ yields

$$\psi(t) \approx \frac{A_i}{A_c + A_i} \omega_i t \tag{4.161}$$

Since the discriminator output is defined by

$$y_D(t) = \frac{K_D}{2\pi} \frac{d\psi}{dt} \tag{4.162}$$

we have

$$y_D(t) = K_D \frac{A_i}{A_c - A_i} f_i, \quad \theta(t) \approx 0 \tag{4.163}$$

This is a positive quantity for $f_i > 0$ and a negative quantity for $f_i < 0$.

If A_i is slightly less than A_c , denoted $A_i \lesssim A_c$, and $\theta(t)$ is near π , a small positive change in $\theta(t)$ will result in a large negative change in $\psi(t)$. The result will be a negative spike appearing at the discriminator output. From Figure 4.29(c) we can write

$$s = A_i(\pi - \theta(t)) \approx (A_c - A_i)\psi(t), \quad \theta(t) \approx \pi \tag{4.164}$$

which can be expressed

$$\psi(t) \approx \frac{A_i(\pi - 2\pi f_i t)}{A_c - A_i} \tag{4.165}$$

Using (4.162), we see that the discriminator output is

$$y_D(t) = -K_D \frac{A_i}{A_c - A_i} f_i, \quad \theta(t) \approx \pi \quad (4.166)$$

This is a negative quantity for $f_i > 0$ and a positive quantity for $f_i < 0$.

If A_i is slightly greater than A_c , denoted $A_i \gtrsim A_c$, and $\theta(t)$ is near π , a small positive change in $\theta(t)$ will result in a large positive change in $\psi(t)$. The result will be a positive spike appearing at the discriminator output. From Figure 4.29(d) we can write

$$s = A_i[\pi - \theta(t)] \approx (A_i - A_c)[\pi - \psi(t)], \quad \theta(t) \approx \pi \quad (4.167)$$

Solving for $\psi(t)$ and differentiating gives the discriminator output

$$y_D(t) \approx -K_D \frac{A_i}{A_c - A_i} f_i \quad (4.168)$$

Note that this is a positive quantity for $f_i > 0$ and a negative quantity for $f_i < 0$.

The phase deviation and discriminator output waveforms are shown in Figure 4.30 for $A_i = 0.1A_c$, $A_i = 0.9A_c$, and $A_i = 1.1A_c$. Figure 4.30(a) illustrates that for small A_i the phase deviation and the discriminator output are nearly sinusoidal as predicted by the results of the small interference analysis given in (4.154) and (4.157). For $A_i = 0.9A_c$, we see that we have a negative spike at the discriminator output as predicted by (4.166). For $A_c = 1.1A_c$, we have a positive spike at the discriminator output as predicted by (4.168). Note that for $A_i > A_c$, the origin of the phasor diagram is encircled as $\theta(t)$ goes from 0 to 2π . In other words, $\psi(t)$ goes from 0 to 2π as $\theta(t)$ goes from 0 to 2π . The origin is not encircled if $A_i < A_c$. Thus, the integral

$$\int_T \left(\frac{d\psi}{dt} \right) dt = \begin{cases} 2\pi, & A_i > A_c \\ 0, & A_i < A_c \end{cases} \quad (4.169)$$

where T is the time required for $\theta(t)$ to go from $\theta(t) = 0$ to $\theta(t) = 2\pi$. In other words, $T = 1/f_i$. Thus, the area under the discriminator output curve is 0 for parts (a) and (b) of Figure 4.30 and $2\pi K_D$ for the discriminator output curve in Figure 4.30(c). The origin encirclement phenomenon will be revisited in Chapter 8 when demodulation of FM signals in the presence of noise is examined. An understanding of the interference results presented here will provide valuable insights when noise effects are considered.

For operation above threshold $A_i \ll A_c$, the severe effect of interference on FM for large f_i can be reduced by placing a filter, called a *de-emphasis filter*, at the FM discriminator output. This filter is typically a simple RC lowpass filter with a 3-dB frequency considerably less than the modulation bandwidth W . The de-emphasis filter effectively reduces the interference for large f_i , as shown in Figure 4.31. For large frequencies, the magnitude of the transfer function of a first-order filter is approximately $1/f$. Since the amplitude of the interference increases linearly with f_i for FM, the output is constant for large f_i , as shown in Figure 4.31.

Since $f_3 < W$, the lowpass de-emphasis filter distorts the message signal in addition to combating interference. The distortion can be avoided by passing the message signal, prior to modulation, through a highpass *pre-emphasis filter* that has a transfer function equal to the reciprocal of the transfer function of the lowpass de-emphasis filter. Since the

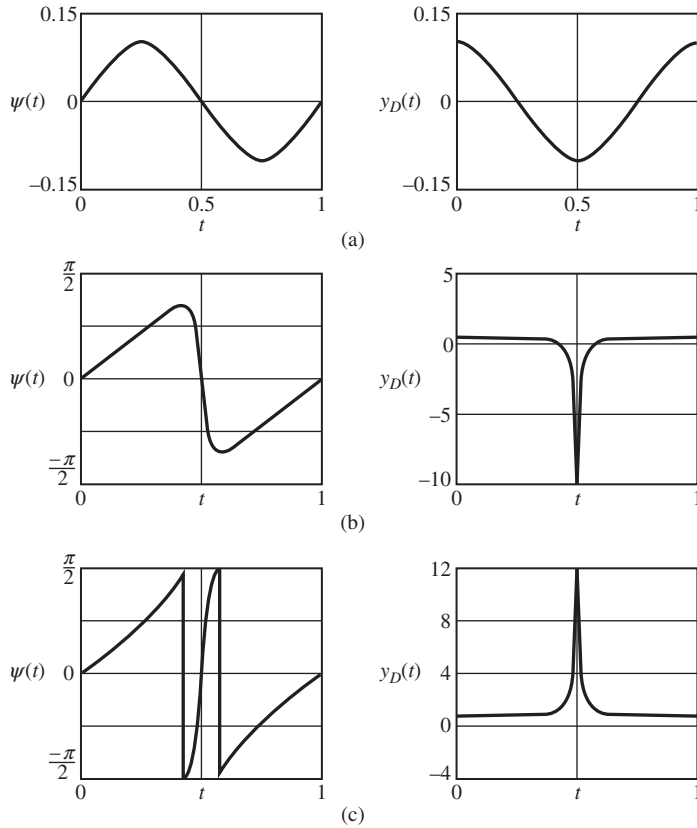


Figure 4.30 Phase deviation and discriminator output due to interference. (a) Phase deviation and discriminator output for $A_i = 0.1A_c$. (b) Phase deviation and discriminator output for $A_i = 0.9A_c$. (c) Phase deviation and discriminator output for $A_i = 1.1A_c$.

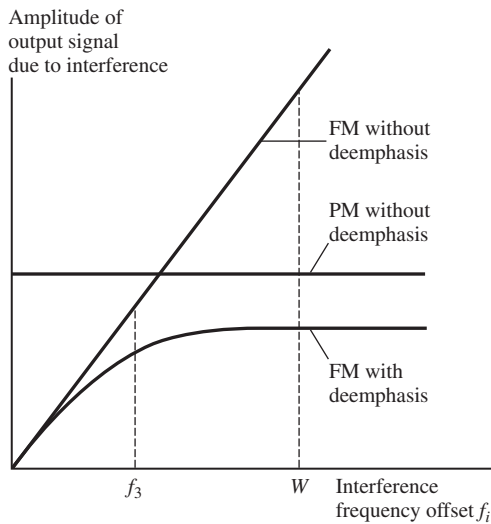


Figure 4.31 Amplitude of discriminator output due to interference.

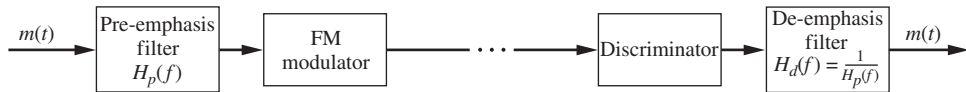


Figure 4.32
Frequency modulation system with pre-emphasis and de-emphasis.

transfer function of the cascade combination of the pre-emphasis and de-emphasis filters is unity, there is no detrimental effect on the modulation. This yields the system shown in Figure 4.32.

The improvement offered by the use of pre-emphasis and de-emphasis is not gained without a price. The highpass pre-emphasis filter amplifies the high-frequency components relative to lower-frequency components, which can result in increased deviation and bandwidth requirements. We shall see in Chapter 8, when the impact of channel noise is studied, that the use of pre-emphasis and de-emphasis often provides significant improvement in system performance with very little added complexity or implementation costs.

The idea of pre-emphasis and/or de-emphasis filtering has found application in a number of areas. For example, signals recorded on long-playing (LP) records are, prior to recording, filtered using a highpass pre-emphasis filter. This attenuates the low-frequency content of the signal being recorded. Since the low-frequency components typically have large amplitudes, the distance between the grooves on the record must be increased to accommodate these large amplitude signals if pre-emphasis filtering were not used. The impact of more widely spaced record grooves is reduced recording time. The playback equipment applies de-emphasis filtering to compensate for the pre-emphasis filtering used in the recording process. In the early days of LP recording, several different pre-emphasis filter designs were used among different record manufacturers. The playback equipment was consequently required to provide for all of the different pre-emphasis filter designs in common use. This later became standardized. With modern digital recording techniques this is no longer an issue.

4.5 ANALOG PULSE MODULATION

As defined in the preceding chapter, analog pulse modulation results when some attribute of a pulse varies continuously in one-to-one correspondence with a sample value. Three attributes can be readily varied: amplitude, width, and position. These lead to pulse-amplitude modulation (PAM), pulse-width modulation (PWM), and pulse-position modulation (PPM) as can be seen by referring back to Figure 3.25. We looked at PAM in the previous chapter. We now briefly look at PWM and PPM.

4.5.1 Pulse-Width Modulation (PWM)

A PWM waveform, as illustrated in Figure 3.25, consists of a sequence of pulses with each pulse having a width proportional to the values of the message signal at the sampling instants. If the message is 0 at the sampling time, the width of the PWM pulse is typically $\frac{1}{2}T_s$. Thus, pulse widths less than $\frac{1}{2}T_s$ correspond to negative sample values, and pulse widths greater

than $\frac{1}{2}T_s$ correspond to positive sample values. The modulation index β is defined so that for $\beta = 1$, the maximum pulse width of the PWM pulses is exactly equal to the sampling period $1/T_s$. Pulse width modulation is seldom used in modern communications systems. However, PWM has found uses in other areas. For example, PWM is used extensively for DC motor control in which motor speed is proportional to the width of the pulses. Large amplitude pulses are therefore avoided. Since the pulses have equal amplitude, the energy in a given pulse is proportional to the pulse width. The sample values can be recovered from a PWM waveform by lowpass filtering.

COMPUTER EXAMPLE 4.5

Due to the complexity of determining the spectrum of a PWM signal we resort to using the FFT to determine the spectrum. The MATLAB program follows.

```
%File: c4ce5.m
clear all;                               %be safe
N = 20000;                                %FFT size
N.samp = 200;                              %200 samples per period
f = 1;                                     %frequency
beta = 0.7;                                %modulation index
period = N/N.samp;                          %sample period (Ts)
Max.width = beta*N/N.samp;                 %maximum width
y = zeros(1,N);                             %initialize
for n=1:N.samp
    x = sin(2*pi*f*(n-1)/N.samp);
    width = (period/2)+round((Max.width/2)*x);
    for k=1:Max.width
        nn = (n-1)*period+k;
        if k<width
            y(nn) = 1;                       %pulse amplitude
        end
    end
end
ymm = y-mean(y);                           %remove mean
z = (1/N)*fft(ymm,N);                       %compute FFT
subplot(211)
stem(0:999,abs(z(1:1000)),'.k')
xlabel('Frequency - Hz.')
ylabel('Amplitude')
subplot(212)
stem(180:220,abs(z(181:221)),'.k')
xlabel('Frequency - Hz.')
ylabel('Amplitude')
%End of script file.
```

In the preceding program the message signal is a sinusoid having a frequency of 1 Hz. The message signal is sampled at 200 samples per period or 200 Hz. The FFT covers 10 periods of the waveform. The spectrum, as determined by the FFT, is illustrated in Figures 4.33(a) and (b). Figure 4.33(a) illustrates the spectrum in the range $0 \leq f \leq 1000$. Since the individual spectral components are spaced 1 Hz apart, corresponding to the 1-Hz sinusoid, they cannot be clearly seen. Figure 4.26(b) illustrates the spectrum in the neighborhood of $f = 200$ Hz. The spectrum in this region reminds us of a Fourier–Bessel spectrum for a sinusoid FM modulated by a pair of sinusoids (see Figure 4.10). We observe that PWM is much like angle modulation.

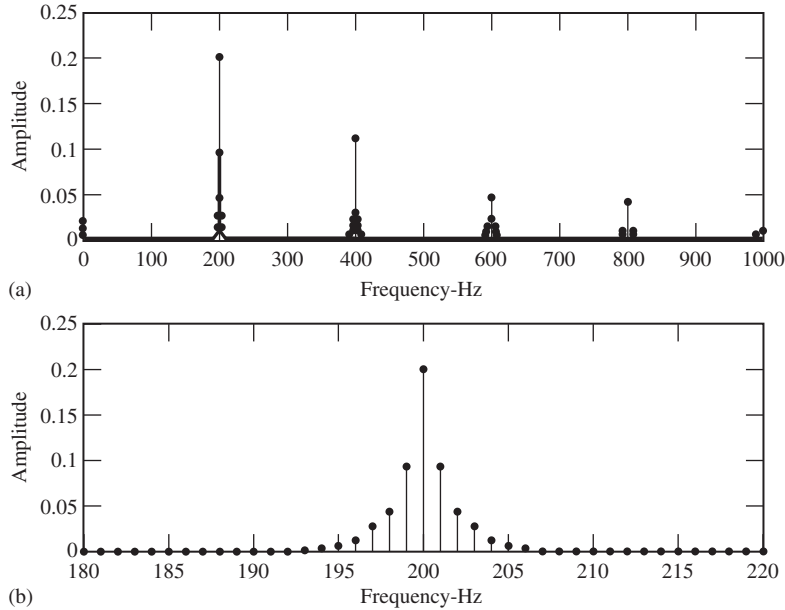


Figure 4.33
Spectrum of a PWM signal. (a) Spectrum for $0 \leq f \leq 1000$ Hz. (b) Spectrum in the neighborhood of $f = 200$.

4.5.2 Pulse-Position Modulation (PPM)

A PPM signal consists of a sequence of pulses in which the pulse displacement from a specified time reference is proportional to the sample values of the information-bearing signal. The PPM signal was illustrated in Figure 3.25 and can be represented by the expression

$$x(t) = g(t - t_n) \quad (4.170)$$

where $g(t)$ represents the shape of the individual pulses, and the occurrence times t_n are related to the values of the message signal $m(t)$ at the sampling instants nT_s , as discussed in the preceding paragraph. The spectrum of a PPM signal is very similar to the spectrum of a PWM signal. (See the computer examples at the end of the chapter.)

If the time axis is slotted so that a given range of sample values is associated with each slot, the pulse positions are quantized, and a pulse is assigned to a given slot depending upon the sample value. Slots are nonoverlapping and are therefore orthogonal. If a given sample value is assigned to one of M slots, the result is M -ary orthogonal communications, which will be studied in detail in Chapter 11. Pulse-position modulation is finding a number of applications in the area of ultra-wideband communications.³ (Note that short-duration pulses require a large bandwidth for transmission.)

³See, for example, R. A. Scholtz, "Multiple Access with Time-Hopping Impulse Modulation," *Proceedings of the IEEE 1993 MILCOM Conference*, 1993, and Reed (2005).

4.6 MULTIPLEXING

In many applications, a large number of data sources are located at a common point, and it is desirable to transmit these signals simultaneously using a single communication channel. This is accomplished using multiplexing. We will now examine several different types of multiplexing, each having advantages and disadvantages.

4.6.1 Frequency-Division Multiplexing

Frequency-division multiplexing (FDM) is a technique whereby several message signals are translated, using modulation, to different spectral locations and added to form a baseband signal. The carriers used to form the baseband are usually referred to as *subcarriers*. If desired, the baseband signal can be transmitted over a single channel using a single modulation process. Several different types of modulation can be used to form the baseband, as illustrated in Figure 4.34. In this example, there are N information signals contained in the baseband. Observation of the baseband spectrum in Figure 4.34(c) suggests that baseband modulator 1 is a DSB modulator with subcarrier frequency f_1 . Modulator 2 is an upper-sideband SSB modulator, and modulator N is an angle modulator.

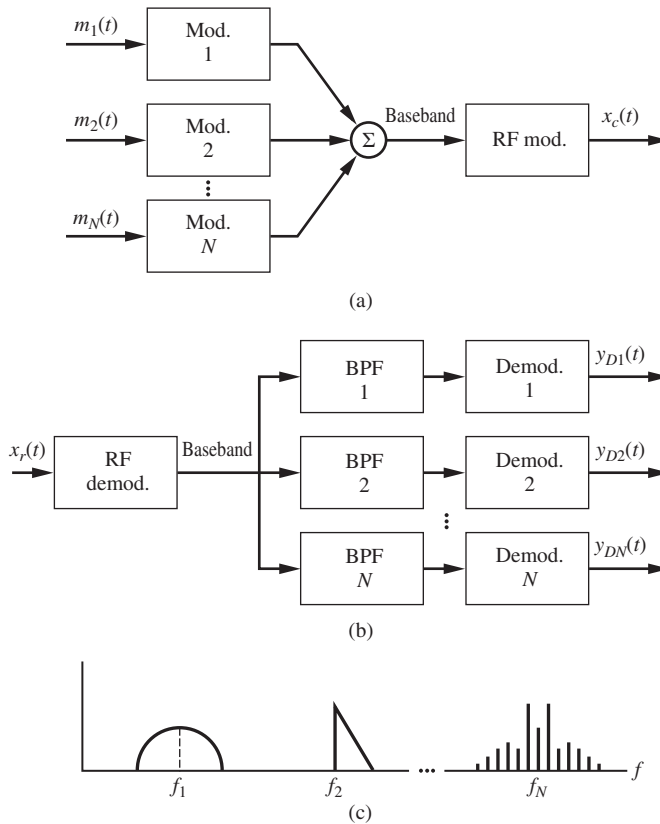


Figure 4.34
Frequency-division
multiplexing. (a) FDM
modulator. (b) FDM
demodulator. (c) Assumed
baseband spectrum.

An FDM demodulator is shown in Figure 4.34(b). The demodulator output is ideally the baseband signal. The individual channels in the baseband are extracted using bandpass filters. The bandpass filter outputs are demodulated in the conventional manner.

Observation of the baseband spectrum illustrates that the baseband bandwidth is equal to the sum of the bandwidths of the modulated signals plus the sum of the *guardbands*, the empty spectral bands between the channels necessary for filtering. This bandwidth is lower bounded by the sum of the bandwidths of the message signals. This bandwidth,

$$B = \sum_{i=1}^N W_i \quad (4.171)$$

where W_i is the bandwidth of $m_i(t)$, is achieved when all baseband modulators are SSB and all guardbands have zero width.

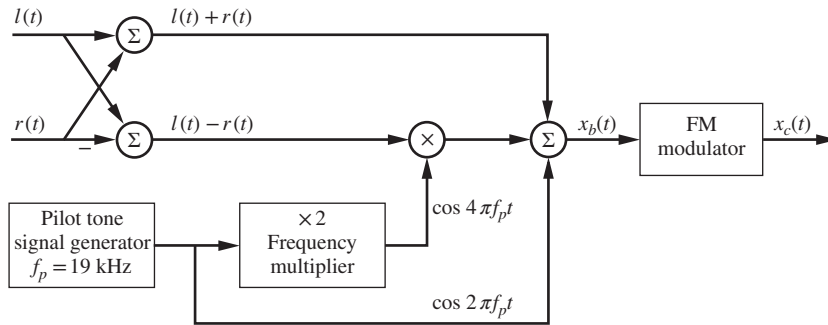
4.6.2 Example of FDM: Stereophonic FM Broadcasting

As an example of FDM, we now consider stereophonic FM broadcasting. A necessary condition established in the early development of stereophonic FM is that stereo FM be compatible with monophonic FM receivers. In other words, the output from a monophonic FM receiver must be the composite (left-channel plus right-channel) stereo signal.

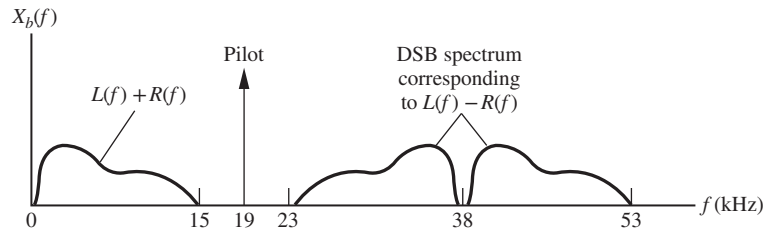
The scheme adopted for stereophonic FM broadcasting is shown in Figure 4.35(a). As can be seen, the first step in the generation of a stereo FM signal is to first form the sum and the difference of the left- and right-channel signals, $l(t) \pm r(t)$. The difference signal, $l(t) - r(t)$, is then translated to 38 kHz using DSB modulation with a carrier derived from a 19-kHz oscillator. A frequency doubler is used to generate a 38-kHz carrier from a 19-kHz oscillator. We previously saw that a PLL could be used to implement this frequency doubler.

The baseband signal is formed by adding the sum and difference signals and the 19-kHz pilot tone. The spectrum of the baseband signal is shown in Figure 4.35(b) for assumed left-channel and right-channel signals. The baseband signal is the input to the FM modulator. It is important to note that if a monophonic FM transmitter, having a message bandwidth of 15 kHz, and a stereophonic FM transmitter, having a message bandwidth of 53 kHz, both have the same constraint on the peak deviation, the deviation ratio D , of the stereophonic FM transmitter is reduced by a factor of $53/15 = 3.53$. The impact of this reduction in the deviation ratio will be seen when we consider noise effects in Chapter 8.

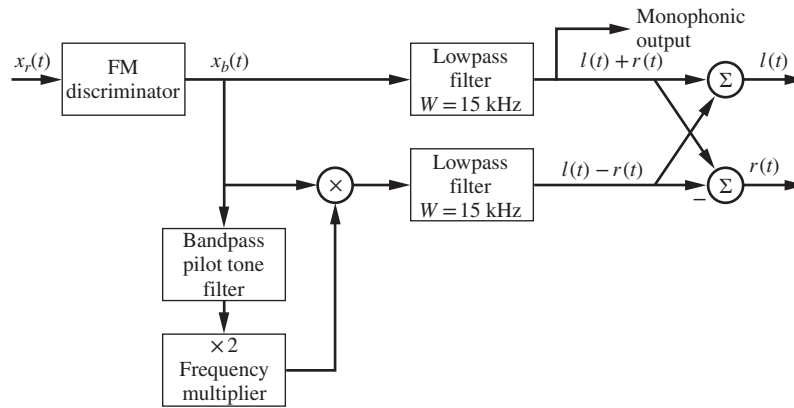
The block diagram of a stereophonic FM receiver is shown in Figure 4.35(c). The output of the FM discriminator is the baseband signal $x_b(t)$, which, under ideal conditions, is identical to the baseband signal at the input to the FM modulator. As can be seen from the spectrum of the baseband signal, the left-plus right-channel signal can be generated by filtering the baseband signal with a lowpass filter having a bandwidth of 15 kHz. Note that this signal constitutes the monophonic output. The left-minus right-channel signal is obtained by coherently demodulating the DSB signal using a 38-kHz demodulation carrier. This coherent demodulation carrier is obtained by recovering the 19-kHz pilot using a bandpass filter and then using a frequency doubler as was done in the modulator. The left-plus right-channel signal and the left-minus right-channel signal are added and subtracted, as shown in Figure 4.35(c) to generate the left-channel signal and the right-channel signal.



(a)



(b)



(c)

Figure 4.35

Stereophonic FM transmitter and receiver. (a) Stereophonic FM transmitter. (b) Single-sided spectrum of FM baseband signal. (c) Stereophonic FM receiver.

4.6.3 Quadrature Multiplexing

Another type of multiplexing is *quadrature multiplexing* (QM), in which quadrature carriers are used for frequency translation. For the system shown in Figure 4.36, the signal

$$x_c(t) = A_c[m_1(t) \cos(2\pi f_c t) + m_2(t) \sin(2\pi f_c t)] \quad (4.172)$$

is a quadrature-multiplexed signal. By sketching the spectra of $x_c(t)$ we see that these spectra overlap in frequency if the spectra of $m_1(t)$ and $m_2(t)$ overlap. Even though frequency

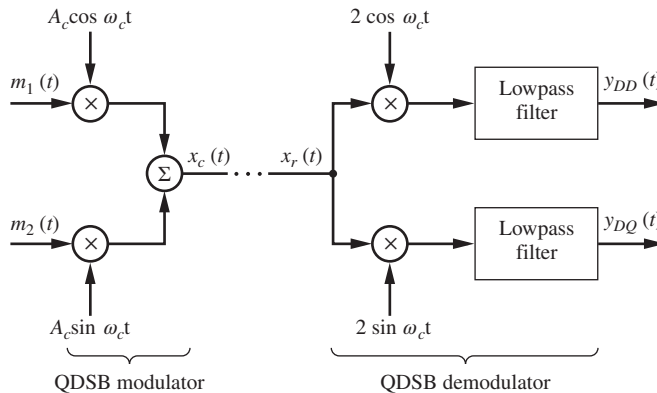


Figure 4.36
Quadrature multiplexing.

translation is used in QM, it is not an FDM technique since the two channels do not occupy disjoint spectral locations. Note that SSB is a QM signal with $m_1(t) = m(t)$ and $m_2(t) = \pm \hat{m}(t)$.

A QM signal is demodulated by using quadrature demodulation carriers. To show this, multiply $x_r(t)$ by $2 \cos(2\pi f_c t + \theta)$. This yields

$$2x_r(t) \cos(2\pi f_c t + \theta) = A_c [m_1(t) \cos \theta - m_2(t) \sin \theta] \\ + A_c [m_1(t) \cos(4\pi f_c t + \theta) + m_2(t) \sin(4\pi f_c t + \theta)] \quad (4.173)$$

The terms on the second line of the preceding equation have spectral content about $2f_c$ and can be removed by using a lowpass filter. The output of the lowpass filter is

$$y_{DD}(t) = A_c [m_1(t) \cos \theta - m_2(t) \sin \theta] \quad (4.174)$$

which yields $m_1(t)$, the desired output for $\theta = 0$. The quadrature channel is demodulated using a demodulation carrier of the form $2 \sin(2\pi f_c t)$.

The preceding result illustrates the effect of a demodulation phase error on QM. The result of this phase error is both an attenuation, which can be time varying, of the desired signal and crosstalk from the quadrature channel. It should be noted that QM can be used to represent both DSB and SSB with appropriate definitions of $m_1(t)$ and $m_2(t)$. We will take advantage of this observation when we consider the combined effect of noise and demodulation phase errors in Chapter 8.

Frequency-division multiplexing can be used with QM by translating pairs of signals, using quadrature carriers, to each subcarrier frequency. Each channel has bandwidth $2W$ and accommodates two message signals, each having bandwidth W . Thus, assuming zero-width guardbands, a baseband of bandwidth NW can accommodate N message signals, each of bandwidth W , and requires $\frac{1}{2}N$ separate subcarrier frequencies.

4.6.4 Comparison of Multiplexing Schemes

We have seen that for all three types of multiplexing studied, the baseband bandwidth is lower-bounded by the total information bandwidth. However, there are advantages and disadvantages to each multiplexing technique.

The basic advantage of FDM is simplicity of implementation, and if the channel is linear, disadvantages are difficult to identify. However, many channels have small, but nonnegligible

nonlinearities. As we saw in Chapter 2, nonlinearities lead to intermodulation distortion. In FDM systems, the result of intermodulation distortion is crosstalk between channels in the baseband.

TDM, which we discussed in the previous chapter, also has a number of inherent disadvantages. Samplers are required, and if continuous data are required by the data user, the continuous waveforms must be reconstructed from the samples. One of the biggest difficulties with TDM is maintaining synchronism between the multiplexing and demultiplexing commutators. The basic advantage of QM is that QM allows simple DSB modulation to be used while at the same time making efficient use of baseband bandwidth. It also allows DC response, which SSB does not. The basic problem with QM is crosstalk between the quadrature channels, which results if perfectly coherent demodulation carriers are not available.

Other advantages and disadvantages of FDM, QM, and TDM will become apparent when we study performance in the presence of noise in Chapter 8.

Further Reading

With the exception of the material on phase-locked loops, the references given in the previous chapter apply equally to this chapter. Once again, there are a wide variety of books available that cover this material and the books cited in Chapter 3 are only a small sample. A number of books are also available that treat the PLL. Examples are Stephens (1998), Egan (2008), Gardner (2005), and Tranter, Thamvichi, and Bose (2010). Additional material of the simulation of PLLs can be found in Tranter, Shanmugan, Rappaport, and Kosbar (2004).

Summary

1. The general expression for an angle-modulated signal is

$$x_c(t) = A_c \cos[2\pi f_c t + \phi(t)]$$

For a PM signal, $\phi(t)$ is given by

$$\phi(t) = k_p m(t)$$

and for an FM signal, it is

$$\phi(t) = 2\pi f_d \int^t m(\alpha) d\alpha$$

where k_p and f_d are the phase and frequency deviation constants, respectively.

2. Angle modulation results in an infinite number of sidebands for sinusoidal modulation. If only a single pair of sidebands is significant, the result is narrowband angle modulation. Narrowband angle modulation, with sinusoidal message, has approximately the same spectrum as an AM signal except for a 180° phase shift of the lower sideband.

3. An angle-modulated carrier with a sinusoidal message signal can be expressed as

$$x_c(t) = A_c \sum_n J_n(\beta) \cos[2\pi(f_c + n f_m)t]$$

The term $J_n(\beta)$ is the Bessel function of the first kind of order n and argument β . The parameter β is known as the *modulation index*. If $m(t) = A \sin \omega_m t$, then $\beta = k_p A$ for PM, and $\beta = f_d A / f_m$ for FM.

4. The power contained in an angle-modulated carrier is $\langle x_c^2(t) \rangle = \frac{1}{2} A_c^2$, if the carrier frequency is large compared to the bandwidth of the modulated carrier.

5. The bandwidth of an angle-modulated signal is, strictly speaking, infinite. However, a measure of the bandwidth can be obtained by defining the power ratio

$$P_r = J_0^2(\beta) + 2 \sum_{n=1}^k J_n^2(\beta)$$

which is the ratio of the total power $\frac{1}{2}A_c^2$ to the power in the bandwidth $B = 2kf_m$. A power ratio of 0.98 yields $B = 2(\beta + 1)f_m$.

6. The deviation ratio of an angle-modulated signal is

$$D = \frac{\text{peak frequency deviation}}{\text{bandwidth of } m(t)}$$

7. Carson's rule for estimating the bandwidth of an angle-modulated carrier with an arbitrary message signal is $B = 2(D + 1)W$.

8. Narrowband-to-wideband conversion is a technique whereby a wideband FM signal is generated from a narrowband FM signal. The system makes use of a frequency multiplier, which, unlike a mixer, multiplies the deviation as well as the carrier frequency.

9. Demodulation of an angle-modulated signal is accomplished through the use of a frequency discriminator. This device yields an output signal proportional to the frequency deviation of the input signal. Placing an integrator at the discriminator output allows PM signals to be demodulated.

10. An FM discriminator can be implemented as a differentiator followed by an envelope detector. Bandpass limiters are used at the differentiator input to eliminate amplitude variations.

11. A PLL is a simple and practical system for the demodulation of angle-modulated signals. It is a feedback control system and is analyzed as such. Phase-locked loops also provide simple implementations of frequency multipliers and frequency dividers.

12. The Costas PLL, which is a variation of the basic PLL, is a system for the demodulation of DSB signals.

13. *Interference*, the presence of undesired signal components, can be a problem in demodulation. Interference at the input of a demodulator results in undesired components at the demodulator output. If the interference is large and

if the demodulator is nonlinear, thresholding can occur. The result of this is a drastic loss of the signal component. In FM systems, the effect of interference is a function of both the amplitude and frequency of the interfering tone. In PM systems, the effect of interference is a function only of the amplitude of the interfering tone. In FM systems interference can be reduced by the use of pre-emphasis and de-emphasis wherein the high-frequency message components are boosted at the transmitter before modulation and the inverse process is done at the receiver after demodulation.

14. Pulse-width modulation results when the width of each carrier pulse is proportional to the value of the message signal at each sampling instant. Demodulation of PWM is also accomplished by lowpass filtering.

15. Pulse-position modulation results when the position of each carrier pulse, as measured by the displacement of each pulse from a fixed reference, is proportional to the value of the message signal at each sampling instant.

16. Multiplexing is a scheme allowing two or more message signals to be communicated simultaneously using a single system.

17. Frequency-division multiplexing results when simultaneous transmission is accomplished by translating message spectra, using modulation to *nonoverlapping* locations in a baseband spectrum. The baseband signal is then transmitted using any carrier modulation method.

18. Quadrature multiplexing results when two message signals are translated, using linear modulation with quadrature carriers, to the same spectral locations. Demodulation is accomplished coherently using quadrature demodulation carriers. A phase error in a demodulation carrier results in serious distortion of the demodulated signal. This distortion has two components: a time-varying attenuation of the desired output signal and crosstalk from the quadrature channel.

Drill Problems

4.1 Find the instantaneous phase of the angle-modulated signals assuming a phase deviation constant k_p , a carrier frequency of f_c , and the following three message signals:

(a) $m_1(t) = 10 \cos(5\pi t)$

(b) $m_2(t) = 10 \cos(5\pi t) + 2 \sin(7\pi t)$

(c) $m_3(t) = 10 \cos(5\pi t) + 2 \sin(7\pi t) + 3 \cos(6.5\pi t)$

4.2 Using the three message signals in the previous drill problem, determine the instantaneous frequency.

4.3 An FM transmitter has a frequency deviation constant of 15 Hz per unit of $m(t)$. Assuming a message signal of $m(t) = 9 \sin(40\pi t)$, write the expression for $x_c(t)$ and determine the maximum phase deviation.

4.4 Using the value of f_d and the expression for $m(t)$ given in the preceding drill problem, determine the expression for the phase deviation.

210 Chapter 4 • Angle Modulation and Multiplexing

4.5 A signal, which is treated as narrowband angle modulation has a modulation index $\beta = 0.2$. Determine the ratio of sideband power to carrier power. Describe the spectrum of the transmitted signal.

4.6 An angle-modulated signal, with sinusoidal $m(t)$ has a modulation index $\beta = 5$. Determine the ratio of sideband power to carrier power assuming that 5 sidebands are transmitted each side of the carrier.

4.7 An FM signal is formed by narrowband-to-wideband conversion. The peak frequency deviation of the narrowband signal is 40 Hz and the bandwidth of the message signal is 200 Hz. The wideband (transmitted) signal is to have a deviation ratio of 6 and a carrier frequency of 1 MHz. Determine the multiplying factor n , the carrier frequency of the narrowband signal, and, using Carson's rule, estimate the bandwidth of the wideband signal.

4.8 A first-order PLL has a total loop gain of 10. Determine the lock range.

4.9 A second-order loop filter, operating in the tracking mode, has a loop gain of 10 and a loop filter transfer function of $(s + a)/s$. Determine the value of a so that the

loop damping factor is 0.8. With this choice of a , what is the loop natural frequency?

4.10 A first-order PLL has a loop gain of 300. The input to the loop instantaneously changes frequency by 40 Hz. Determine the steady-state phase error due to this step change in frequency.

4.11 An FDM system is capable of transmitting a baseband signal having a bandwidth of 100 kHz. One channel is input to the system without modulation (about $f = 0$). Assume that all message signals have a lowpass spectrum with a bandwidth of 2 kHz and that the guardband between channels is 1 kHz. How many channels can be multiplexed together to form the baseband?

4.12 A QM system has two message signals defined by

$$m_1(t) = 5 \cos(8\pi t)$$

and

$$m_2(t) = 8 \sin(12\pi t)$$

Due to a calibration error, the demodulation carriers have a phase error of 10 degrees. Determine the two demodulated message signals.

Problems

Section 4.1

4.1 Let the input to a phase modulator be $m(t) = u(t - t_0)$, as shown in Figure 4.1(a). Assume that the unmodulated carrier is $A_c \cos(2\pi f_c t)$ and that $f_c t_0 = n$, where n is an integer. Sketch accurately the phase modulator output for $k_p = \pi$ and $-\frac{3}{8}\pi$ as was done in Figure 4.1(c) for $k_p = \frac{1}{2}\pi$.

4.2 Repeat the preceding problem for $k_p = -\frac{1}{2}\pi$ and $\frac{3}{8}\pi$.

4.3 Redraw Figure 4.4 assuming $m(t) = A \sin\left(2\pi f_m t + \frac{\pi}{6}\right)$.

4.4 We previously computed the spectrum of the FM signal defined by

$$x_{c1}(t) = A_c \cos[2\pi f_c t + \beta \sin(2\pi f_m t)]$$

Now assume that the modulated signal is given by

$$x_{c2}(t) = A_c \cos[2\pi f_c t + \beta \cos(2\pi f_m t)]$$

Show that the amplitude spectrum of $x_{c1}(t)$ and $x_{c2}(t)$ are identical. Compute the phase spectrum of $x_{c2}(t)$ and compare with the phase spectrum of $x_{c1}(t)$.

4.5 Compute the single-sided amplitude and phase spectra of

$$x_{c3}(t) = A \sin[2\pi f_c t + \beta \sin(2\pi f_m t)]$$

and

$$x_{c4}(t) = A_c \sin[2\pi f_c t + \beta \cos(2\pi f_m t)]$$

Compare the results with Figure 4.5.

4.6 The power of an unmodulated carrier signal is 50 W, and the carrier frequency is $f_c = 40$ Hz. A sinusoidal message signal is used to FM modulate it with index $\beta = 10$. The sinusoidal message signal has a frequency of 5 Hz. Determine the average value of $x_c(t)$. By drawing appropriate spectra, explain this apparent contradiction.

4.7 Given that $J_0(5) = -0.178$ and that $J_1(5) = -0.328$, determine $J_3(5)$ and $J_4(5)$.

4.8 Determine and sketch the spectrum (amplitude and phase) of an angle-modulated signal assuming that the instantaneous phase deviation is $\phi(t) = \beta \sin(2\pi f_m t)$. Also assume $\beta = 10$, $f_m = 30$ Hz, and $f_c = 2000$ Hz.

4.9 A transmitter uses a carrier frequency of 1000 Hz so that the unmodulated carrier is $A_c \cos(2\pi f_c t)$. Determine

both the phase and frequency deviation for each of the following transmitter outputs:

(a) $x_c(t) = \cos[2\pi(1000)t + 40 \sin(5t^2)]$

(b) $x_c(t) = \cos[2\pi(600)t]$

4.10 Repeat the preceding problem assuming that the transmitter outputs are defined by

(a) $x_c(t) = \cos[2\pi(1200)t^2]$

(b) $x_c(t) = \cos[2\pi(900)t + 10\sqrt{t}]$

4.11 An FM modulator has output

$$x_c(t) = 100 \cos \left[2\pi f_c t + 2\pi f_d \int^t m(\alpha) d\alpha \right]$$

where $f_d = 20$ Hz/V. Assume that $m(t)$ is the rectangular pulse $m(t) = 4\Pi \left[\frac{1}{8}(t - 4) \right]$

(a) Sketch the phase deviation in radians.

(b) Sketch the frequency deviation in hertz.

(c) Determine the peak frequency deviation in hertz.

(d) Determine the peak phase deviation in radians.

(e) Determine the power at the modulator output.

4.12 Repeat the preceding problem assuming that $m(t)$ is the triangular pulse $4\Lambda \left[\frac{1}{3}(t - 6) \right]$.

4.13 An FM modulator with $f_d = 10$ Hz/V. Plot the frequency deviation in Hz and the phase deviation in radians for the three message signals shown in Figure 4.37.

4.14 An FM modulator has $f_c = 2000$ Hz and $f_d = 20$ Hz/V. The modulator has input $m(t) = 5 \cos[2\pi(10)t]$.

(a) What is the modulation index?

(b) Sketch, approximately to scale, the magnitude spectrum of the modulator output. Show all frequencies of interest.

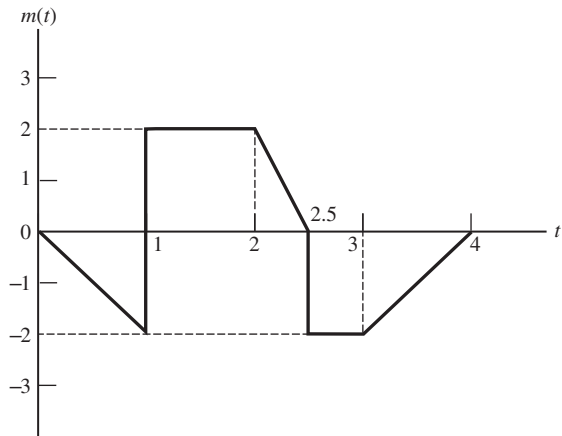
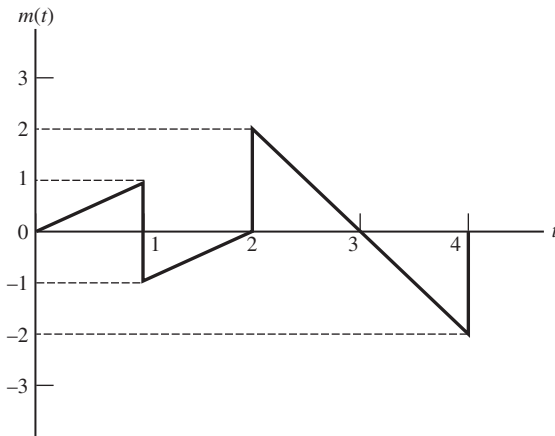
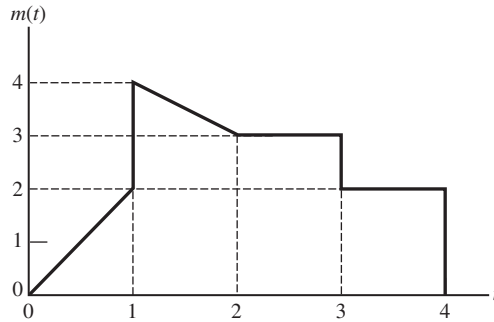


Figure 4.37

212 Chapter 4 • Angle Modulation and Multiplexing

- (c) Is this narrowband FM? Why?
 (d) If the same $m(t)$ is used for a phase modulator, what must k_p be to yield the index given in (a)?

4.15 An audio signal has a bandwidth of 15 kHz. The maximum value of $|m(t)|$ is 10 V. This signal frequency modulates a carrier. Estimate the peak deviation and the bandwidth of the modulator output, assuming that the deviation constant of the modulator is

- (a) 20 Hz/V
 (b) 200 Hz/V
 (c) 2 kHz/V
 (d) 20 kHz/V

4.16 By making use of (4.30) and (4.39), show that

$$\sum_{n=-\infty}^{\infty} J_n^2(\beta) = 1$$

4.17 Prove that $J_n(\beta)$ can be expressed as

$$J_n(\beta) = \frac{1}{\pi} \int_0^{\pi} \cos(\beta \sin x - nx) dx$$

and use this result to show that

$$J_{-n}(\beta) = (-1)^n J_n(\beta)$$

4.18 An FM modulator is followed by an ideal bandpass filter having a center frequency of 500 Hz and a bandwidth of 70 Hz. The gain of the filter is 1 in the passband. The unmodulated carrier is given by $10 \cos(1000\pi t)$, and the message signal is $m(t) = 10 \cos(20\pi t)$. The transmitter frequency-deviation constant f_d is 8 Hz/V.

- (a) Determine the peak frequency deviation in hertz.
 (b) Determine the peak phase deviation in radians.
 (c) Determine the modulation index.
 (d) Determine the power at the filter input and the filter output
 (e) Draw the single-sided spectrum of the signal at the filter input and the filter output. Label the amplitude and frequency of each spectral component.

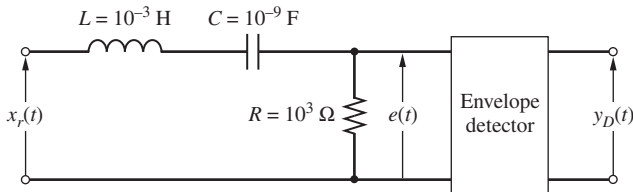


Figure 4.38

4.19 A sinusoidal message signal has a frequency of 250 Hz. This signal is the input to an FM modulator with an index of 8. Determine the bandwidth of the modulator output if a power ratio, P_r , of 0.8 is needed. Repeat for a power ratio of 0.9.

4.20 A narrowband FM signal has a carrier frequency of 110 kHz and a deviation ratio of 0.05. The modulation bandwidth is 10 kHz. This signal is used to generate a wideband FM signal with a deviation ratio of 20 and a carrier frequency of 100 MHz. The scheme utilized to accomplish this is illustrated in Figure 4.12. Give the required value of frequency multiplication, n . Also, fully define the mixer by giving two permissible frequencies for the local oscillator, and define the required bandpass filter (center frequency and bandwidth).

Section 4.2

4.21 Consider the FM discriminator shown in Figure 4.38. The envelope detector can be considered ideal with an infinite input impedance. Plot the magnitude of the transfer function $E(f)/X_r(f)$. From this plot, determine a suitable carrier frequency and the discriminator constant K_D , and estimate the allowable peak frequency deviation of the input signal.

4.22 By adjusting the values of R , L , and C in Figure 4.38, design a discriminator for a carrier frequency of 100 MHz, assuming that the peak frequency deviation is 4 MHz. What is the discriminator constant K_D for your design?

Section 4.3

4.23 Starting with (4.117) verify the steady-state errors given in Table 4.4.

4.24 Using $x_r(t) = m(t) \cos(2\pi f_c t)$ and $e_0(t) = 2 \cos(2\pi f_c t + \theta)$ for the assumed Costas PLL input and VCO output, respectively, verify that all signals shown at the various points in Figure 4.26 are correct. Assuming that the VCO frequency deviation is defined by $d\theta/dt = -K_v e_v(t)$, where $e_v(t)$ is the VCO input and K_v is a positive constant, derive the phase plane. Using the phase plane, verify that the loop locks.

4.25 Using a single PLL, design a system that has an output frequency equal to $\frac{7}{3}f_0$, where f_0 is the input frequency. Describe fully, by sketching, the output of the VCO for your design. Draw the spectrum at the VCO output and at any other point in the system necessary to explain the operation of your design. Describe any filters used in your design by defining the center frequency and the appropriate bandwidth of each.

4.26 A first-order PLL is operating with zero frequency and phase error when a step in frequency of magnitude $\Delta\omega$ is applied. The loop gain K_f is $2\pi(100)$. Determine the steady-state phase error, in degrees, for $\Delta\omega = 2\pi(30)$, $2\pi(50)$, $2\pi(80)$, and $-2\pi(80)$ rad/s. What happens if $\Delta\omega = 2\pi(120)$ rad/s?

4.27 Verify (4.120) by showing that $K_f e^{-K_f t} u(t)$ satisfies all properties of an impulse function in the limit as $K_f \rightarrow \infty$.

4.28 The imperfect second-order PLL is defined as a PLL with the loop filter

$$F(s) = \frac{s + a}{s + \lambda a}$$

in which λ is the offset of the pole from the origin relative to the zero location. In practical implementations λ is small but often cannot be neglected. Use the linear model of the PLL and derive the transfer function for $\Theta(s)/\Phi(s)$. Derive expressions for ω_n and ζ in terms of K_f , a , and λ .

4.29 Assuming the loop filter model for an imperfect second-order PLL described in the preceding problem, derive the steady-state phase errors under the three conditions of θ_0 , f_Δ , and R given in Table 4.4.

4.30 A Costas PLL operates with a small phase error so that $\sin \psi \approx \psi$ and $\cos \psi \approx 1$. Assuming that the low-pass filter preceding the VCO is modeled as $a/(s + a)$, where a is an arbitrary constant, determine the response to $m(t) = u(t - t_0)$.

4.31 In this problem we wish to develop a baseband (lowpass equivalent model) for a Costas PLL. We assume that the loop input is the complex envelope signal

$$\tilde{x}(t) = A_c m(t) e^{j\phi(t)}$$

and that the VCO output is $e^{j\theta(t)}$. Derive and sketch the model giving the signals at each point in the model.

Section 4.4

4.32 Assume that an FM demodulator operates in the presence of sinusoidal interference. Show that the discriminator output is a nonzero constant for each of the following cases: $A_i = A_c$, $A_i = -A_c$, and $A_i \gg A_c$. Determine the FM demodulator output for each of these three cases.

Section 4.5

4.33 A continuous data signal is quantized and transmitted using a PCM system. If each data sample at the receiving end of the system must be known to within $\pm 0.20\%$ of the peak-to-peak full-scale value, how many binary symbols must each transmitted digital word contain? Assume that the message signal is speech and has a bandwidth of 5 kHz. Estimate the bandwidth of the resulting PCM signal (choose k).

Section 4.6

4.34 In an FDM communication system, the transmitted baseband signal is

$$x(t) = m_1(t) \cos(2\pi f_1 t) + m_2(t) \cos(2\pi f_2 t)$$

This system has a second-order nonlinearity between transmitter output and receiver input. Thus, the received baseband signal $y(t)$ can be expressed as

$$y(t) = a_1 x(t) + a_2 x^2(t)$$

Assuming that the two message signals, $m_1(t)$ and $m_2(t)$, have the spectra

$$M_1(f) = M_2(f) = \Pi\left(\frac{f}{W}\right)$$

sketch the spectrum of $y(t)$. Discuss the difficulties encountered in demodulating the received baseband signal. In many FDM systems, the subcarrier frequencies f_1 and f_2 are harmonically related. Describe any additional problems this presents.

Computer Exercises

4.1 Reconstruct Figure 4.7 for the case in which 3 values of the modulation index (0.5, 1, and 5) are achieved by adjusting the peak frequency deviation while holding f_m constant.

4.2 Develop a computer program to generate the amplitude spectrum at the output of an FM modulator assuming a square-wave message signal. Plot the output for various values of the peak deviation. Compare the result with the

214 Chapter 4 • Angle Modulation and Multiplexing

spectrum of a PWM signal and comment on your observations.

4.3 Develop a computer program and use the program to verify the simulation results shown in Figures 4.24 and 4.25.

4.4 Referring to Computer Example 4.4, draw the block diagram of the system represented by the simulation loop, and label the inputs and outputs of the various loop components with the names used in the simulation code. Using this block diagram, verify that the simulation program is correct. What are the sources of error in the simulation program? How can these errors be mitigated?

4.5 Modify the simulation program given in Computer Example 4.4 to allow the sampling frequency to be entered interactively. Examine the effect of using different sampling frequencies by executing the simulation with a range of sampling frequencies. Be sure that you start with a sampling frequency that is clearly too low and gradually increase the sampling frequency until you reach a sampling frequency that is clearly higher than is required for an accurate simulation result. Comment on the results. How do you know that the sampling frequency is sufficiently high?

4.6 Modify the simulation program given in Computer Example 4.4 by replacing the trapezoidal integrator by a rectangular integrator. Show that for sufficiently high sampling frequencies the two PLLs give performances that are essentially equivalent. Also show that for sufficiently small sampling frequencies the two PLLs give performances that

are not equivalent. What does this tell you about selecting a sampling frequency?

4.7 Modify the simulation program given in Computer Example 4.4 so that the phase detector includes a limiter so that the phase detector characteristic is defined by

$$e_d(t) = \begin{cases} A, & \sin[\psi(t)] > A \\ \sin[\psi(t)], & -A \leq \sin[\psi(t)] \leq A \\ -A, & \sin[\psi(t)] < -A \end{cases}$$

where $\psi(t)$ is the phase error $\phi(t) - \theta(t)$ and A is a parameter that can be adjusted by the simulation user. Adjust the value of A and comment on the impact that decreasing A has on the number of cycles slipped and therefore on the time required to achieve phase lock.

4.8 Using the result of Problem 4.26, modify the simulation program given in Computer Example 4.4 so that an imperfect second-order PLL is simulated. Use the same parameter values as in Computer Example 4.4 and let $\lambda = 0.1$. Compare the time required to achieve phase lock.

4.9 A third-order PLL has the unusual property that it is unstable for small loop gain and stable for large loop gain. Use the MATLAB root-locus routine, and appropriately chosen values of a and b , demonstrate this property.

4.10 Using MATLAB, develop a program to simulate a phase-locked loop where the loop output frequency is $\frac{7}{5}f_0$, where f_0 is the input frequency. Demonstrate that the simulated system operates properly.

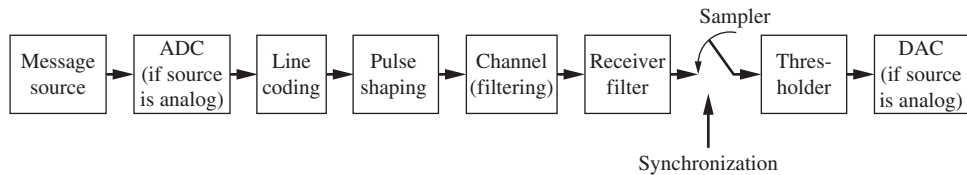
PRINCIPLES OF BASEBAND DIGITAL DATA TRANSMISSION

So far we have dealt primarily with the transmission of analog signals. In this chapter we introduce the idea of transmission of digital data—that is, signals that can assume one of only a finite number of values during each transmission interval. This may be the result of sampling and quantizing an analog signal, as in the case of pulse code modulation discussed in Chapter 4, or it might be the result of the need to transmit a message that is naturally discrete, such as a data or text file. In this chapter, we will discuss several features of a digital data transmission system. One feature that will not be covered in this chapter is the effect of random noise. This will be dealt with in Chapter 8 and following chapters. Another restriction of our discussion is that modulation onto a carrier signal is not assumed—hence, the modifier “baseband.” Thus, the types of data transmission systems to be dealt with utilize signals with power concentrated from zero hertz to a few kilohertz or megahertz, depending on the application. Digital data transmission systems that utilize bandpass signals will be considered in Chapter 9 and following.

■ 5.1 BASEBAND DIGITAL DATA TRANSMISSION SYSTEMS

Figure 5.1 shows a block diagram of a baseband digital data transmission system, which includes several possible signal processing operations. Each will be discussed in detail in future sections of the chapter. For now we give only a short description.

As already mentioned, the analog-to-digital converter (ADC) block is present only if the source produces an analog message. It can be thought of as consisting of two operations—sampling and quantization. The quantization operation can be thought of as broken up into rounding the samples to the nearest quantizing level and then converting them to a binary number representation (designated as 0s and 1s, although their actual waveform representation will be determined by the line code used, to be discussed shortly). The requirements of sampling in order to minimize errors were discussed in Chapter 2, where it was shown that, in order to avoid aliasing, the source had to be lowpass bandlimited, say to W hertz, and the sampling rate had to satisfy $f_s > 2W$ samples per second (sps). If the signal being sampled is not strictly bandlimited or if the sampling rate is less than $2W$ sps, aliasing results. Error characterization due to quantizing will be dealt with in Chapter 8. If the message is analog, necessitating the

**Figure 5.1**

Block diagram of a baseband digital data transmission system.

use of an ADC at the transmitter, the inverse operation must take place at the receiver output in order to convert the digital signal back to analog form (called digital-to-analog conversion, or DAC). As seen in Chapter 2, after converting from binary format to quantized samples, this can be as simple as a lowpass filter or, as analyzed in Problem 2.60, a zero- or higher-order hold operation can be used.

The next block, line coding, will be dealt with in the next section. It is sufficient for now to simply state that the purposes of line coding are varied, and include spectral shaping, synchronization considerations, and bandwidth considerations, among other reasons.

Pulse shaping might be used to further shape the transmitted signal spectrum in order for it to be better accommodated by the transmission channel available. In fact, we will discuss the effects of filtering and how, if inadequate attention is paid to it, severe degradation can result from transmitted pulses interfering with each other. This is termed *intersymbol interference* (ISI) and can very severely impact overall system performance if steps are not taken to counteract it. On the other hand, we will also see that careful selection of the combination of pulse shaping (transmitter filtering) and receiver filtering (it is assumed that any filtering done by the channel is not open to choice) can completely eliminate ISI.

At the output of the receiver filter, it is necessary to synchronize the sampling times to coincide with the received pulse epochs. The samples of the received pulses are then compared with a threshold in order to make a decision as to whether a 0 or a 1 was sent (depending on the line code used, this may require some additional processing). If the data transmission system is operating reliably, these 1–0 decisions are correct with high probability and the resulting DAC output is a close replica of the input message waveform.

Although the present discussion is couched in terms of two possible levels, designated as a 0 or 1, being sent it is found to be advantageous in certain situations to utilize more than two levels. If two levels are used, the data format is referred to as “binary”; if $M > 2$ levels are utilized, the data format is called “ M -ary.” If a binary format is used, the 0–1 symbols are called “bits.” If an M -ary format is used, each transmission is called a “symbol.”

■ 5.2 LINE CODES AND THEIR POWER SPECTRA

5.2.1 Description of Line Codes

The spectrum of a digitally modulated signal is influenced both by the particular baseband data format used to represent the digital data and any additional pulse shaping (filtering) used to prepare the signal for transmission. Several commonly used baseband data formats are illustrated in Figure 5.2. Names for the various data formats shown are given on the vertical axis of the respective sketch of a particular waveform, although these are not the only terms

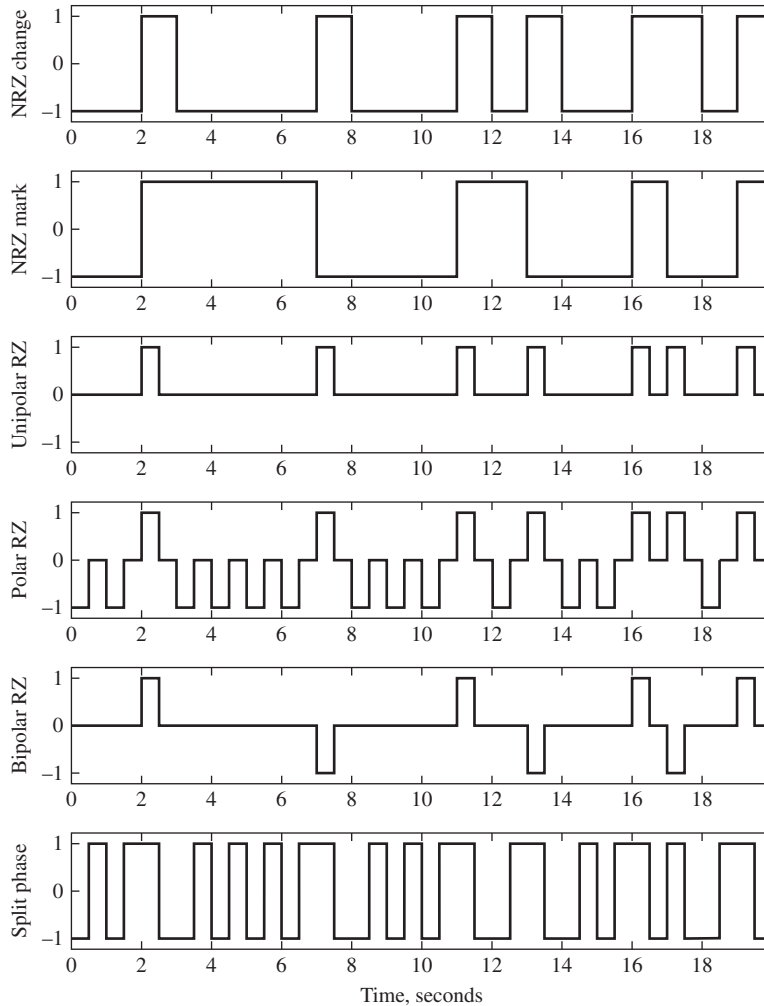


Figure 5.2
Abbreviated list of binary data formats.¹

applied to certain of these. Briefly, during each signaling interval, the following descriptions apply:

- Nonreturn-to-zero (NRZ) change (referred to as NRZ for simplicity)—a 1 is represented by a positive level, A ; a 0 is represented by $-A$
- NRZ mark—a 1 is represented by a change in level (i.e., if the previous level sent was A , $-A$ is sent to represent a 1, and vice versa); a 0 is represented by no change in level
- Unipolar return-to-zero (RZ)—a 1 is represented by a $\frac{1}{2}$ -width pulse (i.e., a pulse that “returns to zero”); a 0 is represented by no pulse

¹Adapted from J. K. Holmes, *Coherent Spread Spectrum Systems*, New York: John Wiley, 1982.

- Polar RZ—a 1 is represented by a positive RZ pulse; a 0 is represented by a negative RZ pulse
- Bipolar RZ—a 0 is represented by a 0 level; 1s are represented by RZ pulses that alternate in sign
- Split phase (Manchester)—a 1 is represented by A switching to $-A$ at $\frac{1}{2}$ the symbol period; a 0 is represented by $-A$ switching to A at $\frac{1}{2}$ the symbol period

Two of the most commonly used formats are NRZ and split phase. Split phase, we note, can be thought of as being obtained from NRZ change by multiplication by a squarewave clock waveform with a period equal to the symbol duration.

Several considerations should be taken into account in choosing an appropriate data format for a given application. Among these are:

- *Self-synchronization*—Is there sufficient timing information built into the code so that synchronizers can be easily designed to extract a timing clock from the code?
- *Power spectrum suitable for the particular channel available*—For example, if the channel does not pass low frequencies, does the power spectrum of the chosen data format have a null at zero frequency?
- *Transmission bandwidth*—If the available transmission bandwidth is scarce, which it often is, a data format should be conservative in terms of bandwidth requirements. Sometimes conflicting requirements may force difficult choices.
- *Transparency*—Every possible data sequence should be faithfully and transparently received, regardless of whether it is infrequent or not.
- *Error detection capability*—Although the subject of forward error correction deals with the design of codes to provide error correction, inherent data correction capability is an added bonus for a given data format.
- *Good bit error probability performance*—There should be nothing about a given data format that makes it difficult to implement minimum error probability receivers.

5.2.2 Power Spectra for Line-Coded Data

It is important to know the spectral occupancy of line-coded data in order to predict the bandwidth requirements for the data transmission system (conversely, given a certain system bandwidth specification, the line code used will imply a certain maximum data rate). We now consider the power spectra for line-coded data assuming that the data source produces a random coin-toss sequence of 1s and 0s, with a binary digit being produced each T seconds (recall that each binary digit is referred to as a bit, which is a contraction for “binary digit”). Since all waveforms are binary in this chapter, we use T without the subscript b for the bit period.

To compute the power spectra for line-coded data, we use a result to be derived in Chapter 7, Section 7.3.4, for the autocorrelation function of pulse-train-type signals. While it may be pedagogically unsound to use a result yet to be described, the avenue suggested to the student is to simply accept the result of Section 7.3.4 for now and concentrate on the results to be derived and the system implications of these results. In particular, a pulse-train signal

of the form

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p(t - kT - \Delta) \quad (5.1)$$

is considered in Section 7.3.4 where $\dots a_{-1}, a_0, a_1, \dots, a_k \dots$ is a sequence of random variables with the averages

$$R_m = \langle a_k a_{k+m} \rangle \quad m = 0, \pm 1, \pm 2, \dots \quad (5.2)$$

The function $p(t)$ is a deterministic pulse-type waveform, T is the separation between pulses, and Δ is a random variable that is independent of the value of a_k and uniformly distributed in the interval $(-T/2, T/2)$. It is shown that the autocorrelation function of such a waveform is

$$R_X(\tau) = \sum_{m=-\infty}^{\infty} R_m r(\tau - mT) \quad (5.3)$$

in which

$$r(\tau) = \frac{1}{T} \int_{-\infty}^{\infty} p(t + \tau) p(t) dt \quad (5.4)$$

The power spectral density is the Fourier transform of $R_X(\tau)$, which is

$$\begin{aligned} S_X(f) &= \mathfrak{F}[R_X(\tau)] = \mathfrak{F}\left[\sum_{m=-\infty}^{\infty} R_m r(\tau - mT)\right] \\ &= \sum_{m=-\infty}^{\infty} R_m \mathfrak{F}[r(\tau - mT)] \\ &= \sum_{m=-\infty}^{\infty} R_m S_r(f) e^{-j2\pi mTf} \\ &= S_r(f) \sum_{m=-\infty}^{\infty} R_m e^{-j2\pi mTf} \end{aligned} \quad (5.5)$$

where $S_r(f) = \mathfrak{F}[r(\tau)]$. Noting that $r(\tau) = \frac{1}{T} \int_{-\infty}^{\infty} p(t + \tau) p(t) dt = \left(\frac{1}{T}\right) p(-t) * p(t)$, we obtain

$$S_r(f) = \frac{|P(f)|^2}{T} \quad (5.6)$$

where $P(f) = \mathfrak{F}[p(t)]$.

EXAMPLE 5.1

In this example we apply the above result to find the power spectral density of NRZ. For NRZ, the pulse shape function is $p(t) = \Pi(t/T)$ so that

$$P(f) = T \operatorname{sinc}(Tf) \quad (5.7)$$

and

$$S_r(f) = \frac{1}{T} |T \operatorname{sinc}(Tf)|^2 = T \operatorname{sinc}^2(Tf) \quad (5.8)$$

The time average $R_m = \langle a_k a_{k+m} \rangle$ can be deduced by noting that, for a given pulse, the amplitude is $+A$ half the time and $-A$ half the time, while, for a sequence of two pulses with a given sign on the first pulse, the second pulse is $+A$ half the time and $-A$ half the time. Thus,

$$R_m = \begin{cases} \frac{1}{2}A^2 + \frac{1}{2}(-A)^2 = A^2, & m = 0 \\ \frac{1}{4}A(A) + \frac{1}{4}A(-A) + \frac{1}{4}(-A)A + \frac{1}{4}(-A)(-A) = 0, & m \neq 0 \end{cases} \quad (5.9)$$

Thus, using (5.8) and (5.9) in (5.5), the power spectral density for NRZ is

$$S_{\text{NRZ}}(f) = A^2 T \operatorname{sinc}^2(Tf) \quad (5.10)$$

This is plotted in Figure 5.3(a) where it is seen that the bandwidth to the first null of the power spectral density is $B_{\text{NRZ}} = 1/T$ hertz. Note that $A = 1$ gives unit power as seen from squaring and averaging the time-domain waveform. ■

EXAMPLE 5.2

The computation of the power spectral density for split phase differs from that for NRZ only in the spectrum of the pulse-shape function because the coefficients R_m are the same as for NRZ. The pulse-shape function for split phase is given by

$$p(t) = \Pi\left(\frac{t+T/4}{T/2}\right) - \Pi\left(\frac{t-T/4}{T/2}\right) \quad (5.11)$$

By applying the time delay and superposition theorems of Fourier transforms, we have

$$\begin{aligned} P(f) &= \frac{T}{2} \operatorname{sinc}\left(\frac{T}{2}f\right) e^{j2\pi(T/4)f} - \frac{T}{2} \operatorname{sinc}\left(\frac{T}{2}f\right) e^{-j2\pi(T/4)f} \\ &= \frac{T}{2} \operatorname{sinc}\left(\frac{T}{2}f\right) (e^{j\pi Tf/2} - e^{-j\pi Tf/2}) \\ &= jT \operatorname{sinc}\left(\frac{T}{2}f\right) \sin\left(\frac{\pi T}{2}f\right) \end{aligned} \quad (5.12)$$

Thus,

$$\begin{aligned} S_r(f) &= \frac{1}{T} \left| jT \operatorname{sinc}\left(\frac{T}{2}f\right) \sin\left(\frac{\pi T}{2}f\right) \right|^2 \\ &= T \operatorname{sinc}^2\left(\frac{T}{2}f\right) \sin^2\left(\frac{\pi T}{2}f\right) \end{aligned} \quad (5.13)$$

Hence, for split phase the power spectral density is

$$S_{\text{SP}}(f) = A^2 T \operatorname{sinc}^2\left(\frac{T}{2}f\right) \sin^2\left(\frac{\pi T}{2}f\right) \quad (5.14)$$

This is plotted in Figure 5.3(b) where it is seen that the bandwidth to the first null of the power spectral density is $B_{\text{SP}} = 2/T$ hertz. However, unlike NRZ, split phase has a null at $f = 0$, which might have favorable implications if the transmission channel does not pass DC. Note that by squaring the time waveform and averaging the result, it is evident that $A = 1$ gives unit power. ■

EXAMPLE 5.3

In this example, we compute the power spectrum of unipolar RZ, which provides the additional challenge of discrete spectral lines. For unipolar RZ, the data correlation coefficients are

$$R_m = \begin{cases} \frac{1}{2}A^2 + \frac{1}{2}(0)^2 = \frac{1}{2}A^2, & m = 0 \\ \frac{1}{4}(A)(A) + \frac{1}{4}(A)(0) + \frac{1}{4}(0)(A) + \frac{1}{4}(0)(0) = \frac{1}{4}A^2, & m \neq 0 \end{cases} \quad (5.15)$$

The pulse-shape function is given by

$$p(t) = \Pi(2t/T) \quad (5.16)$$

Therefore, we have

$$P(f) = \frac{T}{2} \operatorname{sinc}\left(\frac{T}{2}f\right) \quad (5.17)$$

and

$$\begin{aligned} S_r(f) &= \frac{1}{T} \left| \frac{T}{2} \operatorname{sinc}\left(\frac{T}{2}f\right) \right|^2 \\ &= \frac{T}{4} \operatorname{sinc}^2\left(\frac{T}{2}f\right) \end{aligned} \quad (5.18)$$

For unipolar RZ, we therefore have

$$\begin{aligned} S_{\text{URZ}}(f) &= \frac{T}{4} \operatorname{sinc}^2\left(\frac{T}{2}f\right) \left[\frac{1}{2}A^2 + \frac{1}{4}A^2 \sum_{m=-\infty, m \neq 0}^{\infty} e^{-j2\pi m T f} \right] \\ &= \frac{T}{4} \operatorname{sinc}^2\left(\frac{T}{2}f\right) \left[\frac{1}{4}A^2 + \frac{1}{4}A^2 \sum_{m=-\infty}^{\infty} e^{-j2\pi m T f} \right] \end{aligned} \quad (5.19)$$

where $\frac{1}{2}A^2$ has been split between the initial term inside the brackets and the summation (which supplies the term for $m = 0$ in the summation). But from (2.121) we have

$$\sum_{m=-\infty}^{\infty} e^{-j2\pi m T f} = \sum_{m=-\infty}^{\infty} e^{j2\pi m T f} = \frac{1}{T} \sum_{n=-\infty}^{\infty} \delta(f - n/T) \quad (5.20)$$

Thus, $S_{\text{URZ}}(f)$ can be written as

$$\begin{aligned} S_{\text{URZ}}(f) &= \frac{T}{4} \operatorname{sinc}^2\left(\frac{T}{2}f\right) \left[\frac{1}{4}A^2 + \frac{1}{4} \frac{A^2}{T} \sum_{n=-\infty}^{\infty} \delta(f - n/T) \right] \\ &= \frac{A^2 T}{16} \operatorname{sinc}^2\left(\frac{T}{2}f\right) + \frac{A^2}{16} \delta(f) + \frac{A^2}{16} \operatorname{sinc}^2\left(\frac{1}{2}\right) \left[\delta\left(f - \frac{1}{T}\right) + \delta\left(f + \frac{1}{T}\right) \right] \\ &\quad + \frac{A^2}{16} \operatorname{sinc}^2\left(\frac{3}{2}\right) \left[\delta\left(f - \frac{3}{T}\right) + \delta\left(f + \frac{3}{T}\right) \right] + \dots \end{aligned} \quad (5.21)$$

where the fact that $Y(f) \delta(f - f_n) = Y(f_n) \delta(f - f_n)$ for $Y(f)$ continuous at $f = f_n$ has been used to simplify the $\operatorname{sinc}^2\left(\frac{T}{2}f\right) \delta(f - n/T)$ terms. [Note that $\operatorname{sinc}^2\left(\frac{n}{2}\right) = 0$ for n even.]

The power spectrum of unipolar RZ is plotted in Figure 5.3(c) where it is seen that the bandwidth to the first null of the power spectral density is $B_{URZ} = 2/T$ hertz. The reason for the impulses in the spectrum is because the unipolar nature of this waveform is reflected in finite power at DC and harmonics of $1/T$ hertz. This can be a useful feature for synchronization purposes.

Note that for unit power in unipolar RZ, $A = 2$ because the average of the time-domain waveform squared is $\frac{1}{T} \left[\frac{1}{2} \left(A^2 \frac{T}{2} + 0^2 \frac{T}{2} \right) + \frac{1}{2} 0^2 T \right] = \frac{A^2}{4}$. ■

EXAMPLE 5.4

The power spectral density of polar RZ is straightforward to compute based on the results for NRZ. The data correlation coefficients are the same as for NRZ. The pulse-shape function is $p(t) = \Pi(2t/T)$, the same as for unipolar RZ, so $S_r(f) = \frac{T}{4} \text{sinc}^2\left(\frac{T}{2}f\right)$. Thus,

$$S_{PRZ}(f) = \frac{A^2 T}{4} \text{sinc}^2\left(\frac{T}{2}f\right) \quad (5.22)$$

The power spectrum of polar RZ is plotted in Figure 5.3(d) where it is seen that the bandwidth to the first null of the power spectral density is $B_{PRZ} = 2/T$ hertz. Unlike polar RZ, there are no discrete spectral lines. Note that by squaring and averaging the time-domain waveform, we get $\frac{1}{T} \left(A^2 \frac{T}{2} + 0^2 \frac{T}{2} \right) = \frac{A^2}{2}$, so $A = \sqrt{2}$ for unit average power. ■

EXAMPLE 5.5

The final line code for which we will compute the power spectrum is bipolar RZ. For $m = 0$, the possible $a_k a_k$ products are $AA = (-A)(-A) = A^2$ —each of which occurs $\frac{1}{4}$ the time and $(0)(0) = 0$ which occurs $\frac{1}{2}$ the time. For $m = \pm 1$, the possible data sequences are $(1, 1)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$ for which the possible $a_k a_{k+1}$ products are $-A^2$, 0 , 0 , and 0 , respectively, each of which occurs with probability $\frac{1}{4}$. For $m > 1$ the possible products are A^2 and $-A^2$, each of which occurs with probability $\frac{1}{8}$, and $\pm A(0)$, and $(0)(0)$, each of which occur with probability $\frac{1}{4}$. Thus, the data correlation coefficients become

$$R_m = \begin{cases} \frac{1}{4}A^2 + \frac{1}{4}(-A)^2 + \frac{1}{2}(0)^2 = \frac{1}{2}A^2, & m = 0 \\ \frac{1}{4}(-A)^2 + \frac{1}{4}(A)(0) + \frac{1}{4}(0)(A) + \frac{1}{4}(0)(0) = -\frac{A^2}{4}, & m = \pm 1 \\ \frac{1}{8}A^2 + \frac{1}{8}(-A^2) + \frac{1}{4}(A)(0) + \frac{1}{4}(-A)(0) + \frac{1}{4}(0)(0) = 0, & |m| > 1 \end{cases} \quad (5.23)$$

The pulse-shape function is

$$p(t) = \Pi(2t/T) \quad (5.24)$$

Therefore, we have

$$P(f) = \frac{T}{2} \text{sinc}\left(\frac{T}{2}f\right) \quad (5.25)$$

and

$$\begin{aligned} S_r(f) &= \frac{1}{T} \left| \frac{T}{2} \text{sinc}\left(\frac{T}{2}f\right) \right|^2 \\ &= \frac{T}{4} \text{sinc}^2\left(\frac{T}{2}f\right) \end{aligned} \quad (5.26)$$

Therefore, for bipolar RZ we have

$$\begin{aligned}
 S_{\text{BPRZ}}(f) &= S_r(f) \sum_{m=-\infty}^{\infty} R_m e^{-j2\pi m T f} \\
 &= \frac{A^2 T}{8} \text{sinc}^2\left(\frac{T}{2} f\right) \left(1 - \frac{1}{2} e^{j2\pi T f} - \frac{1}{2} e^{-j2\pi T f}\right) \\
 &= \frac{A^2 T}{8} \text{sinc}^2\left(\frac{T}{2} f\right) [1 - \cos(2\pi T f)] \\
 &= \frac{A^2 T}{4} \text{sinc}^2\left(\frac{T}{2} f\right) \sin^2(\pi T f)
 \end{aligned} \tag{5.27}$$

which is shown in Figure 5.3(e).

Note that by squaring the time-domain waveform and accounting for it being 0 for the time when logic 0s are sent and it being 0 half the time when logic 1s are sent, we get for the power

$$\frac{1}{T} \left[\frac{1}{2} \left(\frac{1}{2} A^2 \frac{T}{2} + \frac{1}{2} (-A)^2 \frac{T}{2} + 0^2 \frac{T}{2} \right) + \frac{1}{2} 0^2 T \right] = \frac{A^2}{4} \tag{5.28}$$

so $A = 2$ for unit average power. ■

Typical power spectra are shown in Figure 5.3 for all of the data modulation formats shown in Figure 5.2, assuming a random (coin toss) bit sequence. For data formats lacking power spectra with significant frequency content at multiples of the bit rate, $1/T$, nonlinear operations are required to generate power at a frequency of $1/T$ Hz or multiples thereof for symbol synchronization purposes. Note that split phase guarantees at least one zero crossing per bit interval, but requires twice the transmission bandwidth of NRZ. Around 0 Hz, NRZ possesses significant power. Generally, no data format possesses all the desired features listed in Section 5.2.1, and the choice of a particular data format will involve trade-offs.

COMPUTER EXAMPLE 5.1

A MATLAB script file for plotting the power spectra of Figure 5.3 is given below.

```

% File: c5ce1.m
%
clf
ANRZ = 1;
T = 1;
f = -40:.005:40;
SNRZ = ANRZ^2*T*(sinc(T*f)).^2;
areaNRZ = trapz(f, SNRZ) % Area of NRZ spectrum as check
ASP = 1;
SSP = ASP^2*T*(sinc(T*f/2)).^2.*(sin(pi*T*f/2)).^2;
areaSP = trapz(f, SSP) % Area of split-phase spectrum as check
AURZ = 2;
SURZc = AURZ^2*T/16*(sinc(T*f/2)).^2;
areaRZc = trapz(f, SURZc)
fdisc = -40:1:40;
SURZd = zeros(size(fdisc));
SURZd = AURZ^2/16*(sinc(fdisc/2)).^2;
areaRZ = sum(SURZd)+areaRZc % Area of unipolar return-to-zero spect as
check

```

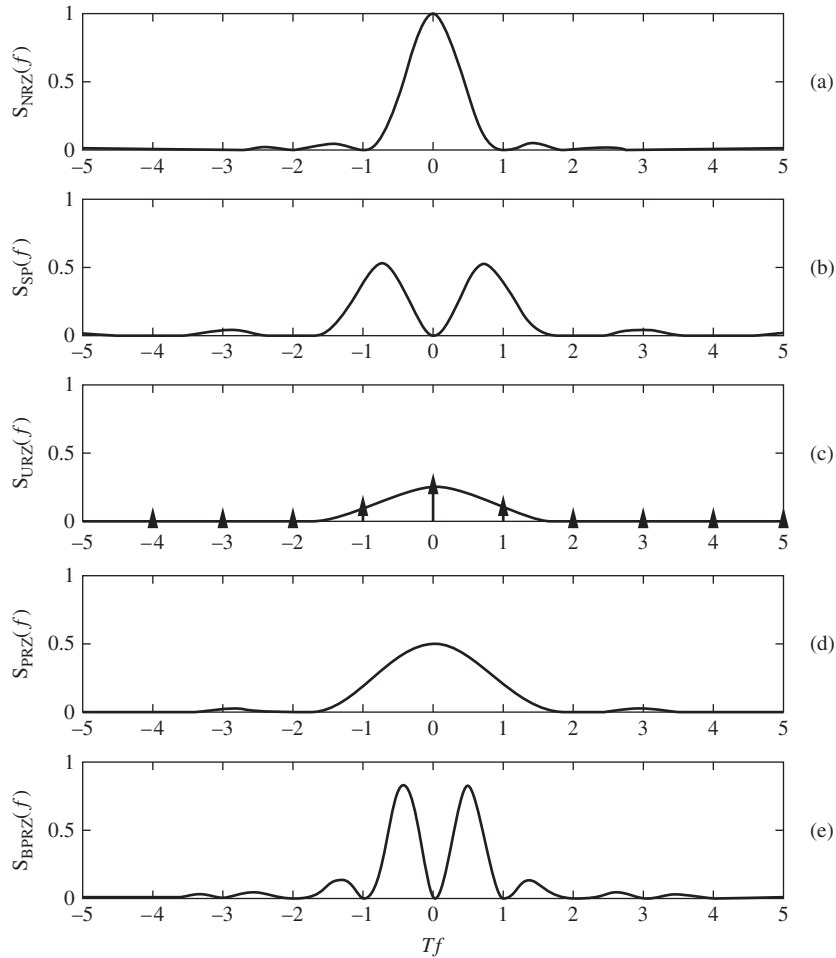


Figure 5.3
Power spectra for line-coded binary data formats.

```

APRZ = sqrt(2);
SPRZ = APRZ^2*T/4*(sinc(T*f/2)).^2;
areaSPRZ = trapz(f, SPRZ) % Area of polar return-to-zero spectrum as
check
ABPRZ = 2;
SBPRZ = ABPRZ^2*T/4*((sinc(T*f/2)).^2).*(sin(pi*T*f)).^2;
areaBPRZ = trapz(f, SBPRZ) % Area of bipolar return-to-zero spectrum
as check
subplot(5,1,1), plot(f, SNRZ), axis([-5, 5, 0, 1]), ylabel('S.N.R.Z(f)')
subplot(5,1,2), plot(f, SSP), axis([-5, 5, 0, 1]), ylabel('S.S.P(f)')
subplot(5,1,3), plot(f, SURZc), axis([-5, 5, 0, 1]), yla-
bel('S.U.R.Z(f)')
hold on
subplot(5,1,3), stem(fdisc, SURZd, ' '), axis([-5, 5, 0, 1])
subplot(5,1,4), plot(f, SPRZ), axis([-5, 5, 0, 1]), ylabel('S.P.R.Z(f)')

```

```
subplot(5,1,5), plot(f, SBPRZ), axis([-5, 5, 0, 1]),
xlabel('Tf'), ylabel('S.B.P.R.Z(f)')
% End of script file
```

5.3 EFFECTS OF FILTERING OF DIGITAL DATA—ISI

One source of degradation in a digital data transmission system has already been mentioned and termed *intersymbol interference*, or ISI. ISI results when a sequence of signal pulses is passed through a channel with a bandwidth insufficient to pass the significant spectral components of the signal. Example 2.20 illustrated the response of a lowpass RC filter to a rectangular pulse. For an input of

$$x_1(t) = A\Pi\left(\frac{t-T/2}{T}\right) = A[u(t) - u(t-T)] \quad (5.29)$$

the output of the filter was found to be

$$y_1(t) = A\left[1 - \exp\left(-\frac{t}{RC}\right)\right]u(t) - A\left[1 - \exp\left(-\frac{t-T}{RC}\right)\right]u(t-T) \quad (5.30)$$

This is plotted in Figure 2.16(a), which shows that the output is more “smeared out” the smaller T/RC is [although not in exactly the same form as (2.182), they are in fact equivalent]. In fact, by superposition, a sequence of two pulses of the form

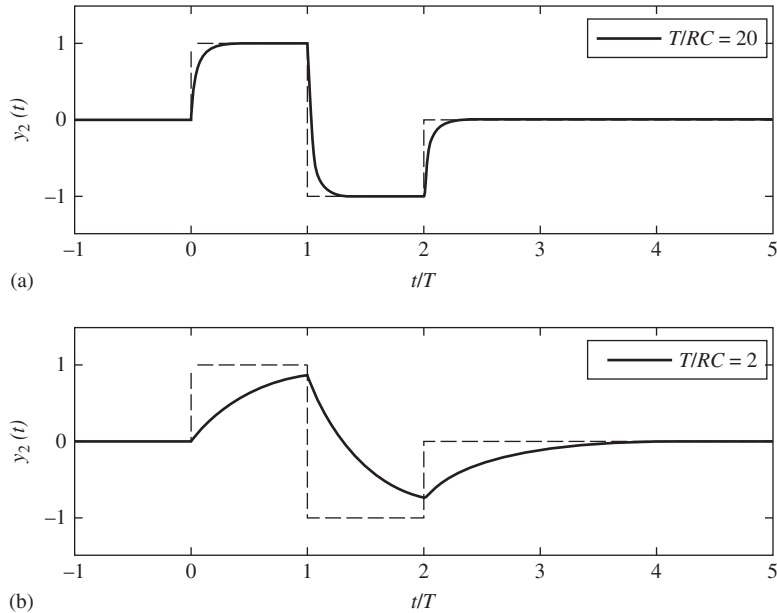
$$\begin{aligned} x_2(t) &= A\Pi\left(\frac{t-T/2}{T}\right) - A\Pi\left(\frac{t-3T/2}{T}\right) \\ &= A[u(t) - 2u(t-T) + u(t-2T)] \end{aligned} \quad (5.31)$$

will result in the response

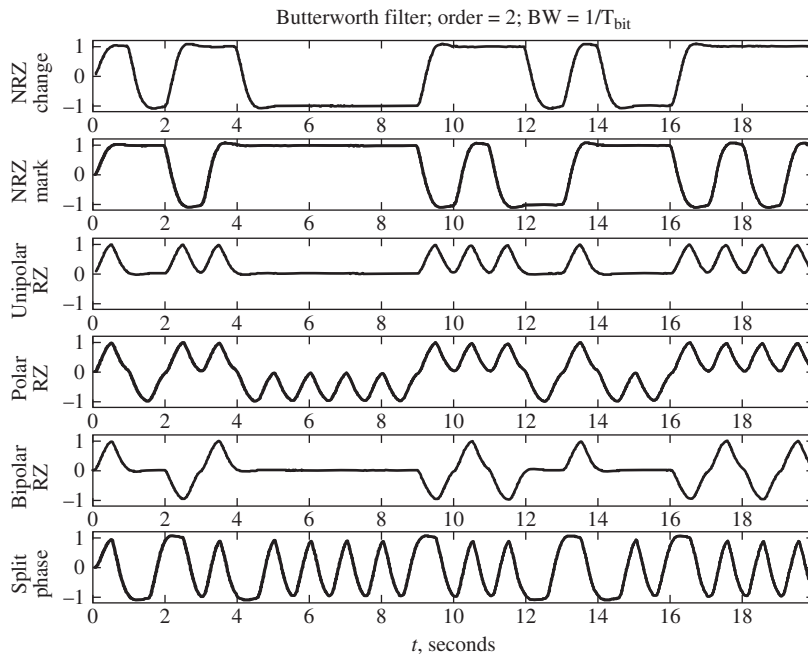
$$\begin{aligned} y_2(t) &= A\left[1 - \exp\left(-\frac{t}{RC}\right)\right]u(t) - 2A\left[1 - \exp\left(-\frac{t-T}{RC}\right)\right]u(t-T) \\ &\quad + A\left[1 - \exp\left(-\frac{t-2T}{RC}\right)\right]u(t-2T) \end{aligned} \quad (5.32)$$

At a simple level, this illustrates the idea of ISI. If the channel, represented by the lowpass RC filter, has only a single pulse at its input, there is no problem from the transient response of the channel. However, when two or more pulses are input to the channel in time sequence [in the case of the input $x_2(t)$, a positive pulse followed by a negative one], the transient response due to the initial pulse interferes with the responses due to the trailing pulses. This is illustrated in Figure 5.4 where the two-pulse response (5.32) is plotted for two values of T/RC , the first of which results in negligible ISI and the second of which results in significant ISI in addition to distortion of the output pulses. In fact, the smaller T/RC , the more severe the ISI effects are because the time constant, RC , of the filter is large compared with the pulse width, T .

To consider a more realistic example, we reconsider the line codes of Figure 5.2. These waveforms are shown filtered by a lowpass, second-order Butterworth filter in Figure 5.5 for the filter 3-dB frequency equal to $f_3 = 1/T_{\text{bit}} = 1/T$ and in Figure 5.6 for $f_3 = 0.5/T$.

**Figure 5.4**

Response of a lowpass RC filter to a positive rectangular pulse followed by a negative rectangular pulse to illustrate the concept of ISI: (a) $T/RC = 20$; (b) $T/RC = 2$.

**Figure 5.5**

Data sequences formatted with various line codes passed through a channel represented by a second-order lowpass Butterworth filter of bandwidth one bit rate.

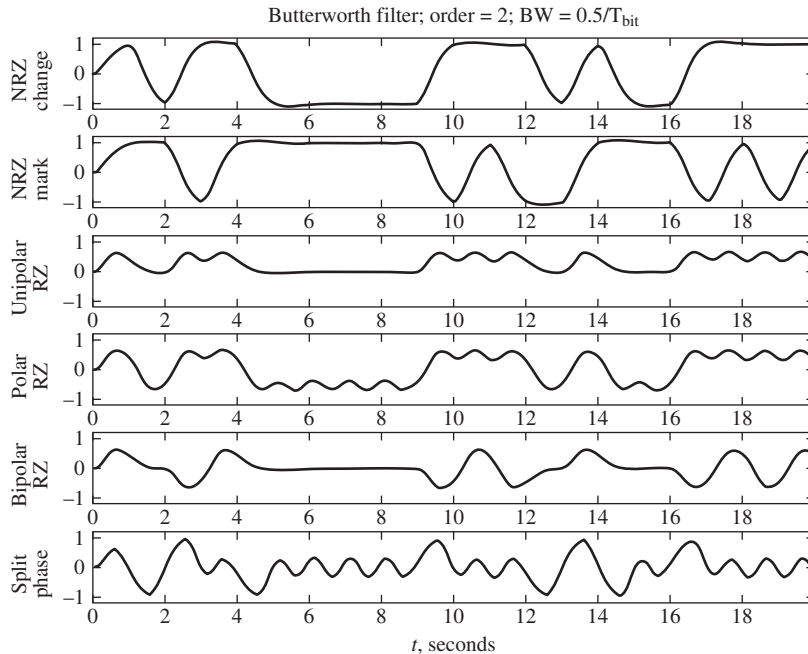


Figure 5.6

Data sequences formatted with various line codes passed through a channel represented by a second-order lowpass Butterworth filter of bandwidth one-half bit rate.

The effects of ISI are evident. In Figure 5.5 the bits are fairly discernible, even for data formats using pulses of width $T/2$ (i.e., all the RZ cases and split phase). In Figure 5.6, the NRZ cases have fairly distinguishable bits, but the RZ and split-phase formats suffer greatly from ISI. Recall that from the plots of Figure 5.3 and the analysis that led to them, the RZ and split-phase formats occupy essentially twice the bandwidth of the NRZ formats for a given data rate.

The question about what can be done about ISI naturally arises. One perhaps surprising solution is that with proper transmitter and receiver filter design (the filter representing the channel is whatever it is) the effects of ISI can be completely eliminated. We investigate this solution in the following section. Another somewhat related solution is the use of special filtering at the receiver called equalization. At a very rudimentary level, an equalization filter can be looked at as having the inverse of the channel filter frequency response, or a close approximation to it. We consider one form of equalization filtering in Section 5.5.

■ 5.4 PULSE SHAPING: NYQUIST'S CRITERION FOR ZERO ISI

In this section we examine designs for the transmitter and receiver filters that shape the overall signal pulse-shape function so as to ideally eliminate interference between adjacent pulses. This is formally stated as Nyquist's criterion for zero ISI.

5.4.1 Pulses Having the Zero ISI Property

To see how one might implement this approach, we recall the sampling theorem, which gives a theoretical maximum spacing between samples to be taken from a signal with an ideal lowpass spectrum in order that the signal can be reconstructed exactly from the sample values. In particular, the transmission of a lowpass signal with bandwidth W hertz can be viewed as sending a minimum of $2W$ independent sps. If these $2W$ sps represent $2W$ independent pieces of data, this transmission can be viewed as sending $2W$ pulses per second through a channel represented by an ideal lowpass filter of bandwidth W . The transmission of the n th piece of information through the channel at time $t = nT = n/(2W)$ is accomplished by sending an impulse of amplitude a_n . The output of the channel due to this impulse at the input is

$$y_n(t) = a_n \operatorname{sinc} \left[2W \left(t - \frac{n}{2W} \right) \right] \quad (5.33)$$

For an input consisting of a train of impulses spaced by $T = 1/(2W)$ s, the channel output is

$$y(t) = \sum_n y_n(t) = \sum_n a_n \operatorname{sinc} \left[2W \left(t - \frac{n}{2W} \right) \right] \quad (5.34)$$

where $\{a_n\}$ is the sequence of sample values (i.e., the information). If the channel output is sampled at time $t_m = m/2W$, the sample value is a_m because

$$\operatorname{sinc}(m - n) = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases} \quad (5.35)$$

which results in all terms in (5.34) except the m th being zero. In other words, the m th sample value at the output is not affected by preceding or succeeding sample values; it represents an independent piece of information.

Note that the bandlimited channel implies that the time response due to the n th impulse at the input is infinite in extent; a waveform cannot be simultaneously bandlimited and time-limited. It is of interest to inquire if there are any bandlimited waveforms other than $\operatorname{sinc}(2Wt)$ that have the property of (5.35), that is, that their zero crossings are spaced by $T = 1/(2W)$ seconds. One such family of pulses are those having raised cosine spectra. Their time response is given by

$$p_{\text{RC}}(t) = \frac{\cos(\pi\beta t/T)}{1 - (2\beta t/T)^2} \operatorname{sinc} \left(\frac{t}{T} \right) \quad (5.36)$$

and their spectra by

$$P_{\text{RC}}(f) = \begin{cases} T, & |f| \leq \frac{1-\beta}{2T} \\ \frac{T}{2} \left\{ 1 + \cos \left[\frac{\pi T}{\beta} \left(|f| - \frac{1-\beta}{2T} \right) \right] \right\}, & \frac{1-\beta}{2T} < |f| \leq \frac{1+\beta}{2T} \\ 0, & |f| > \frac{1+\beta}{2T} \end{cases} \quad (5.37)$$

where β is called the roll-off factor. Figure 5.7 shows this family of spectra and the corresponding pulse responses for several values of β . Note that zero crossings for $p_{\text{RC}}(t)$ occur at least every T seconds. If $\beta = 1$, the single-sided bandwidth of $P_{\text{RC}}(f)$ is $\frac{1}{T}$ hertz (just substitute $\beta = 1$ into (5.37)), which is twice that for the case of $\beta = 0$ [$\operatorname{sinc}(t/T)$ pulse]. The price paid for the raised cosine roll-off with increasing frequency of $P_{\text{RC}}(f)$, which may be easier to realize as practical filters in the transmitter and receiver, is increased bandwidth.

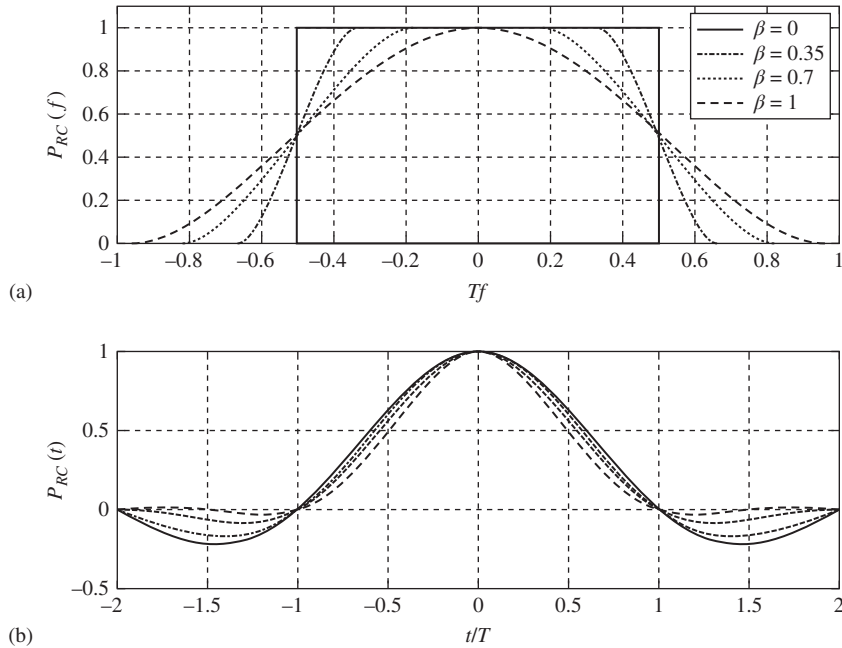


Figure 5.7
 (a) Raised cosine spectra and (b) corresponding pulse responses.

Also, $p_{RC}(t)$ for $\beta = 1$ has a narrow main lobe with very low side lobes. This is advantageous in that interference with neighboring pulses is minimized if the sampling instants are slightly in error. Pulses with raised cosine spectra are used extensively in the design of digital communication systems.

5.4.2 Nyquist's Pulse-Shaping Criterion

Nyquist's pulse-shaping criterion states that a pulse-shape function $p(t)$, having a Fourier transform $P(f)$ that satisfies the criterion

$$\sum_{k=-\infty}^{\infty} P\left(f + \frac{k}{T}\right) = T, \quad |f| \leq \frac{1}{2T} \quad (5.38)$$

results in a pulse-shape function with sample values

$$p(nT) = \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \quad (5.39)$$

Using this result, we can see that no adjacent pulse interference will result if the received data stream is represented as

$$y(t) = \sum_{n=-\infty}^{\infty} a_n p(t - nT) \quad (5.40)$$

and the sampling at the receiver occurs at integer multiples of T seconds at the pulse epochs. For example, to obtain the $n = 10$ th sample, one simply sets $t = 10T$ in (5.40), and the resulting sample is a_{10} , given that the result of Nyquist's pulse-shaping criterion of (5.39) holds.

The proof of Nyquist's pulse-shaping criterion follows easily by making use of the inverse Fourier representation for $p(t)$, which is

$$p(t) = \int_{-\infty}^{\infty} P(f) \exp(j2\pi ft) df \quad (5.41)$$

For the n th sample value, this expression can be written as

$$p(nT) = \sum_{k=-\infty}^{\infty} \int_{-(2k+1)/2T}^{(2k+1)/2T} P(f) \exp(j2\pi fnT) df \quad (5.42)$$

where the inverse Fourier transform integral for $p(t)$ has been broken up into contiguous frequency intervals of length $1/T$ Hz. By the change of variables $u = f - k/T$, (5.42) becomes

$$\begin{aligned} p(nT) &= \sum_{k=-\infty}^{\infty} \int_{-1/2T}^{1/2T} P\left(u + \frac{k}{T}\right) \exp(j2\pi nTu) du \\ &= \int_{-1/2T}^{1/2T} \sum_{k=-\infty}^{\infty} P\left(u + \frac{k}{T}\right) \exp(j2\pi nTu) du \end{aligned} \quad (5.43)$$

where the order of integration and summation has been reversed. By hypothesis

$$\sum_{k=-\infty}^{\infty} P(u + k/T) = T \quad (5.44)$$

between the limits of integration, so that (5.43) becomes

$$\begin{aligned} p(nT) &= \int_{-1/2T}^{1/2T} T \exp(j2\pi nTu) du = \text{sinc}(n) \\ &= \begin{cases} 1, & n = 0 \\ 0, & n \neq 0 \end{cases} \end{aligned} \quad (5.45)$$

which completes the proof of Nyquist's pulse-shaping criterion.

With the aid of this result, it is now apparent why the raised-cosine pulse family is free of intersymbol interference, even though the family is by no means unique. Note that what is excluded from the raised-cosine spectrum for $|f| < \frac{1}{T}$ hertz is filled by the spectral translate tail for $|f| > \frac{1}{T}$ hertz. Example 5.6 illustrates this for a simpler, although more impractical, spectrum than the raised-cosine spectrum.

EXAMPLE 5.6

Consider the triangular spectrum

$$P_{\Delta}(f) = T \Lambda(Tf) \quad (5.46)$$

It is shown in Figure 5.8(a) and in Figure 5.8(b) $\sum_{k=-\infty}^{\infty} P_{\Delta}\left(f + \frac{k}{T}\right)$ is shown where it is evident that the sum is a constant. Using the transform pair $\Lambda(t/B) \leftrightarrow B \operatorname{sinc}^2(Bf)$ and duality to get the transform pair $p_{\Delta}(t) = \operatorname{sinc}^2(t/T) \leftrightarrow T\Lambda(Tf) = P_{\Delta}(f)$, we see that this pulse-shape function does indeed have the zero-ISI property because $p_{\Delta}(nT) = \operatorname{sinc}^2(n) = 0$, $n \neq 0$, n integer.

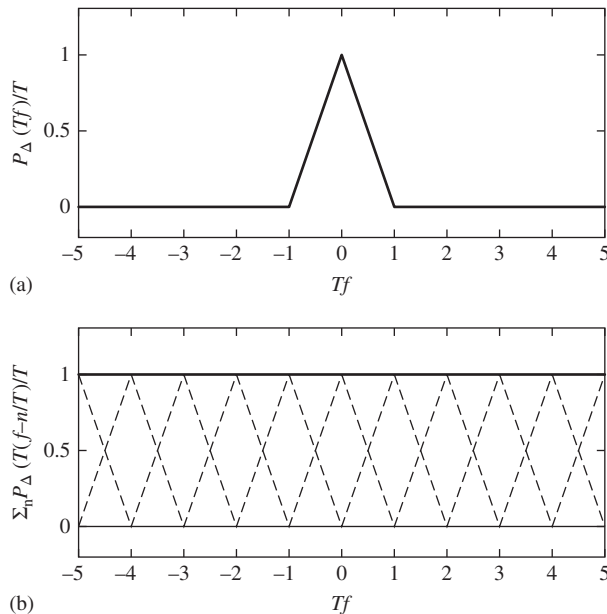


Figure 5.8

Illustration that a triangular spectrum (a), satisfies Nyquist's zero-ISI criterion (b).

5.4.3 Transmitter and Receiver Filters for Zero ISI

Consider the simplified pulse transmission system of Figure 5.9. A source produces a sequence of sample values $\{a_n\}$. Note that these are not necessarily quantized or binary digits, but they could be. For example, two bits per sample could be sent with four possible levels, representing 00, 01, 10, and 11. In the simplified transmitter model under consideration here, the k th sample value multiplies a unit impulse occurring at time kT and this weighted impulse train is the input to a transmitter filter with impulse response $h_T(t)$ and corresponding frequency response $H_T(f)$. The noise for now is assumed to be zero (effects of noise will be considered in Chapter 9). Thus, the input signal to the transmission channel, represented by a filter having

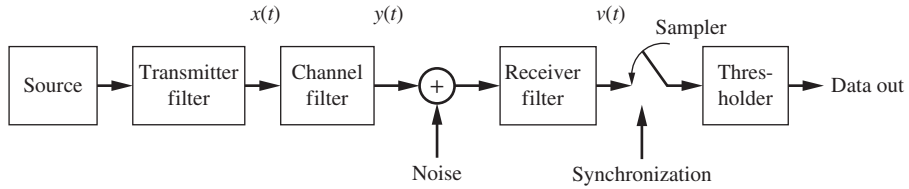


Figure 5.9 Transmitter, channel, and receiver cascade illustrating the implementation of a zero-ISI communication system.

impulse response $h_C(t)$ and corresponding frequency response $H_C(f)$, for all time is

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} a_k \delta(t - kT) * h_T(t) \\ &= \sum_{k=-\infty}^{\infty} a_k h_T(t - kT) \end{aligned} \quad (5.47)$$

The output of the channel is

$$y(t) = x(t) * h_C(t) \quad (5.48)$$

and the output of the receiver filter is

$$v(t) = y(t) * h_R(t) \quad (5.49)$$

We want the output of the receiver filter to have the zero-ISI property and, to be specific, we set

$$v(t) = \sum_{k=-\infty}^{\infty} a_k A p_{RC}(t - kT - t_d) \quad (5.50)$$

where $p_{RC}(t)$ is the raised-cosine pulse function, t_d represents the delay introduced by the cascade of filters, and A represents an amplitude scale factor. Putting this all together, we have

$$A p_{RC}(t - t_d) = h_T(t) * h_C(t) * h_R(t) \quad (5.51)$$

or, by Fourier-transforming both sides, we have

$$A P_{RC}(f) \exp(-j2\pi f t_d) = H_T(f) H_C(f) H_R(f) \quad (5.52)$$

In terms of amplitude responses this becomes

$$A P_{RC}(f) = |H_T(f)| |H_C(f)| |H_R(f)| \quad (5.53)$$

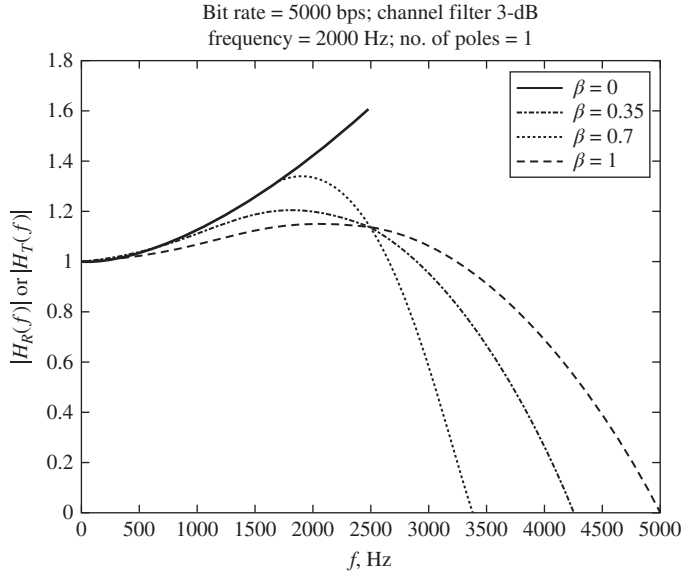


Figure 5.10

Transmitter and receiver filter amplitude responses that implement the zero-ISI condition assuming a first-order Butterworth channel filter and raised-cosine pulse shapes.

Now $|H_C(f)|$ is fixed (the channel is whatever it is) and $P_{RC}(f)$ is specified. Suppose we want the transmitter and receiver filter amplitude responses to be the same. Then, solving (5.46) with $|H_T(f)| = |H_R(f)|$, we have

$$|H_T(f)|^2 = |H_R(f)|^2 = \frac{AP_{RC}(f)}{|H_C(f)|} \quad (5.54)$$

or

$$|H_T(f)| = |H_R(f)| = \frac{A^{1/2}P_{RC}^{1/2}(f)}{|H_C(f)|^{1/2}} \quad (5.55)$$

This amplitude response is shown in Figure 5.10 for raised-cosine spectra of various roll-off factors and for a channel filter assumed to have a first-order Butterworth amplitude response. We have not accounted for the effects of additive noise. If the noise spectrum is flat, the only change would be another multiplicative constant. The constants are arbitrary since they multiply both signal and noise alike.

5.5 ZERO-FORCING EQUALIZATION

In the previous section, it was shown how to choose transmitter and receiver filter amplitude responses, given a certain channel filter, to provide output pulses satisfying the zero-ISI condition. In this section, we present a procedure for designing a filter that will accept a channel output pulse response not satisfying the zero-ISI condition and produce a pulse at its output that has N zero-valued samples on either side of its maximum sample value taken to

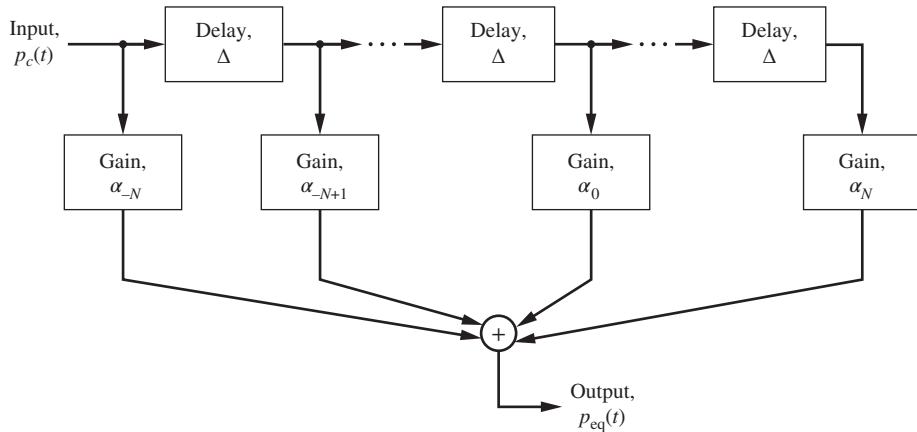


Figure 5.11
A transversal filter implementation for equalization of intersymbol interference.

be 1 for convenience. This filter will be called a zero-forcing equalizer. We specialize our considerations of an equalization filter to a particular form—a transversal or tapped-delay-line filter. Figure 5.11 shows the block diagram of such a filter.

There are at least two reasons for considering a transversal structure for the purpose of equalization. First, it is simple to analyze. Second, it is easy to mechanize by electronic means (i.e., transmission line delays and analog multipliers) at high frequencies and by digital signal processors at lower frequencies.

Let the pulse response of the channel output be $p_c(t)$. The output of the equalizer in response to $p_c(t)$ is

$$p_{\text{eq}}(t) = \sum_{n=-N}^N \alpha_n p_c(t - n\Delta) \quad (5.56)$$

where Δ is the tap spacing and the total number of transversal filter taps is $2N + 1$. We want $p_{\text{eq}}(t)$ to satisfy Nyquist's pulse-shaping criterion, which we will call the *zero-ISI condition*. Since the output of the equalizer is sampled every T seconds, it is reasonable that the tap spacing be $\Delta = T$. The zero-ISI condition therefore becomes

$$\begin{aligned} p_{\text{eq}}(mT) &= \sum_{n=-N}^N \alpha_n p_c[(m-n)T] \\ &= \begin{cases} 1, & m = 0 \\ 0, & m \neq 0 \end{cases} \quad m = 0, \pm 1, \pm 2, \dots, \pm N \end{aligned} \quad (5.57)$$

Note that the zero-ISI condition can be satisfied at only $2N$ time instants because there are only $2N + 1$ coefficients to be selected in (5.57) and the output of the filter for $t = 0$ is forced

to be 1. Defining the matrices (actually column matrices or vectors for the first two)

$$[P_{\text{eq}}] = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \left. \begin{array}{l} \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \\ \text{ } \end{array} \right\} \begin{array}{l} N \text{ zeros} \\ \\ \\ N \text{ zeros} \end{array} \quad (5.58)$$

$$[A] = \begin{bmatrix} \alpha_{-N} \\ \alpha_{-N+1} \\ \vdots \\ \alpha_N \end{bmatrix} \quad (5.59)$$

and

$$[P_c] = \begin{bmatrix} p_c(0) & p_c(-T) & \cdots & p_c(-2NT) \\ p_c(T) & p_c(0) & \cdots & p_c[(-2N+1)T] \\ \vdots & & & \vdots \\ p_c(2NT) & & & p_c(0) \end{bmatrix} \quad (5.60)$$

it follows that (5.57) can be written as the matrix equation

$$[P_{\text{eq}}] = [P_c][A] \quad (5.61)$$

The method of solution of the zero-forcing coefficients is now clear. Since $[P_{\text{eq}}]$ is specified by the zero-ISI condition, all we must do is multiply through by the inverse of $[P_c]$. The desired coefficient matrix $[A]$ is then the middle column of $[P_c]^{-1}$, which follows by multiplying $[P_c]^{-1}$ times $[P_{\text{eq}}]$:

$$[A] = [P_c]^{-1}[P_{\text{eq}}] = [P_c]^{-1} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \text{middle column of } [P_c]^{-1} \quad (5.62)$$

EXAMPLE 5.7

Consider a channel for which the following sample values of the channel pulse response are obtained:

$$\begin{aligned} p_c(-3T) &= 0.02 & p_c(-2T) &= -0.05 & p_c(-T) &= 0.2 & p_c(0) &= 1.0 \\ p_c(T) &= 0.3 & p_c(2T) &= -0.07 & p_c(3T) &= 0.03 \end{aligned}$$

The matrix $[P_c]$ is

$$[P_c] = \begin{bmatrix} 1.0 & 0.2 & -0.05 \\ 0.3 & 1.0 & 0.2 \\ -0.07 & 0.3 & 1.0 \end{bmatrix} \quad (5.63)$$

and the inverse of this matrix is

$$[P_c]^{-1} = \begin{bmatrix} 1.0815 & -0.2474 & 0.1035 \\ -0.3613 & 1.1465 & -0.2474 \\ 0.1841 & -0.3613 & 1.0815 \end{bmatrix} \quad (5.64)$$

Thus, by (5.62)

$$[A] = \begin{bmatrix} 1.0815 & -0.2474 & 0.1035 \\ -0.3613 & 1.1465 & -0.2474 \\ 0.1841 & -0.3613 & 1.0815 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.2474 \\ 1.1465 \\ -0.3613 \end{bmatrix} \quad (5.65)$$

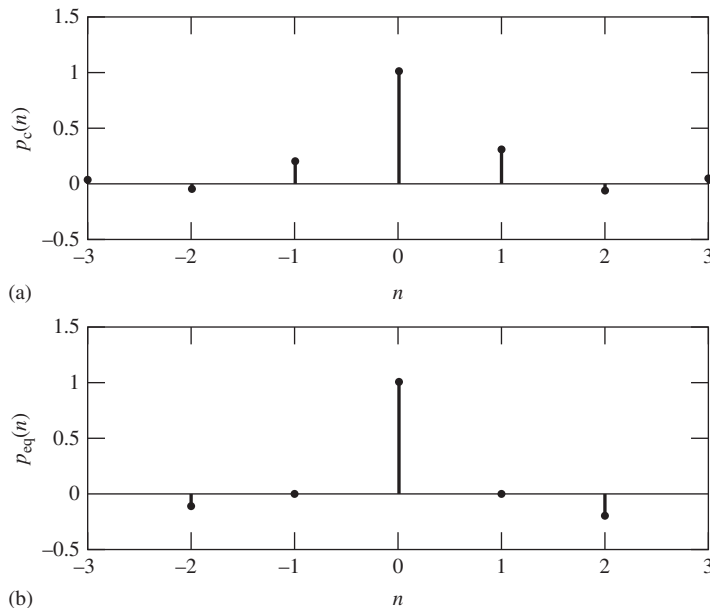


Figure 5.12

Samples for (a) an assumed channel response and (b) the output of a zero-forcing equalizer of length 3.

Using these coefficients, the equalizer output is

$$p_{\text{eq}}(m) = -0.2474p_c[(m+1)T] + 1.1465p_c(mT) - 0.3613p_c[(m-1)T], \quad m = \dots, -1, 0, 1, \dots \quad (5.66)$$

Putting values in shows that $p_{\text{eq}}(0) = 1$ and that the single samples on either side of $p_{\text{eq}}(0)$ are zero. Samples more than one away from the center sample are not necessarily zero for this example. Calculation using the extra samples for $p_c(nT)$ gives $p_c(-2T) = -0.1140$ and $p_c(2T) = -0.1961$. Samples for the channel and the equalizer outputs are shown in Figure 5.12. ■

5.6 EYE DIAGRAMS

We now consider eye diagrams that, although not a quantitative measure of system performance, are simple to construct and give significant insight into system performance. An eye diagram is constructed by plotting overlapping k -symbol segments of a baseband signal. In other words, an eye diagram can be displayed on an oscilloscope by triggering the time sweep of the oscilloscope, as shown in Figure 5.13, at times $t = nkT_s$ where T_s is the symbol period, kT_s is the eye period, and n is an integer. A simple example will demonstrate the process of generating an eye diagram.

EXAMPLE 5.8

Consider the eye diagram of a bandlimited digital NRZ baseband signal. In this example the signal is generated by passing an NRZ waveform through a third-order Butterworth filter as illustrated in Figure 5.13. The filter bandwidth is normalized to the symbol rate. In other words, if the symbol rate of the NRZ waveform is 1000 symbols per second, and the normalized filter bandwidth is $B_N = 0.6$, the filter bandwidth is 600 hertz. The eye diagrams corresponding to the signal at the filter output are those illustrated in Figure 5.14 for normalized bandwidths, B_N , of 0.4, 0.6, 1.0, and 2.0. Each of the four eye diagrams span $k = 4$ symbols. Sampling is performed at 20 samples/symbol and therefore the sampling index ranges from 1 to 80 as shown. The effect of bandlimiting by the filter, leading to intersymbol interference, on the eye diagram is clearly seen.

We now look at an eye diagram in more detail. Figure 5.15 shows the top pane of Figure 5.14 ($B_N = 0.4$), in which two symbols are illustrated rather than four. Observation of Figure 5.15 suggests that the eye diagram is composed of two fundamental waveforms, each of which approximates a sine wave. One waveform goes through two periods in the two symbol eyes and the other waveform goes through a single period. A little thought shows that the high-frequency waveform corresponds to the binary sequences 01 or 10 while the low-frequency waveform corresponds to the binary sequences 00 or 11.

Also shown in Figure 5.15 is the optimal sampling time, which is when the eye is most open. Note that for significant bandlimiting the eye will be more closed due to intersymbol interference. This shrinkage of the eye opening due to ISI is labeled amplitude jitter, A_j . Referring back to Figure 5.14 we see that increasing the filter bandwidth decreases the amplitude jitter. When we consider the effects of

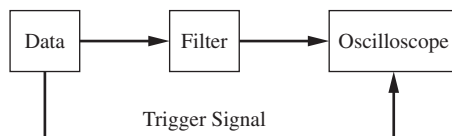


Figure 5.13
Simple technique for generating an eye diagram for a bandlimited signal.

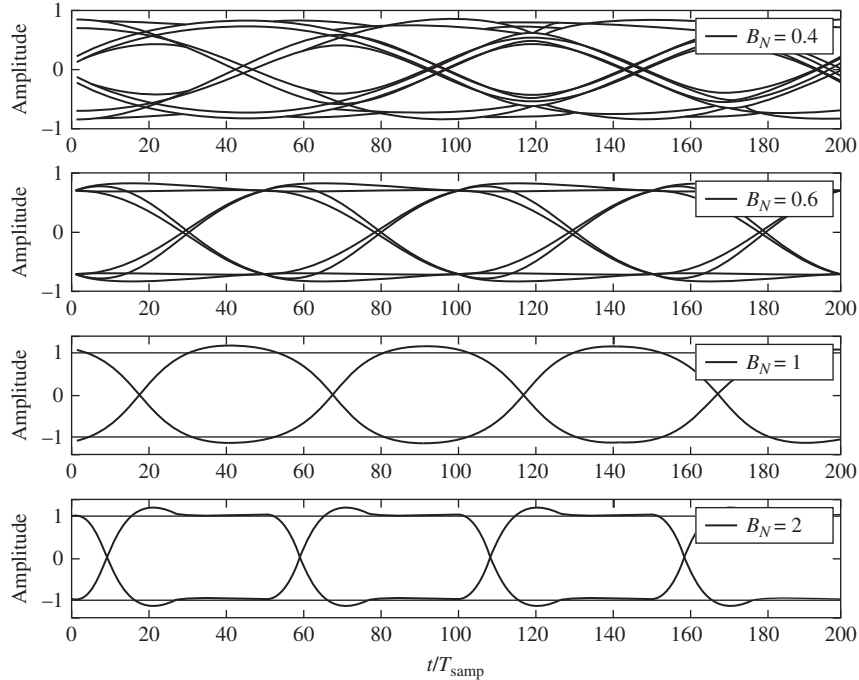


Figure 5.14
Eye diagrams for $B_N = 0.4, 0.6, 1.0,$ and 2.0 .

noise in later chapters of this book, we will see that if the vertical eye opening is reduced, the probability of symbol error increases. Note also that ISI leads to timing jitter, denoted T_j in Figure 5.15, which is a perturbation of the zero crossings of the filtered signal. Also note that a large slope of the signal at the zero crossings will result in a more open eye and that increasing this slope is accomplished by increasing the signal bandwidth. If the signal bandwidth is decreased leading to increased intersymbol interference, T_j increases and synchronization becomes more difficult. As we will see in later chapters, increasing the bandwidth of a channel often results in increased noise levels. This leads to both an increase in timing jitter and amplitude jitter. Thus, many trade-offs exist in the design of communication systems, several of which will be explored in later sections of this book.

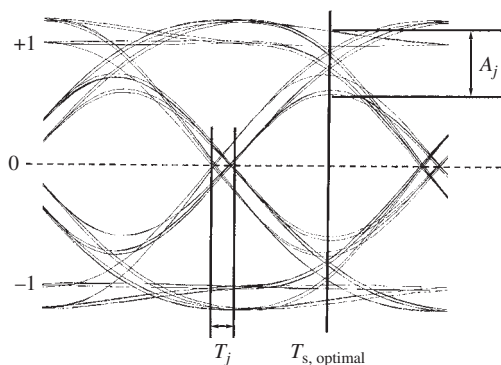


Figure 5.15
Two-symbol eye diagrams for $B_N = 0.4$.

COMPUTER EXAMPLE 5.2

The eye diagrams illustrated in Figure 5.15 were generated using the following MATLAB code:

```
% File: c5ce2.m
clf
nsym = 1000; nsamp = 50; bw = [0.4 0.6 1 2];
for k = 1:4
    lambda = bw(k);
    [b,a] = butter(3,2*lambda/nsamp);
    l = nsym*nsamp; % Total sequence length
    y = zeros(1,l-nsamp+1); % Initialize output vector
    x = 2*round(rand(1,nsym))-1; % Components of x = +1 or -1
    for i = 1:nsym % Loop to generate info symbols
        kk = (i-1)*nsamp+1;
        y(kk) = x(i);
    end
    datavector = conv(y,ones(1,nsamp)); % Each symbol is nsamp long
    filtout = filter(b, a, datavector);
    datamatrix = reshape(filtout, 4*nsamp, nsym/4);
    datamatrix1 = datamatrix(:, 6:(nsym/4));
    subplot(4,1,k), plot(datamatrix1, 'k'), ylabel('Amplitude'), ...
    axis([0 200 -1.4 1.4]), legend(['\itB.N} = ', num2str(lambda)])
    if k == 4
        xlabel('\itt/T}.s.a.m.p')
    end
end
% End of script file.
```

Note: The bandwidth values shown on Figure 5.14 were added using an editor after the figure was generated. Figure 5.15 was generated from the top pane of Figure 5.14 using an editor. ■

5.7 SYNCHRONIZATION

We now briefly look at the important subject of synchronization. There are many different levels of synchronization in a communications system. Coherent demodulation requires carrier synchronization as we discussed in the preceding chapter where we noted that a Costas PLL could be used to demodulate a DSB signal. In a digital communications system bit or symbol synchronization gives us knowledge of the starting and ending times of discrete-time symbols. This is a necessary step in data recovery. When block coding is used for error correction in a digital communications system, knowledge of the initial symbols in the code words must be identified for decoding. This process is known as word synchronization. In addition, groups of symbols are often grouped together to form data frames and frame synchronization is required to identify the starting and ending symbols in each data frame. In this section we focus on symbol synchronization. Other types of synchronization will be considered later in this book.

Three general methods exist by which symbol synchronization² can be obtained. These are (1) derivation from a primary or secondary standard (for example, transmitter and receiver

²See Stiffler (1971), Part II, or Lindsey and Simon (1973), Chapter 9, for a more extensive discussion.

slaved to a master timing source), (2) utilization of a separate synchronization signal (pilot clock), and (3) derivation from the modulation itself, referred to as *self-synchronization*. In this section we explore two self-synchronization techniques.

As we saw earlier in this chapter (see Figure 5.2), several binary data formats, such as polar RZ and split phase, guarantee a level transition within every symbol period that may aid in synchronization. For other data formats a discrete spectral component is present at the symbol frequency. A phase-locked loop, such as we studied in the preceding chapter, can then be used to track this component in order to recover symbol timing. For data formats that do not have a discrete spectral line at the symbol frequency, a nonlinear operation is performed on the signal in order to generate such a spectral component. A number of techniques are in common use for accomplishing this. The following examples illustrate two basic techniques, both of which make use of the PLL for timing recovery. Techniques for acquiring symbol synchronization that are similar in form to the Costas loop are also possible and will be discussed in Chapter 10.³

COMPUTER EXAMPLE 5.3

To demonstrate the first method we assume that a data signal is represented by an NRZ signal that has been bandlimited by passing it through a bandlimited channel. If this NRZ signal is squared, a component is generated at the symbol frequency. The component generated at the symbol frequency can then be phase tracked by a PLL in order to generate the symbol synchronization as illustrated by the following MATLAB simulation:

```
% File: c5ce3.m
nsym = 1000; nsamp = 50; lambda = 0.7;
[b,a] = butter(3,2*lambda/nsamp);
l = nsym*nsamp; % Total sequence length
y = zeros(1,l-nsamp+1); % Initialize output vector
x = 2*round(rand(1,nsym))-1; % Components of x = +1 or -1
for i = 1:nsym % Loop to generate info symbols
    k = (i-1)*nsamp+1;
    y(k) = x(i);
end
datavector1 = conv(y,ones(1,nsamp)); % Each symbol is nsamp long
subplot(3,1,1), plot(datavector1(1,200:799),'k', 'LineWidth', 1.5)
axis([0 600 -1.4 1.4]), ylabel('Amplitude')
filtout = filter(b,a,datavector1);
datavector2 = filtout.*filtout;
subplot(3,1,2), plot(datavector2(1,200:799),'k', 'LineWidth', 1.5)
ylabel('Amplitude')
y = fft(datavector2);
yy = abs(y)/(nsym*nsamp);
subplot(3,1,3), stem(yy(1,1:2*nsym),'k')
xlabel('FFT Bin'), ylabel('Spectrum')
% End of script file.
```

The results of executing the preceding MATLAB program are illustrated in Figure 5.16. Assume that the 1000 symbols generated by the MATLAB program occur in a time span of 1 s. Thus, the symbol rate

³Again, see Stiffler (1971) or Lindsey and Simon (1973).

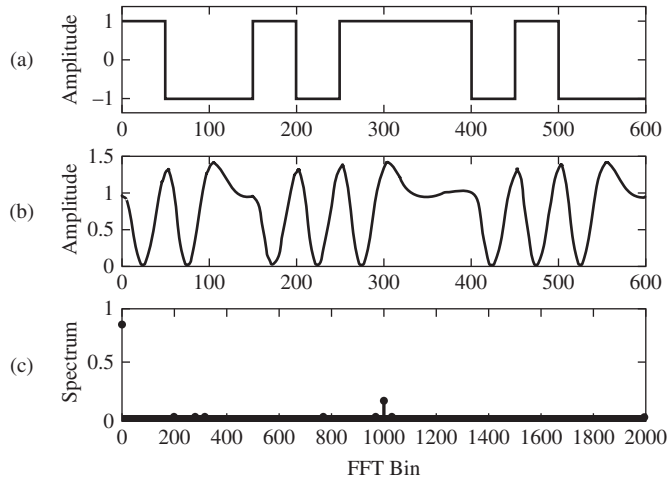


Figure 5.16

Simulation results for Computer Example 5.2: (a) NRZ waveform; (b) NRZ waveform filtered and squared; (c) FFT of squared NRZ waveform.

is 1000 symbols/s and, since the NRZ signal is sampled at 50 samples/symbol, the sampling frequency is 50,000 samples/second. Figure 5.16(a) illustrates 600 samples of the NRZ signal. Filtering by a third-order Butterworth filter having a bandwidth of twice the symbol rate and squaring this signal results in the signal shown in Figure 5.16(b). The second-order harmonic created by the squaring operation can clearly be seen by observing a data segment consisting of alternating data symbols. The spectrum, generated using the FFT algorithm, is illustrated in Figure 5.16(c). Two spectral components can clearly be seen; a component at DC (0 Hz), which results from the squaring operation, and a component at 1000 Hz, which represents the component at the symbol rate. This component is tracked by a PLL to establish symbol timing.

It is interesting to note that a sequence of alternating data states, e.g., 101010..., will result in an NRZ waveform that is a square wave. If the spectrum of this square wave is determined by forming the Fourier series, the period of the square wave will be twice the symbol period. The frequency of the fundamental will therefore be one-half the symbol rate. The squaring operation doubles the frequency to the symbol rate of 1000 symbols/s.

COMPUTER EXAMPLE 5.4

To demonstrate a second self-synchronization method, consider the system illustrated in Figure 5.17. Because of the nonlinear operation provided by the delay-and-multiply operation power is produced at the symbol frequency. The following MATLAB program simulates the symbol synchronizer:

```
% File: c5ce4.m
nsym = 1000; nsamp = 50;           % Make nsamp even
m = nsym*nsamp;
y = zeros(1,m-nsamp+1);           % Initialize output vector
x = 2*round(rand(1,nsym))-1;       % Components of x = +1 or -1
for i = 1:nsym                       % Loop to generate info symbols
    k = (i-1)*nsamp+1;
    y(k) = x(i);
```

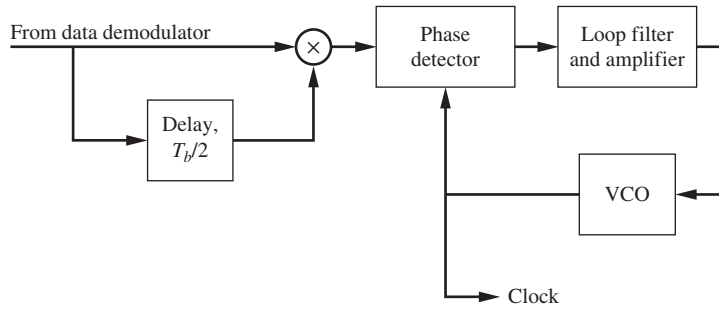


Figure 5.17

System for deriving a symbol clock simulated in Computer Example 5.4.

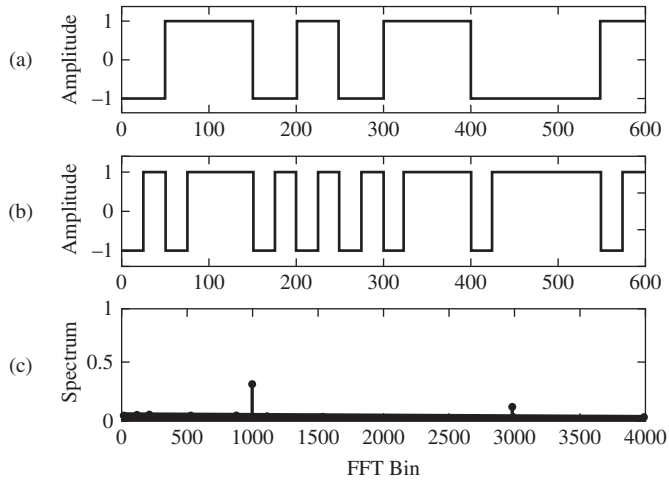


Figure 5.18

Simulation results for Computer Example 5.4: (a) data waveform; (b) data waveform multiplied by a half-bit delayed version of itself; (c) FFT spectrum of (b).

```

end
datavector1 = conv(y,ones(1,nsamp)); % Make symbols nsamp samples long
subplot(3,1,1), plot(datavector1(1,200:10000),'k', 'LineWidth', 1.5)
axis([0 600 -1.4 1.4]), ylabel('Amplitude')
datavector2 = [datavector1(1,m-nsamp/2+1:m) datavector1(1,1:m-
nsamp/2)];
datavector3 = datavector1.*datavector2;
subplot(3,1,2), plot(datavector3(1,200:10000),'k', 'LineWidth', 1.5),
axis([0 600 -1.4 1.4]), ylabel('Amplitude')
y = fft(datavector3);
yy = abs(y)/(nsym*nsamp);
subplot(3,1,3), stem(yy(1,1:4*nsym),'k.')
xlabel('FFT Bin'), ylabel('Spectrum')
% End of script file.

```

The data waveform is shown in Figure 5.18(a), and this waveform multiplied by its delayed version is shown in Figure 5.18(b). The spectral component at 1000 Hz, as seen in Figure 5.18(c), represents the symbol-rate component and is tracked by a PLL for timing recovery.

5.8 CARRIER MODULATION OF BASEBAND DIGITAL SIGNALS

The baseband digital signals considered in this chapter are typically transmitted using RF carrier modulation. As in the case of analog modulation considered in the preceding chapter, the fundamental techniques are based on amplitude, phase, or frequency modulation. This is illustrated in Figure 5.19 for the case in which the data bits are represented by an NRZ data format. Six bits are shown corresponding to the data sequence 101001. For digital amplitude modulation, known as amplitude-shift keying (ASK), the carrier amplitude is determined by the data bit for that interval. For digital phase modulation, known as phase-shift keying (PSK), the excess phase of the carrier is established by the data bit. The phase changes can clearly be seen in Figure 5.19. For digital frequency modulation, known as frequency-shift keying

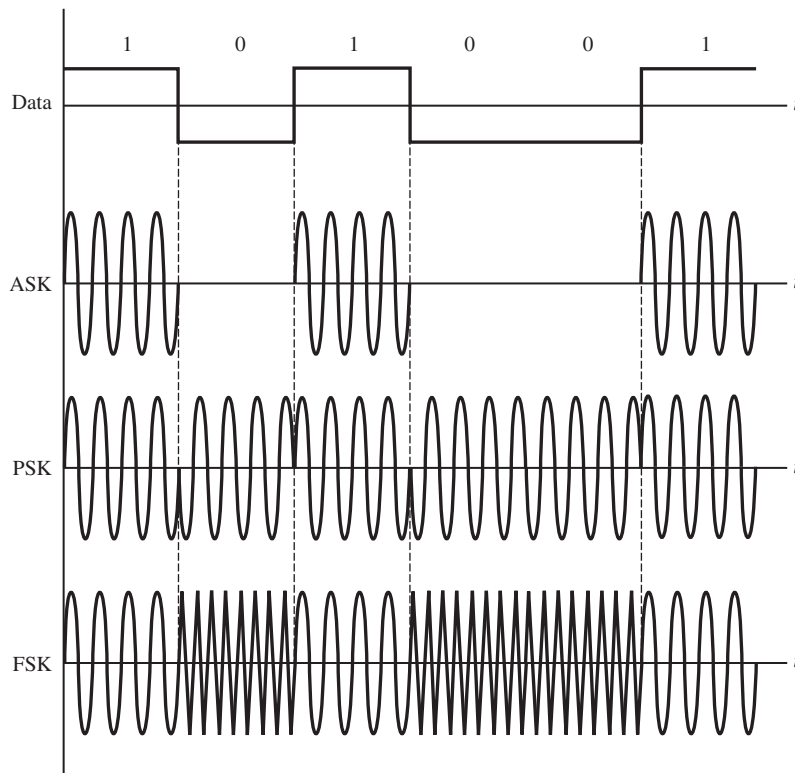


Figure 5.19
Examples of digital modulation schemes.

(FSK), the carrier frequency deviation is established by the data bit. To illustrate the similarity to the material studied in Chapters 3 and 4, note that the ASK RF signal can be represented by

$$x_{\text{ASK}}(t) = A_c[1 + d(t)] \cos(2\pi f_c t) \quad (5.67)$$

where $d(t)$ is the NRZ waveform. Note that this is identical to AM modulation with the only essential difference being the definition of the message signal. PSK and FSK can be similarly represented by

$$x_{\text{PSK}}(t) = A_c \cos \left[2\pi f_c t + \frac{\pi}{2} d(t) \right] \quad (5.68)$$

and

$$x_{\text{FSK}}(t) = A_c \cos \left[2\pi f_c t + k_f \int^t d(\alpha) d\alpha \right] \quad (5.69)$$

respectively. We therefore see that many of the concepts introduced in Chapters 3 and 4 carry over to digital data systems. These techniques will be studied in detail in Chapters 9 and 10. However, a major concern of both analog and digital communication systems is system performance in the presence of channel noise and other random disturbances. In order to have the tools required to undertake a study of system performance, we interrupt our discussion of communication systems to study random variables and stochastic processes.

Further Reading

Further discussions on the topics of this chapter may be found in Ziemer and Peterson (2001), Couch (2013), Proakis and Salehi (2005), and Anderson (1998).

Summary

1. The block diagram of the baseband model of a digital communications systems contains several components not present in the analog systems studied in the preceding chapters. The underlying message signal may be analog or digital. If the message signal is analog, an analog-to-digital converter must be used to convert the signal from analog to digital form. In such cases a digital-to-analog converter is usually used at the receiver output to convert the digital data back to analog form. Three operations covered in detail in this chapter were line coding, pulse shaping, and symbol synchronization.
2. Digital data can be represented using a number of formats, generally referred to as line codes. The two basic classifications of line codes are those that do not have an amplitude transition within each symbol period and those that do have an amplitude transition within each symbol period. A number of possibilities exist within each of these classifications. Two of the most popular data formats are NRZ (nonreturn to zero), which does not have an amplitude transition within each symbol period and split phase,

which does have an amplitude transition within each symbol period. The power spectral density corresponding to various data formats is important because of the impact on transmission bandwidth. Data formats having an amplitude transition within each symbol period may simplify symbol synchronization at the cost of increased bandwidth. Thus, bandwidth versus ease of synchronization are among the trade-offs available in digital transmission system design.

3. A major source of performance degradation in a digital system is intersymbol interference or ISI. Distortion due to ISI results when the bandwidth of a channel is not sufficient to pass all significant spectral components of the channel input signal. Channel equalization is often used to combat the effects of ISI. Equalization, in its simplest form, can be viewed as filtering the channel output using a filter having a frequency response function that is the inverse of the frequency response function of the channel.

4. A number of pulse shapes satisfy the Nyquist pulse-shaping criterion and result in zero ISI. A simple example is the pulse defined by $p(t) = \text{sinc}(t/T)$, where T is the

sampling (symbol) period. Zero ISI results since $p(t) = 1$ for $t = 0$ and $p(t) = 0$ for $t = nT, n \neq 0$.

5. A popular technique for implementing zero-ISI conditions is to use identical filters in both the transmitter and receiver. If the frequency response function of the channel is known and the underlying pulse shape is defined, the frequency response function of the transmitter/receiver filters can easily be found so that the Nyquist zero-ISI condition is satisfied. This technique is typically used with pulses having raised cosine spectra.

6. A zero-forcing equalizer is a digital filter that operates upon a channel output to produce a sequence of samples satisfying the Nyquist zero-ISI condition. The implementation takes the form of a tapped delay line, or transversal, filter. The tap weights are determined by the inverse of the matrix defining the pulse response of the channel. Attributes of the zero-forcing equalizer include ease of implementation and ease of analysis.

7. Eye diagrams are formed by overlaying segments of signals representing k data symbols. The eye diagrams, while not a quantitative measure of system performance, provide a qualitative measure of system performance. Signals with large vertical eye openings display lower levels of intersymbol interference than those with smaller vertical openings. Eyes with small horizontal openings have high levels of timing jitter, which makes symbol synchronization more difficult.

8. Many levels of synchronization are required in digital communication systems, including carrier, symbol, word, and frame synchronization. In this chapter we considered only symbol synchronization. Symbol synchronization is typically accomplished by using a PLL to track a component in the data signal at the symbol frequency. If the data format does not have discrete spectral lines at the symbol rate or multiples thereof, a nonlinear operation must be applied to the data signal in order to generate a spectral component at the symbol rate.

Drill Problems

5.1 Which data formats, for a random (coin toss) data stream, have (a) zero dc level; (b) built in redundancy that could be used for error checking; (c) discrete spectral lines present in their power spectra; (d) nulls in their spectra at zero frequency; (e) the most compact power spectra (measured to first null of their power spectra)?

- (i) NRZ change;
- (ii) NRZ mark;
- (iii) Unipolar RZ;
- (iv) Polar RZ;
- (v) Bipolar RZ;
- (vi) Split phase.

5.2 Tell which binary data format(s) shown in Figure 5.2 satisfy the following properties, assuming random (fair coin toss) data:

- (a) Zero DC level;
- (b) A zero crossing for each data bit;
- (c) Binary 0 data bits represented by 0 voltage level for transmission and the waveform has nonzero DC level;
- (d) Binary 0 data bits represented by 0 voltage level for transmission and the waveform has zero DC level;

(e) The spectrum is zero at frequency zero ($f = 0$ Hz);

(f) The spectrum has a discrete spectral line at frequency zero ($f = 0$ Hz).

5.3 Explain what happens to a line-coded data sequence when passed through a severely bandlimited channel.

5.4 What is meant by a pulse having the zero-ISI property? What must be true of the pulse spectrum in order that it have this property?

5.5 Which of the following pulse spectra have inverse Fourier transforms with the zero-ISI property?

(a) $P_1(f) = \Pi(Tf)$ where T is the pulse duration;

(b) $P_2(f) = \Lambda(Tf/2)$;

(c) $P_3(f) = \Pi(2Tf)$;

(d) $P_4(f) = \Pi(Tf) + \Pi(2Tf)$.

5.6 True or false: The zero-ISI property exists only for pulses with raised cosine spectra.

5.7 How many total samples of the incoming pulse are required to force the following number of zeros on either side of the middle sample for a zero-forcing equalizer?

(a) 1; (b) 3; (c) 4; (d) 7; (e) 8; (f) 10.

5.8 Choose the correct adjective: A wider bandwidth channel implies (more) (less) timing jitter.

5.9 Choose the correct adjective: A narrower bandwidth channel implies (more) (less) amplitude jitter.

5.10 Judging from the results of Figures 5.16 and 5.18, which method for generating a spectral component at the

data clock frequency generates a higher-power one: the squarer or the delay-and-multiply circuit?

5.11 Give advantages and disadvantages of the carrier modulation methods illustrated in Figure 5.19.

Problems

Section 5.1

5.1 Given the channel features or objectives below. For each part, tell which line code(s) is (are) the best choice(s).

- (a) The channel frequency response has a null at $f = 0$ hertz.
- (b) The channel has a passband from 0 to 10 kHz and it is desired to transmit data through it at 10,000 bits/s.
- (c) At least one zero crossing per bit is desired for synchronization purposes.
- (d) Built-in redundancy is desired for error-checking purposes.
- (e) For simplicity of detection, distinct positive pulses are desired for ones and distinct negative pulses are desired for zeros.
- (f) A discrete spectral line at the bit rate is desired from which to derive a clock at the bit rate.

5.2 For the ± 1 -amplitude waveforms of Figure 5.2, show that the average powers are:

- (a) NRZ change--- $P_{\text{ave}} = 1 \text{ W}$;
- (b) NRZ mark--- $P_{\text{ave}} = 1 \text{ W}$;
- (c) Unipolar RZ--- $P_{\text{ave}} = \frac{1}{4} \text{ W}$;
- (d) Polar RZ--- $P_{\text{ave}} = \frac{1}{2} \text{ W}$;
- (e) Bipolar RZ--- $P_{\text{ave}} = \frac{1}{4} \text{ W}$;
- (f) Split phase--- $P_{\text{ave}} = 1 \text{ W}$;

5.3

- (a) Given the random binary data sequence 0 1 1 0 0 0 1 0 1 1. Provide waveform sketches for:
 - (i) NRZ change;
 - (ii) Split phase.
- (b) Demonstrate satisfactorily that the split-phase waveform can be obtained from the NRZ waveform by multiplying the NRZ waveform by a ± 1 -valued clock signal of period T .

5.4 For the data sequence of Problem 5.3 provide a waveform sketch for NRZ mark.

5.5 For the data sequence of Problem 5.3 provide waveform sketches for:

- (a) Unipolar RZ;
- (b) Polar RZ;
- (c) Bipolar RZ.

5.6 A channel of bandwidth 4 kHz is available. Determine the data rate that can be accommodated for the following line codes (assume a bandwidth to the first spectral null):

- (a) NRZ change;
- (b) Split phase;
- (c) Unipolar RZ and polar RZ
- (d) Bipolar RZ.

Section 5.2

5.7 Given the step response for a second-order Butterworth filter as in Problem 2.65c, use the superposition and time-invariance properties of a linear time-invariant system to write down the filter's response to the input

$$x(t) = u(t) - 2u(t - T) + u(t - 2T)$$

where $u(t)$ is the unit step. Plot as a function of t/T for (a) $f_3 T = 20$ and (b) $f_3 T = 2$.

5.8 Using the superposition and time-invariance properties of an RC filter, show that (5.27) is the response of a lowpass RC filter to (5.26) given that the filter's response to a unit step is $[1 - \exp(-t/RC)] u(t)$.

Section 5.3

5.9 Show that (5.32) is an ideal rectangular spectrum for $\beta = 0$. What is the corresponding pulse-shape function?

5.10 Show that (5.31) and (5.32) are Fourier-transform pairs.

5.11 Sketch the following spectra and tell which ones satisfy Nyquist's pulse-shape criterion. For those that do,

find the appropriate sample interval, T , in terms of W . Find the corresponding pulse-shape function $p(t)$. (Recall that $\Pi\left(\frac{f}{A}\right)$ is a unit-high rectangular pulse from $-\frac{A}{2}$ to $\frac{A}{2}$; $\Lambda\left(\frac{f}{B}\right)$ is a unit-high triangle from $-B$ to B .)

- (a) $P_1(f) = \Pi\left(\frac{f}{2W}\right) + \Pi\left(\frac{f}{W}\right)$
- (b) $P_2(f) = \Lambda\left(\frac{f}{2W}\right) + \Pi\left(\frac{f}{W}\right)$
- (c) $P_3(f) = \Pi\left(\frac{f}{4W}\right) - \Lambda\left(\frac{f}{W}\right)$
- (d) $P_4(f) = \Pi\left(\frac{f-W}{W}\right) + \Pi\left(\frac{f+W}{W}\right)$
- (e) $P_5(f) = \Lambda\left(\frac{f}{2W}\right) - \Lambda\left(\frac{f}{W}\right)$

5.12 If $|H_C(f)| = [1 + (f/5000)^2]^{-1/2}$, provide a plot for $|H_T(f)| = |H_R(f)|$ assuming the pulse spectrum $P_{RC}(f)$ with $\frac{1}{T} = 5000$ Hz for (a) $\beta = 1$; (b) $\beta = \frac{1}{2}$.

5.13 It is desired to transmit data at 9 kbps over a channel of bandwidth 7 kHz using raised-cosine pulses. What is the maximum value of the roll-off factor, β , that can be used?

5.14

- (a) Show by a suitable sketch that the trapezoidal spectrum given below satisfies Nyquist's pulse-shaping criterion:

$$P(f) = 2\Lambda(f/2W) - \Lambda(f/W)$$

- (b) Find the pulse-shape function corresponding to this spectrum.

Section 5.4

5.15 Given the following channel pulse response samples:

$$\begin{aligned} p_c(-3T) &= 0.001 & p_c(-2T) &= -0.01 & p_c(-T) &= 0.1 & p_c(0) &= 1.0 \\ p_c(T) &= 0.2 & p_c(2T) &= -0.02 & p_c(3T) &= 0.005 \end{aligned}$$

- (a) Find the tap coefficients for a three-tap zero-forcing equalizer.
- (b) Find the output samples for $mT = -2T, -T, 0, T,$ and $2T$.

5.16 Repeat Problem 5.15 for a five-tap zero-forcing equalizer.

5.17 A simple model for a multipath communications channel is shown in Figure 5.20(a).

- (a) Find $H_c(f) = Y(f)/X(f)$ for this channel and plot $|H_c(f)|$ for $\beta = 1$ and 0.5.
- (b) In order to equalize, or undo, the channel-induced distortion, an equalization filter is used. Ideally, its frequency response function should be

$$H_{eq}(f) = \frac{1}{H_c(f)}$$

if the effects of noise are ignored and only distortion caused by the channel is considered. A tapped-delay-line or transversal filter, as shown in Figure 5.20(b), is commonly used to approximate $H_{eq}(f)$. Write down a series expression for $H'_{eq}(f) = Z(f)/Y(f)$.

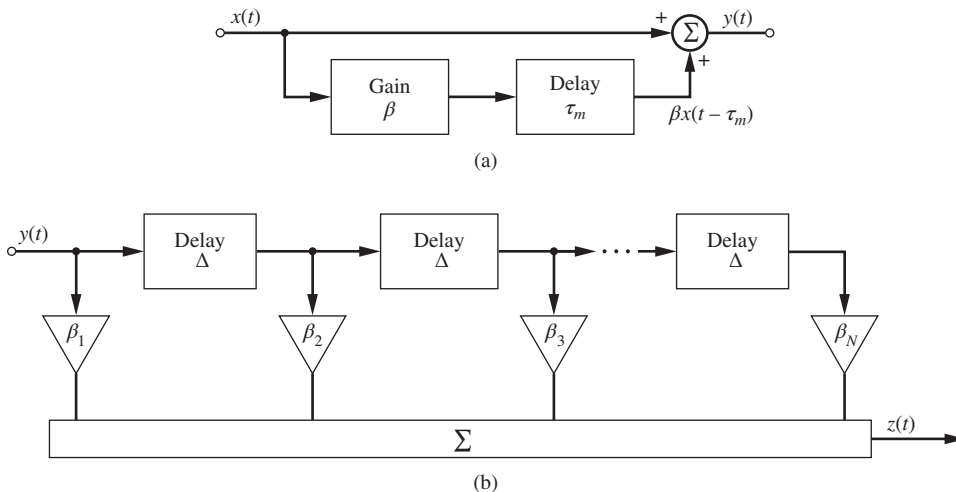


Figure 5.20

- (c) Using $(1+x)^{-1} = 1 - x + x^2 - x^3 + \dots$, $|x| < 1$, find a series expression for $1/H_c(f)$. Equating this with $H_{eq}(f)$ found in part (b), find the values for $\beta_1, \beta_2, \dots, \beta_N$, assuming $\tau_m = \Delta$.

5.18 Given the following channel pulse response:

$$\begin{aligned} p_c(-4T) &= -0.01; p_c(-3T) = 0.02; p_c(-2T) \\ &= -0.05; p_c(-T) = 0.07; p_c(0) = 1; \\ p_c(T) &= -0.1; p_c(2T) = 0.07; p_c(3T) \\ &= -0.05; p_c(4T) = 0.03; \end{aligned}$$

- (a) Find the tap weights for a three-tap zero-forcing equalizer.
 (b) Find the output samples for $mT = -2T, -T, 0, T, 2T$.

5.19 Repeat Problem 5.18 for a five-tap zero-forcing equalizer.

Section 5.5

5.20 In a certain digital data transmission system the probability of a bit error as a function of timing jitter is given by

$$P_E = \frac{1}{4} \exp(-z) + \frac{1}{4} \exp\left[-z \left(1 - 2 \frac{|\Delta T|}{T}\right)\right]$$

where z is the signal-to-noise ratio, $|\Delta T|$, is the timing jitter, and T is the bit period. From observations of an eye diagram for the system, it is determined that $|\Delta T|/T = 0.05$ (5%).

- (a) Find the value of signal-to-noise ratio, z_0 , that gives a probability of error of 10^{-6} for a timing jitter of 0.
 (b) With the jitter of 5%, tell what value of signal-to-noise ratio, z_1 , is necessary to maintain the probability of error at 10^{-6} . Express the ratio z_1/z_0 in dB, where $[z_1/z_0]_{\text{dB}} = 10 \log_{10}(z_1/z_0)$. Call this the degradation due to jitter.
 (c) Recalculate parts (a) and (b) for a probability of error of 10^{-4} . Is the degradation due to jitter better or worse than for a probability of error of 10^{-6} ?

5.21

- (a) Using the superposition and time-invariance properties of a linear time-invariant system find the response of a lowpass RC filter to the input

$$x(t) = u(t) - 2u(t - T) + 2u(t - 2T) - u(t - 3T)$$

Plot for $T/RC = 0.4, 0.6, 1, 2$ on separate axes. Use MATLAB to do so.

- (b) Repeat for $-x(t)$. Plot on the same set of axes as in part a.
 (c) Repeat for $x(t) = u(t)$.
 (d) Repeat for $x(t) = -u(t)$.

Note that you have just constructed a rudimentary eye diagram.

5.22 It is desired to transmit data ISI free at 10 kbps for which pulses with a raised-cosine spectrum are used. If the channel bandwidth is limited to 5 kHz, ideal lowpass, what is the allowed roll-off factor, β ?

5.23

- (a) For ISI-free signaling using pulses with raised-cosine spectra, give the relation of the roll-off factor, β , to data rate, $R = 1/T$, and channel bandwidth, f_{max} (assumed to be ideal lowpass).
 (b) What must be the relationship between R and f_{max} for realizable raised-cosine spectra pulses?

Section 5.6

5.24 Rewrite the MATLAB simulation of Example 5.8 for the case of an absolute-value type of nonlinearity. Is the spectral line at the bit rate stronger or weaker than for the square-law type of nonlinearity?

5.25 Assume that the bit period of Example 5.8 is $T = 1$ second. That means that the sampling rate is $f_s = 10$ sps because $\text{nsamp} = 10$ in the program. Assuming that a $N_{\text{FFT}} = 5000$ point FFT was used to produce Figure 5.16 and that the 5000th point corresponds to f_s justify that the FFT output at bin 1000 corresponds to the bit rate of $1/T = 1$ bit per second in this case.

Section 5.7

5.26 Referring to (5.68), it is sometimes desirable to leave a residual carrier component in a PSK-modulated waveform for carrier synchronization purposes at the receiver. Thus, instead of (5.68), we would have

$$x_{\text{PSK}}(t) = A_c \cos \left[2\pi f_c t + \alpha \frac{\pi}{2} d(t) \right], \quad 0 < \alpha < 1$$

Find α so that 10% of the power of $x_{\text{PSK}}(t)$ is in the carrier (unmodulated) component.

(Hint: Use $\cos(u+v)$ to write $x_{\text{PSK}}(t)$ as two terms, one dependent on $d(t)$ and the other independent of $d(t)$. Make

use of the facts that $d(t) = \pm 1$ and cosine is even and sine is odd.)

5.27 Referring to (5.69) and using the fact that $d(t) = \pm 1$ in T -second intervals, find the value of k_f such that the

peak frequency deviation of $x_{\text{FSK}}(t)$ is 10,000 Hz if the bit rate is 1000 bits per second.

Computer Exercises

5.1 Write a MATLAB program that will produce plots like those shown in Figure 5.2 assuming a random binary data sequence. Include as an option a Butterworth channel filter whose number of poles and bandwidth (in terms of bit rate) are inputs.

5.2 Write a MATLAB program that will produce plots like those shown in Figure 5.10. The Butterworth channel filter poles and 3-dB frequency should be inputs as well as the roll-off factor, β .

5.3 Write a MATLAB program that will compute the weights of a transversal-filter zero-

forcing equalizer for a given input pulse sample sequence.

5.4 A symbol synchronizer uses a fourth-power device instead of a squarer. Modify the MATLAB program of Computer Example 5.3 accordingly and show that a useful spectral component is generated at the output of the fourth-power device. Rewrite the program to be able to select between square-law, fourth-power law, and delay-and-multiply with delay of one-half bit period. Compare the relative strengths of the spectral line at the bit rate to the line at DC. Which is the best bit sync on this basis?

CHAPTER 6

OVERVIEW OF PROBABILITY AND
RANDOM VARIABLES

The objective of this chapter is to review probability theory in order to provide a background for the mathematical description of random signals. In the analysis and design of communication systems it is necessary to develop mathematical models for random signals and noise, or *random processes*, which will be accomplished in Chapter 7.

6.1 WHAT IS PROBABILITY?

Two intuitive notions of probability may be referred to as the equally likely outcomes and relative-frequency approaches.

6.1.1 Equally Likely Outcomes

The equally likely outcomes approach defines probability as follows: if there are N possible *equally likely* and *mutually exclusive* outcomes (that is, the occurrence of a given outcome precludes the occurrence of any of the others) to a random, or chance, experiment and if N_A of these outcomes correspond to an event A of interest, then the probability of event A , or $P(A)$, is

$$P(A) = \frac{N_A}{N} \quad (6.1)$$

There are practical difficulties with this definition of probability. One must be able to break the chance experiment up into two or more equally likely outcomes and this is not always possible. The most obvious experiments fitting these conditions are card games, dice, and coin tossing. Philosophically, there is difficulty with this definition in that use of the words “equally likely” really amounts to saying something about being equally probable, which means we are using probability to define probability.

Although there are difficulties with the equally likely definition of probability, it is useful in engineering problems when it is reasonable to list N equally likely, mutually exclusive outcomes. The following example illustrates its usefulness in a situation where it applies.

EXAMPLE 6.1

Given a deck of 52 playing cards, (a) What is the probability of drawing the ace of spades? (b) What is the probability of drawing a spade?

Solution

(a) Using the principle of equal likelihood, we have one favorable outcome in 52 possible outcomes. Therefore, $P(\text{ace of spades}) = \frac{1}{52}$. (b) Again using the principle of equal likelihood, we have 13 favorable outcomes in 52, and $P(\text{spade}) = \frac{13}{52} = \frac{1}{4}$. ■

6.1.2 Relative Frequency

Suppose we wish to assess the probability of an unborn child being a boy. Using the classical definition, we predict a probability of $\frac{1}{2}$, since there are two possible mutually exclusive outcomes, which from outward appearances appear equally probable. However, yearly birth statistics for the United States consistently indicate that the ratio of males to total births is about 0.51. This is an example of the relative-frequency approach to probability.

In the *relative-frequency approach*, we consider a random experiment, enumerate all possible outcomes, repeatedly perform the experiment, and take the ratio of the number of outcomes, N_A , favorable to an event of interest, A , to the total number of trials, N . As an approximation of the probability of A , $P(A)$, we define the limit of N_A/N , called the *relative frequency* of A , as $N \rightarrow \infty$, as $P(A)$:

$$P(A) \triangleq \lim_{N \rightarrow \infty} \frac{N_A}{N} \quad (6.2)$$

This definition of probability can be used to estimate $P(A)$. However, since the infinite number of experiments implied by (6.2) cannot be performed, only an approximation to $P(A)$ is obtained. Thus, the relative-frequency notion of probability is useful for estimating a probability but is not satisfactory as a mathematical basis for probability.

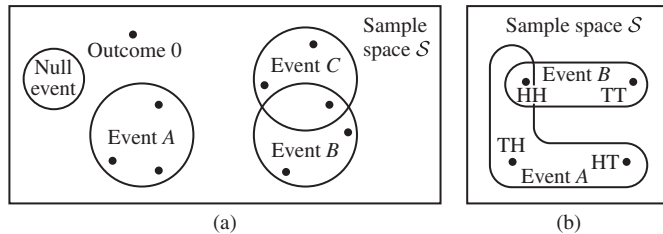
The following example fixes these ideas and will be referred to later in this chapter.

EXAMPLE 6.2

Consider the simultaneous tossing of two fair coins. Thus, on any given trial, we have the possible outcomes HH, HT, TH, and TT, where, for example, HT denotes a head on the first coin and a tail on the second coin. (We imagine that numbers are painted on the coins so we can tell them apart.) What is the probability of two heads on any given trial?

Solution

By distinguishing between the coins, the correct answer, using equal likelihood, is $\frac{1}{4}$. Similarly, it follows that $P(\text{HT}) = P(\text{TH}) = P(\text{TT}) = \frac{1}{4}$. ■

**Figure 6.1**

Sample spaces. (a) Pictorial representation of an arbitrary sample space. Points show outcomes; circles show events. (b) Sample-space representation for the tossing of two coins.

6.1.3 Sample Spaces and the Axioms of Probability

Because of the difficulties mentioned for the preceding two definitions of probability, mathematicians prefer to approach probability on an axiomatic basis. The axiomatic approach, which is general enough to encompass both the equally likely and relative-frequency definitions of probability, will now be briefly described.

A chance experiment can be viewed geometrically by representing its possible outcomes as elements of a space referred to as a sample space S . An *event* is defined as a collection of outcomes. An impossible collection of outcomes is referred to as the *null event*, ϕ . Figure 6.1(a) shows a representation of a sample space. Three events of interest, A , B , and C , which do not encompass the entire sample space, are shown.

A specific example of a chance experiment might consist of measuring the dc voltage at the output terminals of a power supply. The sample space for this experiment would be the collection of all possible numerical values for this voltage. On the other hand, if the experiment is the tossing of two coins, as in Example 6.2, the sample space would consist of the four outcomes HH , HT , TH , and TT enumerated earlier. A sample-space representation for this experiment is shown in Figure 6.1(b). Two events of interest, A and B , are shown. Event A denotes at least one head, and event B consists of the coins matching. Note that A and B encompass all possible outcomes for this particular example.

Before proceeding further, it is convenient to summarize some useful notation from set theory. The event “ A or B or both” will be denoted as $A \cup B$ or sometimes as $A + B$. The event “both A and B ” will be denoted either as $A \cap B$ or sometimes as (A, B) or AB (called the “joint event” A and B). The event “not A ” will be denoted \bar{A} . An event such as $A \cup B$, which is composed of two or more events, will be referred to as a *compound event*. In set theory terminology, “mutually exclusive events” are referred to as “disjoint sets”; if two events, A and B , are mutually exclusive, then $A \cap B = \phi$.

In the axiomatic approach, a *measure*, called *probability* is somehow assigned to the events of a sample space¹ such that this measure possesses the properties of probability. The properties or axioms of this probability measure are chosen to yield a satisfactory theory such that results from applying the theory will be consistent with experimentally observed phenomena. A set of satisfactory axioms is the following:

Axiom 1

$P(A) \geq 0$ for all events A in the sample space S .

¹For example, by the relative-frequency or the equally likely approaches.

Axiom 2

The probability of all possible events occurring is unity, $P(S) = 1$.

Axiom 3

If the occurrence of A precludes the occurrence of B , and vice versa (that is, A and B are mutually exclusive), then $P(A \cup B) = P(A) + P(B)$.²

It is emphasized that this approach to probability does not give us the number $P(A)$; it must be obtained by some other means.

6.1.4 Venn Diagrams

It is sometimes convenient to visualize the relationships between various events for a chance experiment in terms of a *Venn diagram*. In such diagrams, the sample space is indicated as a rectangle, with the various events indicated by circles or ellipses. Such a diagram looks exactly as shown in Figure 6.1(a), where it is seen that events B and C are not mutually exclusive, as indicated by the overlap between them, whereas event A is mutually exclusive of events B and C .

6.1.5 Some Useful Probability Relationships

Since it is true that $A \cup \bar{A} = S$ and that A and \bar{A} are mutually exclusive, it follows by Axioms 2 and 3 that $P(A) + P(\bar{A}) = P(S) = 1$, or

$$P(\bar{A}) = 1 - P(A) \quad (6.3)$$

A generalization of Axiom 3 to events that are not mutually exclusive is obtained by noting that $A \cup B = A \cup (B \cap \bar{A})$, where A and $B \cap \bar{A}$ are disjoint (this is most easily seen by using a Venn diagram). Therefore, Axiom 3 can be applied to give

$$P(A \cup B) = P(A) + P(B \cap \bar{A}) \quad (6.4)$$

Similarly, we note from a Venn diagram that the events $A \cap B$ and $B \cap \bar{A}$ are disjoint and that $(A \cap B) \cup (B \cap \bar{A}) = B$ so that

$$P(A \cap B) + P(B \cap \bar{A}) = P(B) \quad (6.5)$$

Solving for $P(B \cap \bar{A})$ from (6.5) and substituting into (6.4) yields the following for $P(A \cup B)$:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (6.6)$$

This is the desired generalization of Axiom 3.

Now consider two events A and B , with individual probabilities $P(A) > 0$ and $P(B) > 0$, respectively, and joint event probability $P(A \cap B)$. We define the *conditional probability* of

²This can be generalized to $P(A \cup B \cup C) = P(A) + P(B) + P(C)$ for A , B , and C mutually exclusive by considering $B_1 = B \cup C$ to be a composite event in Axiom 3 and applying Axiom 3 twice: i.e., $P(A \cup B_1) = P(A) + P(B_1) = P(A) + P(B) + P(C)$. Clearly, in this way we can generalize this result to any finite number of mutually exclusive events.

event A given that event B occurred as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (6.7)$$

Similarly, the conditional probability of event B given that event A has occurred is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (6.8)$$

Putting Equations (6.7) and (6.8) together, we obtain

$$P(A|B) P(B) = P(B|A) P(A) \quad (6.9)$$

or

$$P(B|A) = \frac{P(B) P(A|B)}{P(A)} \quad (6.10)$$

This is a special case of *Bayes' rule*.

Finally, suppose that the occurrence or nonoccurrence of B in no way influences the occurrence or nonoccurrence of A . If this is true, A and B are said to be *statistically independent*. Thus, if we are given B , this tells us nothing about A and therefore, $P(A|B) = P(A)$. Similarly, $P(B|A) = P(B)$. From Equation (6.7) or (6.8) it follows that, for such events,

$$P(A \cap B) = P(A)P(B) \quad (6.11)$$

Equation (6.11) will be taken as the definition of statistically independent events.

EXAMPLE 6.3

Referring to Example 6.2, suppose A denotes at least one head and B denotes a match. The sample space is shown in Figure 6.1(b). To find $P(A)$ and $P(B)$, we may proceed in several different ways.

Solution

First, if we use equal likelihood, there are three outcomes favorable to A (that is, HH, HT, and TH) among four possible outcomes, yielding $P(A) = \frac{3}{4}$. For B , there are two favorable outcomes in four possibilities, giving $P(B) = \frac{1}{2}$.

As a second approach, we note that, if the coins do not influence each other when tossed, the outcomes on separate coins are statistically independent with $P(H) = P(T) = \frac{1}{2}$. Also, event A consists of any of the mutually exclusive outcomes HH, TH, and HT, giving

$$P(A) = \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) = \frac{3}{4} \quad (6.12)$$

by (6.11) and Axiom 3, generalized. Similarly, since B consists of the mutually exclusive outcomes HH and TT,

$$P(B) = \left(\frac{1}{2} \cdot \frac{1}{2}\right) + \left(\frac{1}{2} \cdot \frac{1}{2}\right) = \frac{1}{2} \quad (6.13)$$

again through the use of (6.11) and Axiom 3. Also, $P(A \cap B) = P(\text{at least one head and a match}) = P(\text{HH}) = \frac{1}{4}$.

Next, consider the probability of at least one head given a match, $P(A|B)$. Using Bayes' rule, we obtain

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2} \quad (6.14)$$

which is reasonable, since given B , the only outcomes under consideration are HH and TT, only one of which is favorable to event A . Next, finding $P(B|A)$, the probability of a match given at least one head, we obtain

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \quad (6.15)$$

Checking this result using the principle of equal likelihood, we have one favorable event among three candidate events (HH, TH, and HT), which yields a probability of $\frac{1}{3}$. We note that

$$P(A \cap B) \neq P(A)P(B) \quad (6.16)$$

Thus, events A and B are not statistically independent, although the events H and T on either coin are independent.

Finally, consider the joint probability $P(A \cup B)$. Using (6.6), we obtain

$$P(A \cup B) = \frac{3}{4} + \frac{1}{2} - \frac{1}{4} = 1 \quad (6.17)$$

Remembering that $P(A \cup B)$ is the probability of at least one head, or a match, or both, we see that this includes all possible outcomes, thus confirming the result. ■

EXAMPLE 6.4

This example illustrates the reasoning to be applied when trying to determine if two events are independent. A single card is drawn at random from a deck of cards. Which of the following pairs of events are independent? (a) The card is a club, and the card is black. (b) The card is a king, and the card is black.

Solution

We use the relationship $P(A \cap B) = P(A|B)P(B)$ (always valid) and check it against the relation $P(A \cap B) = P(A)P(B)$ (valid only for independent events). For part (a), we let A be the event that the card is a club and B be the event that it is black. Since there are 26 black cards in an ordinary deck of cards, 13 of which are clubs, the conditional probability $P(A|B)$ is $\frac{13}{26}$ (given we are considering only black cards, we have 13 favorable outcomes for the card being a club). The probability that the card is black is $P(B) = \frac{26}{52}$, because half the cards in the 52-card deck are black. The probability of a club (event A), on the other hand, is $P(A) = \frac{13}{52}$ (13 cards in a 52-card deck are clubs). In this case,

$$P(A|B)P(B) = \frac{13}{26} \frac{26}{52} \neq P(A)P(B) = \frac{13}{52} \frac{26}{52} \quad (6.18)$$

so the events are not independent.

For part (b), we let A be the event that a king is drawn, and event B be that it is black. In this case, the probability of a king given that the card is black is $P(A|B) = \frac{2}{26}$ (two cards of the 26 black cards are kings). The probability of a king is simply $P(A) = \frac{4}{52}$ (four kings in the 52-card deck) and $P(B) = P(\text{black}) = \frac{26}{52}$. Hence,

$$P(A|B)P(B) = \frac{2}{26} \frac{26}{52} = P(A)P(B) = \frac{4}{52} \frac{26}{52} \quad (6.19)$$

which shows that the events king and black are statistically independent. ■

EXAMPLE 6.5

As an example more closely related to communications, consider the transmission of binary digits through a channel as might occur, for example, in computer networks. As is customary, we denote the two possible symbols as 0 and 1. Let the probability of receiving a zero, given a zero was sent, $P(0r|0s)$, and the probability of receiving a 1, given a 1 was sent, $P(1r|1s)$, be

$$P(0r|0s) = P(1r|1s) = 0.9 \quad (6.20)$$

Thus, the probabilities $P(1r|0s)$ and $P(0r|1s)$ must be

$$P(1r|0s) = 1 - P(0r|0s) = 0.1 \quad (6.21)$$

and

$$P(0r|1s) = 1 - P(1r|1s) = 0.1 \quad (6.22)$$

respectively. These probabilities characterize the channel and would be obtained through experimental measurement or analysis. Techniques for calculating them for particular situations will be discussed in Chapters 9 and 10.

In addition to these probabilities, suppose that we have determined through measurement that the probability of sending a 0 is

$$P(0s) = 0.8 \quad (6.23)$$

and therefore the probability of sending a 1 is

$$P(1s) = 1 - P(0s) = 0.2 \quad (6.24)$$

Note that once $P(0r|0s)$, $P(1r|1s)$, and $P(0s)$ are specified, the remaining probabilities are calculated using Axioms 2 and 3.

The next question we ask is, ‘‘If a 1 was received, what is the probability, $P(1s|1r)$, that a 1 was sent?’’ Applying Bayes’ rule, we find that

$$P(1s|1r) = \frac{P(1r|1s)P(1s)}{P(1r)} \quad (6.25)$$

To find $P(1r)$, we note that

$$P(1r, 1s) = P(1r|1s)P(1s) = 0.18 \quad (6.26)$$

and

$$P(1r, 0s) = P(1r|0s)P(0s) = 0.08 \quad (6.27)$$

Thus,

$$P(1r) = P(1r, 1s) + P(1r, 0s) = 0.18 + 0.08 = 0.26 \quad (6.28)$$

and

$$P(1s|1r) = \frac{(0.9)(0.2)}{0.26} = 0.69 \quad (6.29)$$

Similarly, one can calculate $P(0s|1r) = 0.31$, $P(0s|0r) = 0.97$, and $P(1s|0r) = 0.03$. For practice, you should go through the necessary calculations. ■

6.1.6 Tree Diagrams

Another handy device for determining probabilities of compound events is a *tree diagram*, particularly if the compound event can be visualized as happening in time sequence. This device is illustrated by the following example.

EXAMPLE 6.6

Suppose five cards are drawn without replacement from a standard 52-card deck. What is the probability that three of a kind results?

Solution

The tree diagram for this chance experiment is shown in Figure 6.2. On the first draw we focus on a particular card, denoted as X , which we either draw or do not. The second draw results in four possible events of interest: a card is drawn that matches the first card with probability $\frac{3}{51}$ or a match is not obtained with probability $\frac{48}{51}$. If some card other than X was drawn on the first draw, then X results with probability $\frac{4}{51}$ on the second draw (lower half of Figure 6.2). At this point, 50 cards are left in the deck. If we follow the upper branch, which corresponds to a match of the first card, two events of interest are again possible: another match that will be referred to as a *triple* with probability of $\frac{2}{50}$ on that draw, or a card that does not match the first two with probability $\frac{48}{50}$. If a card other than X was obtained on the second draw, then X occurs with probability $\frac{4}{50}$ if X was obtained on the first draw, and probability $\frac{46}{50}$ if it was not. The remaining branches are filled in similarly. Each path through the tree will either result in success or failure, and the probability of drawing the cards along a particular path will be the product of the separate probabilities along each path. Since a particular sequence of draws resulting in success is mutually exclusive of the sequence of draws resulting in any other success, we simply add up all the products of probabilities along all paths that result in success. In addition to these sequences involving card X , there are 12 others involving other face values that result in three of a kind. Thus, we multiply the result obtained from Figure 6.2 by 13. The probability of drawing three cards of the same value, in any order, is then given by

$$\begin{aligned} P(3 \text{ of a kind}) &= 13 \frac{(10)(4)(3)(2)(48)(47)}{(52)(51)(50)(49)(48)} \\ &= 0.02257 \end{aligned} \quad (6.30)$$

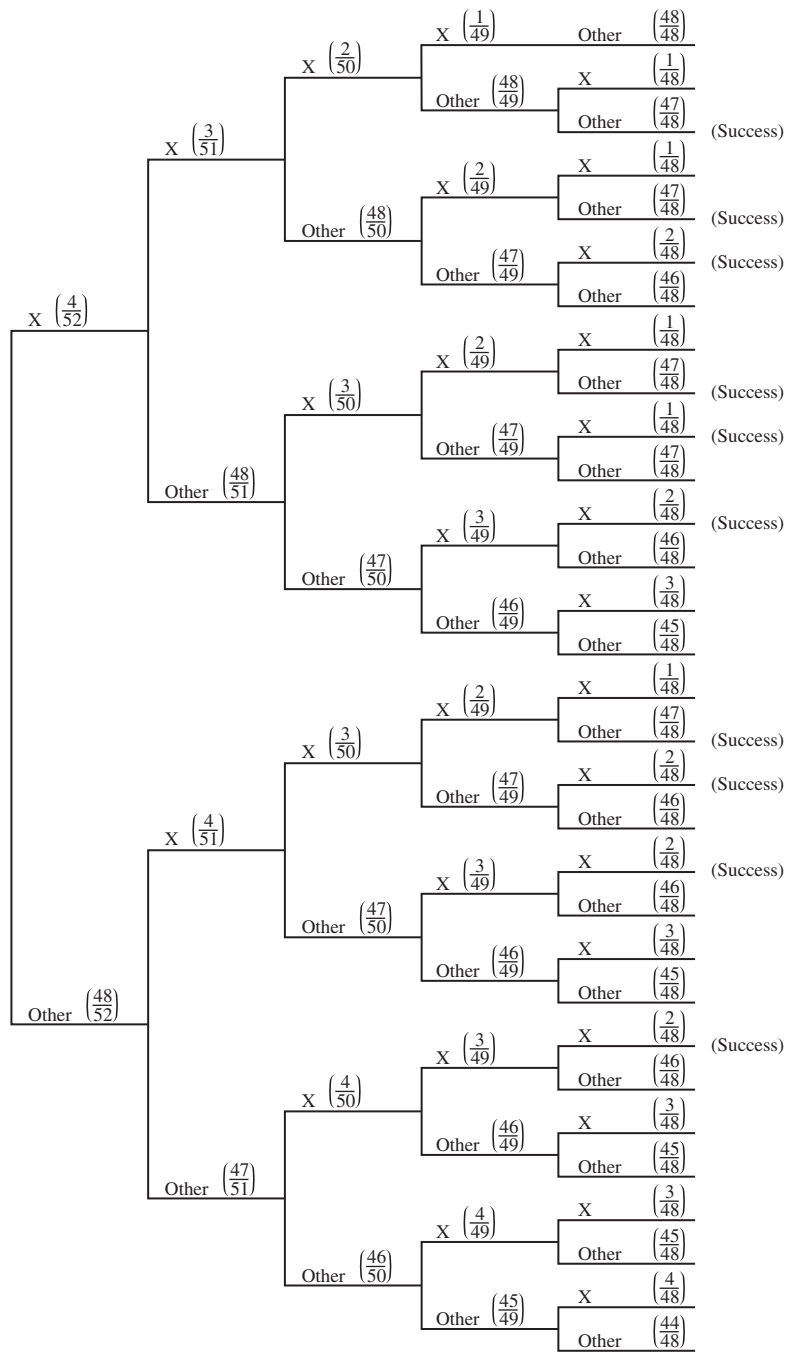


Figure 6.2
A card-drawing problem illustrating the use of a tree diagram.

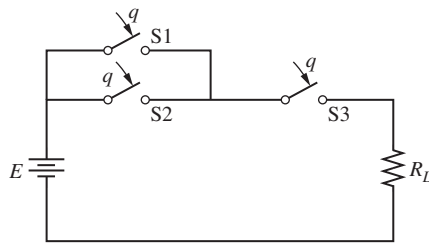


EXAMPLE 6.7

Another type of problem very closely related to those amenable to tree-diagram solutions is a reliability problem. Reliability problems can result from considering the overall failure of a system composed of several components each of which may fail with a certain probability p . An example is shown in Figure 6.3, where a battery is connected to a load through the series-parallel combination of relay switches, each of which may fail to close with probability p (or close with probability $q = 1 - p$). The problem is to find the probability that current flows in the load. From the diagram, it is clear that a circuit is completed if S1 or S2 and S3 are closed. Therefore,

$$\begin{aligned}
 P(\text{success}) &= P(\text{S1 or S2 and S3 closed}) \\
 &= P(\text{S1 or S2 or both closed})P(\text{S3 closed}) \\
 &= [1 - P(\text{both switches open})]P(\text{S3 closed}) \\
 &= (1 - p^2)q
 \end{aligned} \tag{6.31}$$

where it is assumed that the separate switch actions are statistically independent.

**Figure 6.3**

Circuit illustrating the calculation of reliability.

6.1.7 Some More General Relationships

Some useful formulas for a somewhat more general case than those considered above will now be derived. Consider an experiment composed of compound events (A_i, B_j) that are mutually exclusive. The totality of all these compound events, $i = 1, 2, \dots, M$, $j = 1, 2, \dots, N$, composes the entire sample space (that is, the events are said to be exhaustive or to form a partition of the sample space). For example, the experiment might consist of rolling a pair of dice with $(A_i, B_j) = (\text{number of spots showing on die 1}, \text{number of spots showing on die 2})$.

Suppose the probability of the joint event (A_i, B_j) is $P(A_i, B_j)$. Each compound event can be thought of as a simple event, and if the probabilities of all these mutually exclusive, exhaustive events are summed, a probability of 1 will be obtained, since the probabilities of all possible outcomes have been included. That is,

$$\sum_{i=1}^M \sum_{j=1}^N P(A_i, B_j) = 1 \tag{6.32}$$

Now consider a particular event B_j . Associated with this particular event, we have M possible mutually exclusive, but not exhaustive outcomes $(A_1, B_j), (A_2, B_j), \dots, (A_M, B_j)$. If we sum over the corresponding probabilities, we obtain the probability of B_j irrespective of the

outcome on A . Thus,

$$P(B_j) = \sum_{i=1}^M P(A_i, B_j) \quad (6.33)$$

Similar reasoning leads to the result

$$P(A_i) = \sum_{j=1}^N P(A_i, B_j) \quad (6.34)$$

$P(A_i)$ and $P(B_j)$ are referred to as *marginal probabilities*.

Suppose the conditional probability of B_m given A_n , $P(B_m|A_n)$, is desired. In terms of the joint probabilities $P(A_i, B_j)$, we can write this conditional probability as

$$P(B_m|A_n) = \frac{P(A_n, B_m)}{\sum_{j=1}^N P(A_n, B_j)} \quad (6.35)$$

which is a more general form of Bayes' rule than that given by (6.10).

EXAMPLE 6.8

A certain experiment has the joint and marginal probabilities shown in Table 6.1. Find the missing probabilities.

Solution

Using $P(B_1) = P(A_1, B_1) + P(A_2, B_1)$, we obtain $P(B_1) = 0.1 + 0.1 = 0.2$. Also, since $P(B_1) + P(B_2) + P(B_3) = 1$, we have $P(B_3) = 1 - 0.2 - 0.5 = 0.3$. Finally, using $P(A_1, B_3) + P(A_2, B_3) = P(B_3)$, we get $P(A_1, B_3) = 0.3 - 0.1 = 0.2$, and therefore, $P(A_1) = 0.1 + 0.4 + 0.2 = 0.7$

Table 6.1 $P(A_i, B_j)$

B_j				
A_i	B_1	B_2	B_3	$P(A_i)$
A_1	0.1	0.4	?	?
A_2	0.1	0.1	0.1	0.3
$P(B_j)$?	0.5	?	1

6.2 RANDOM VARIABLES AND RELATED FUNCTIONS

6.2.1 Random Variables

In the applications of probability it is often more convenient to work in terms of numerical outcomes (for example, the number of errors in a digital data message) rather than nonnumerical outcomes (for example, failure of a component). Because of this, we introduce the idea of a *random variable*, which is defined as a rule that assigns a numerical value to each possible

Table 6.2 Possible Random Variables

Outcome: S_i	R.V. No. 1: $X_1(S_i)$	R.V. No. 2: $X_2(S_i)$
$S_1 = \text{heads}$	$X_1(S_1) = 1$	$X_2(S_1) = \pi$
$S_2 = \text{tails}$	$X_1(S_2) = -1$	$X_2(S_2) = \sqrt{2}$

outcome of a chance experiment. (The term *random variable* is a misnomer; a random variable is really a function, since it is a rule that assigns the members of one set to those of another.)

As an example, consider the tossing of a coin. Possible assignments of random variables are given in Table 6.2. These are examples of *discrete random variables* and are illustrated in Figure 6.4(a).

As an example of a *continuous random variable*, consider the spinning of a pointer, such as is typically found in children's games. A possible assignment of a random variable would be the angle Θ_1 in radians, that the pointer makes with the vertical when it stops. Defined in this fashion, Θ_1 has values that continuously increase with rotation of the pointer. A second possible random variable, Θ_2 , would be Θ_1 minus integer multiples of 2π radians, such that

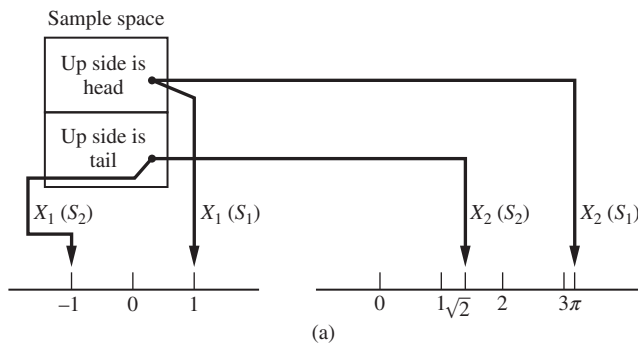
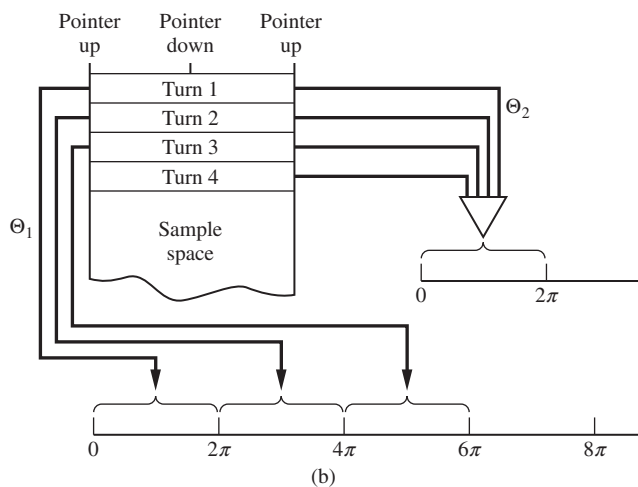


Figure 6.4

Pictorial representation of sample spaces and random variables.

(a) Coin-tossing experiment.

(b) Pointer-spinning experiment.



$0 \leq \Theta_2 < 2\pi$, which is commonly denoted as Θ_1 modulo 2π . These random variables are illustrated in Figure 6.4(b).

At this point, we introduce a convention that will be adhered to, for the most part, throughout this book. Capital letters (X , Θ , and so on) denote random variables, and the corresponding lowercase letters (x , θ , and so on) *denote the values that the random variables take on or running values for them*.

6.2.2 Probability (Cumulative) Distribution Functions

We need some way of probabilistically describing random variables that works equally well for discrete and continuous random variables. One way of accomplishing this is by means of the *cumulative-distribution function (cdf)*.

Consider a chance experiment with which we have associated a random variable X . The cdf $F_X(x)$ is defined as

$$F_X(x) = \text{probability that } X \leq x = P(X \leq x) \quad (6.36)$$

We note that $F_X(x)$ is a function of x , not of the random variable X . But $F_X(x)$ also depends on the assignment of the random variable X , which accounts for the subscript.

The cdf has the following properties:

Property 1. $0 \leq F_X(x) \leq 1$, with $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.

Property 2. $F_X(x)$ is continuous from the right; that is, $\lim_{x \rightarrow x_0^+} F_X(x) = F_X(x_0)$.

Property 3. $F_X(x)$ is a nondecreasing function of x ; that is, $F_X(x_1) \leq F_X(x_2)$ if $x_1 < x_2$.

The reasonableness of the preceding properties is shown by the following considerations.

Since $F_X(x)$ is a probability, it must, by the previously stated axioms, lie between 0 and 1, inclusive. Since $X = -\infty$ excludes all possible outcomes of the experiment, $F_X(-\infty) = 0$, and since $X = \infty$ includes all possible outcomes, $F_X(\infty) = 1$, which verifies Property 1.

For $x_1 < x_2$, the events $X \leq x_1$ and $x_1 < X \leq x_2$ are mutually exclusive; furthermore, $X \leq x_2$ implies $X \leq x_1$ or $x_1 < X \leq x_2$. By Axiom 3, therefore,

$$P(X \leq x_2) = P(X \leq x_1) + P(x_1 < X \leq x_2)$$

or

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) \quad (6.37)$$

Since probabilities are nonnegative, the left-hand side of (6.37) is nonnegative. Thus, we see that Property 3 holds.

The reasonableness of the right-continuity property is shown as follows. Suppose the random variable X takes on the value x_0 with probability P_0 . Consider $P(X \leq x)$. If $x < x_0$, the event $X = x_0$ is not included, no matter how close x is to x_0 . When $x = x_0$, we include the event $X = x_0$, which occurs with probability P_0 . Since the events $X \leq x < x_0$ and $X = x_0$ are mutually exclusive, $P(X \leq x)$ must jump by an amount P_0 when $x = x_0$, as shown in Figure 6.5. Thus, $F_X(x) = P(X \leq x)$ is continuous from the right. This is illustrated in Figure 6.5 by the dot on the curve to the right of the jump. What is more useful for our purposes, however, is that the *magnitude of any jump of $F_X(x)$, say at x_0 , is equal to the probability that $X = x_0$* .

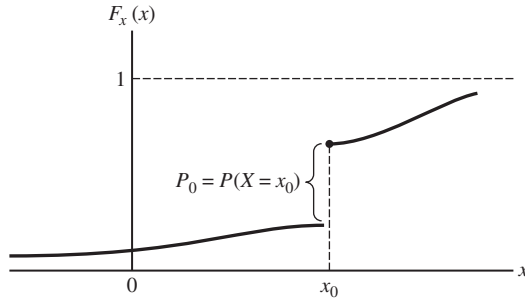


Figure 6.5

Illustration of the jump property of $f_X(x)$.

6.2.3 Probability-Density Function

From (6.37) we see that the cdf of a random variable is a complete and useful description for the computation of probabilities. However, for purposes of computing statistical averages, the probability-density function (pdf), $f_X(x)$, of a random variable, X , is more convenient. The pdf of X is defined in terms of the cdf of X by

$$f_X(x) = \frac{dF_X(x)}{dx} \quad (6.38)$$

Since the cdf of a discrete random variable is discontinuous, its pdf, mathematically speaking, does not exist at the points of discontinuity. By representing the derivative of a jump-discontinuous function at a point of discontinuity by a delta function of area equal to the magnitude of the jump, we can define pdfs for discrete random variables. In some books, this problem is avoided by defining a *probability mass function* for a discrete random variable, which consists simply of lines equal in magnitude to the probabilities that the random variable takes on at its possible values.

Recalling that $f_X(-\infty) = 0$, we see from (6.38) that

$$f_X(x) = \int_{-\infty}^x f_X(\eta) d\eta \quad (6.39)$$

That is, the *area* under the pdf from $-\infty$ to x is the probability that the observed value will be less than or equal to x .

From (6.38), (6.39), and the properties of $f_X(x)$, we see that the pdf has the following properties:

$$f_X(x) = \frac{dF_X(x)}{dx} \geq 0 \quad (6.40)$$

$$\int_{-\infty}^{\infty} f_X(x) dx = 1 \quad (6.41)$$

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x) dx \quad (6.42)$$

To obtain another enlightening and very useful interpretation of $f_X(x)$, we consider (6.42) with $x_1 = x - dx$ and $x_2 = x$. The integral then becomes $f_X(x) dx$, so

$$f_X(x) dx = P(x - dx < X \leq x) \quad (6.43)$$

That is, the ordinate at any point x on the pdf curve multiplied by dx gives the probability of the random variable X lying in an infinitesimal range around the point x assuming that $f_X(x)$ is continuous at x .

The following two examples illustrate cdfs and pdfs for discrete and continuous cases, respectively.

EXAMPLE 6.9

Suppose two fair coins are tossed and X denotes the number of heads that turn up. The possible outcomes, the corresponding values of X , and the respective probabilities are summarized in Table 6.3. The cdf and pdf for this experiment and random variable definition are shown in Figure 6.6. The properties of the cdf and pdf for discrete random variables are demonstrated by this figure, as a careful examination will reveal. It is emphasized that the cdf and pdf change if the definition of the random variable or the probability assigned is changed.

Table 6.3 Outcomes and Probabilities

Outcome	X	$P(X = x_j)$
TT	$x_1 = 0$	$\frac{1}{4}$
TH	$x_2 = 1$	$\frac{1}{2}$
HT		
HH	$x_3 = 2$	$\frac{1}{4}$

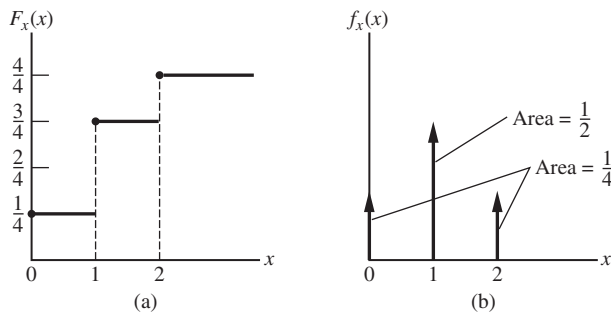


Figure 6.6

The cdf (a) and pdf (b) for a coin-tossing experiment.

EXAMPLE 6.10

Consider the pointer-spinning experiment described earlier. We assume that any one stopping point is not favored over any other and that the random variable Θ is defined as the angle that the pointer makes with the vertical, modulo 2π . Thus, Θ is limited to the range $[0, 2\pi)$, and for any two angles θ_1 and θ_2 in $[0, 2\pi)$, we have

$$P(\theta_1 - \Delta\theta < \Theta \leq \theta_1) = P(\theta_2 - \Delta\theta < \Theta \leq \theta_2) \quad (6.44)$$

by the assumption that the pointer is equally likely to stop at any angle in $[0, 2\pi)$. In terms of the pdf $f_\Theta(\theta)$, this can be written, using (6.37), as

$$f_\Theta(\theta_1) = f_\Theta(\theta_2), \quad 0 \leq \theta_1, \theta_2 < 2\pi \quad (6.45)$$

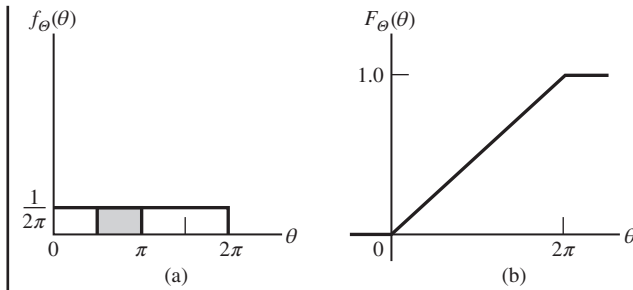


Figure 6.7
The pdf (a) and cdf (b) for a pointer-spinning experiment.

Thus, in the interval $[0, 2\pi)$, $f_{\Theta}(\theta)$ is a constant, and outside $[0, 2\pi)$, $f_{\Theta}(\theta)$ is zero by the modulo 2π condition (this means that angles less than or equal to 0 or greater than 2π are impossible). By (6.35), it follows that

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{2\pi}, & 0 \leq \theta < 2\pi \\ 0, & \text{otherwise} \end{cases} \quad (6.46)$$

The pdf $f_{\Theta}(\theta)$ is shown graphically in Figure 6.7(a). The cdf $F_{\Theta}(\theta)$ is easily obtained by performing a graphical integration of $f_{\Theta}(\theta)$ and is shown in Figure 6.7(b).

To illustrate the use of these graphs, suppose we wish to find the probability of the pointer landing anywhere in the interval $[\frac{1}{2}\pi, \pi]$. The desired probability is given either as the area under the pdf curve from $\frac{1}{2}\pi$ to π , shaded in Figure 6.7(a), or as the value of the ordinate at $\theta = \pi$ minus the value of the ordinate at $\theta = \frac{1}{2}\pi$ on the cdf curve. The probability that the pointer lands exactly at $\frac{1}{2}\pi$, however, is 0. ■

6.2.4 Joint cdfs and pdfs

Some chance experiments must be characterized by two or more random variables. The cdf or pdf description is readily extended to such cases. For simplicity, we will consider only the case of two random variables.

To give a specific example, consider the chance experiment in which darts are repeatedly thrown at a target, as shown schematically in Figure 6.8. The point at which the dart lands on the target must be described in terms of two numbers. In this example, we denote the impact point by the two random variables X and Y , whose values are the xy coordinates of the point where the dart sticks, with the origin being fixed at the bull's eye. The joint cdf of X and Y is defined as

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) \quad (6.47)$$

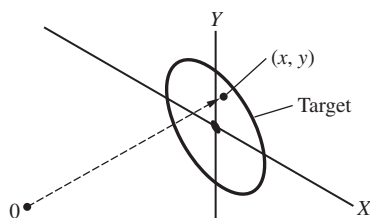


Figure 6.8
The dart-throwing experiment.

where the comma is interpreted as “and.” The *joint pdf* of X and Y is defined as

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} \quad (6.48)$$

Just as we did in the case of single random variables, we can show that

$$P(x_1 < X < x_2, y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x, y) dx dy \quad (6.49)$$

which is the two-dimensional equivalent of (6.42). Letting $x_1 = y_1 = -\infty$ and $x_2 = y_2 = \infty$, we include the entire sample space. Thus,

$$F_{XY}(\infty, \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1 \quad (6.50)$$

Letting $x_1 = x - dx$, $x_2 = x$, $y_1 = y - dy$, and $y_2 = y$, we obtain the following enlightening special case of (6.49):

$$f_{XY}(x, y) dx dy = P(x - dx < X \leq x, y - dy < Y \leq y) \quad (6.51)$$

Thus the probability of finding X in an infinitesimal interval around x while simultaneously finding Y in an infinitesimal interval around y is $f_{XY}(x, y) dx dy$.

Given a joint cdf or pdf, we can obtain the cdf or pdf of one of the random variables using the following considerations. The cdf for X irrespective of the value Y takes on is simply

$$\begin{aligned} F_X(x) &= P(X \leq x, -\infty < Y < \infty) \\ &= F_{XY}(x, \infty) \end{aligned} \quad (6.52)$$

By similar reasoning, the cdf for Y alone is

$$F_Y(y) = F_{XY}(\infty, y) \quad (6.53)$$

$F_X(x)$ and $F_Y(y)$ are referred to as *marginal cdfs*. Using (6.49) and (6.50), we can express (6.52) and (6.53) as

$$F_X(x) = \int_{-\infty}^{\infty} \int_{-\infty}^x f_{XY}(x', y') dx' dy' \quad (6.54)$$

and

$$F_Y(y) = \int_{-\infty}^y \int_{-\infty}^{\infty} f_{XY}(x', y') dx' dy' \quad (6.55)$$

respectively. Since

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{and} \quad f_Y(y) = \frac{dF_Y(y)}{dy} \quad (6.56)$$

we obtain

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y') dy' \quad (6.57)$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x', y) dx' \quad (6.58)$$

from (6.54) and (6.55), respectively. Thus, to obtain the marginal pdfs $f_X(x)$ and $f_Y(y)$ from the joint pdf $f_{XY}(x, y)$, we simply integrate out the undesired variable (or variables for more than two random variables). Hence, the joint cdf or pdf contains all the information possible about the joint random variables X and Y . Similar results hold for more than two random variables.

Two random variables are statistically independent (or simply independent) if the values one takes on do not influence the values that the other takes on. Thus, for any x and y , it must be true that

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \quad (6.59)$$

or, in terms of cdfs,

$$F_{XY}(x, y) = F_X(x)F_Y(y) \quad (6.60)$$

That is, the joint cdf of independent random variables factors into the product of the separate marginal cdfs. Differentiating both sides of (6.59) with respect to first x and then y , and using the definition of the pdf, we obtain

$$f_{XY}(x, y) = f_X(x)f_Y(y) \quad (6.61)$$

which shows that the joint pdf of independent random variables also factors. If two random variables are not independent, we can write their joint pdf in terms of conditional pdfs $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ as

$$\begin{aligned} f_{XY}(x, y) &= f_X(x) f_{Y|X}(y|x) \\ &= f_Y(y) f_{X|Y}(x|y) \end{aligned} \quad (6.62)$$

These relations *define* the conditional pdfs of two random variables. An intuitively satisfying interpretation of $f_{X|Y}(x|y)$ is

$$f_{X|Y}(x|y)dx = P(x - dx < X \leq x \text{ given } Y = y) \quad (6.63)$$

with a similar interpretation for $f_{Y|X}(y|x)$. Equation (6.62) is reasonable in that if X and Y are dependent, a given value of Y should influence the probability distribution for X . On the other hand, if X and Y are independent, information about one of the random variables tells us nothing about the other. Thus, for independent random variables,

$$f_{X|Y}(x|y) = f_X(x) \text{ and } f_{Y|X}(y|x) = f_Y(y), \text{ independent random variables} \quad (6.64)$$

which could serve as an alternative definition of statistical independence. The following example illustrates the preceding ideas.

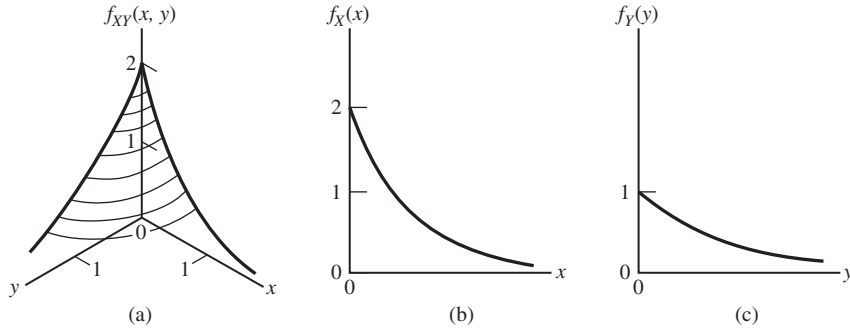
EXAMPLE 6.11

Two random variables X and Y have the joint pdf

$$f_{XY}(x, y) = \begin{cases} Ae^{-(2x+y)}, & x, y \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (6.65)$$

where A is a constant. We evaluate A from

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1 \quad (6.66)$$

**Figure 6.9**

Joint and marginal pdfs for two random variables. (a) Joint pdf. (b) Marginal pdf for X . (c) Marginal pdf for Y .

Since

$$\int_0^{\infty} \int_0^{\infty} e^{-(2x+y)} dx dy = \frac{1}{2} \quad (6.67)$$

$A = 2$. We find the marginal pdfs from (6.51) and (6.52) as follows:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y) dy = \begin{cases} \int_0^{\infty} 2e^{-(2x+y)} dy, & x \geq 0 \\ 0, & x < 0 \end{cases} \\ &= \begin{cases} 2e^{-2x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \end{aligned} \quad (6.68)$$

$$f_Y(y) = \begin{cases} e^{-y}, & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (6.69)$$

These joint and marginal pdfs are shown in Figure 6.9. From these results, we note that X and Y are statistically independent since $f_{XY}(x, y) = f_X(x)f_Y(y)$.

We find the joint cdf by integrating the joint pdf on both variables, using (6.42) and (6.40), which gives

$$\begin{aligned} F_{XY}(x, y) &= \int_{-\infty}^y \int_{-\infty}^x f_{XY}(x', y') dx' dy' \\ &= \begin{cases} (1 - e^{-2x})(1 - e^{-y}), & x, y \geq 0 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (6.70)$$

Dummy variables are used in the integration to avoid confusion. Note that $F_{XY}(-\infty, -\infty) = 0$ and $F_{XY}(\infty, \infty) = 1$, as they should, since the first case corresponds to the probability of an impossible event and the latter corresponds to the inclusion of all possible outcomes. We also can use the result for $F_{XY}(x, y)$ to obtain

$$F_X(x) = F_{XY}(x, \infty) = \begin{cases} (1 - e^{-2x}), & x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (6.71)$$

and

$$F_Y(y) = F_{XY}(\infty, y) = \begin{cases} (1 - e^{-y}), & y \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (6.72)$$

Also note that the joint cdf factors into the product of the marginal cdfs, as it should, for statistically independent random variables.

The conditional pdfs are

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \begin{cases} 2e^{-2x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (6.73)$$

and

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \begin{cases} e^{-y}, & y \geq 0 \\ 0, & y < 0 \end{cases} \quad (6.74)$$

They are equal to the respective marginal pdfs, as they should be for independent random variables. ■

EXAMPLE 6.12

To illustrate the processes of normalization of joint pdfs, finding marginal from joint pdfs, and checking for statistical independence of the corresponding random variables, we consider the joint pdf

$$f_{XY}(x, y) = \begin{cases} \beta xy, & 0 \leq x \leq y, 0 \leq y \leq 4 \\ 0, & \text{otherwise} \end{cases} \quad (6.75)$$

For independence, the joint pdf should be the product of the marginal pdfs.

Solution

This example is somewhat tricky because of the limits, so a diagram of the pdf is given in Figure 6.10. We find the constant β by normalizing the volume under the pdf to unity by integrating $f_{XY}(x, y)$ over all x and y . This gives

$$\begin{aligned} \beta \int_0^4 y \left[\int_0^y x \, dx \right] dy &= \beta \int_0^4 y \frac{y^2}{2} dy \\ &= \beta \frac{y^4}{2 \times 4} \Big|_0^4 \\ &= 32\beta = 1 \end{aligned}$$

so $\beta = \frac{1}{32}$.

We next proceed to find the marginal pdfs. Integrating over x first and checking Figure 6.10 to obtain the proper limits of integration, we obtain

$$\begin{aligned} f_Y(y) &= \int_0^y \frac{xy}{32} dx, \quad 0 \leq y \leq 4 \\ &= \begin{cases} y^3/64, & 0 \leq y \leq 4 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (6.76)$$

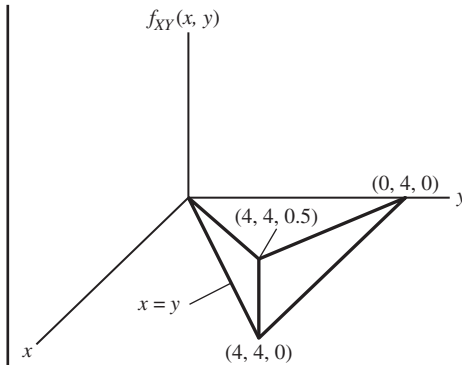


Figure 6.10
Probability-density function for Example 6.12.

The pdf on X is similarly obtained as

$$\begin{aligned}
 f_X(x) &= \int_x^4 \frac{xy}{32} dy, \quad 0 \leq y \leq 4 \\
 &= \begin{cases} (x/4) [1 - (x/4)^2] & 0 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases} \quad (6.77)
 \end{aligned}$$

A little work shows that both marginal pdfs integrate to 1, as they should.

It is clear that the product of the marginal pdfs is not equal to the joint pdf, so the random variables X and Y are not statistically independent. ■

6.2.5 Transformation of Random Variables

Situations are often encountered where the pdf (or cdf) of a random variable X is known and we desire the pdf of a second random variable Y defined as a function of X , for example,

$$Y = g(X) \quad (6.78)$$

We initially consider the case where $g(X)$ is a monotonic function of its argument (for example, it is either nondecreasing or nonincreasing as the independent variable ranges from $-\infty$ to ∞), a restriction that will be relaxed shortly.

A typical function is shown in Figure 6.11. The probability that X lies in the range $(x - dx, x)$ is the same as the probability that Y lies in the range $(y - dy, y)$, where $y = g(x)$.

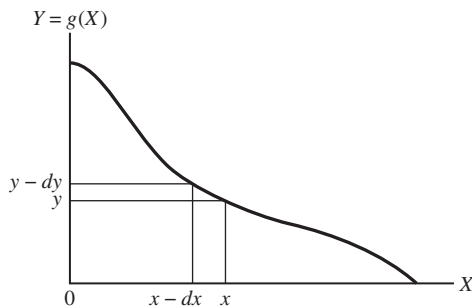


Figure 6.11
A typical monotonic transformation of a random variable.

Using (6.43), we obtain

$$f_X(x) dx = f_Y(y) dy \quad (6.79)$$

if $g(X)$ is monotonically increasing, and

$$f_X(x) dx = -f_Y(y) dy \quad (6.80)$$

if $g(X)$ is monotonically decreasing, since an *increase* in x results in a *decrease* in y . Both cases are taken into account by writing

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|_{x=g^{-1}(y)} \quad (6.81)$$

where $x = g^{-1}(y)$ denotes the inversion of (6.78) for x in terms of y .

EXAMPLE 6.13

To illustrate the use of (6.81), let us consider the pdf of Example 6.10, namely

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{2\pi} & 0 \leq \theta \leq 2\pi \\ 0, & \text{otherwise} \end{cases} \quad (6.82)$$

Assume that the random variable Θ is transformed to the random variable Y according to

$$Y = -\left(\frac{1}{\pi}\right)\Theta + 1 \quad (6.83)$$

Since $\theta = -\pi y + \pi$, $\frac{d\theta}{dy} = -\pi$ and the pdf of Y , by (6.81), is

$$f_Y(y) = f_{\Theta}(\theta = -\pi y + \pi) |-\pi| = \begin{cases} \frac{1}{2} & -1 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (6.84)$$

Note that from (6.83), $\Theta = 2\pi$ gives $Y = -1$ and $\Theta = 0$ gives $Y = 1$, so we would expect the pdf of Y to be nonzero only in the interval $[-1, 1]$; furthermore, since the transformation is linear, it is not surprising that the pdf of Y is uniform as is the pdf of Θ . ■

Consider next the case of $g(x)$ nonmonotonic as illustrated in Figure 6.12. For the case shown, the infinitesimal interval $(y - dy, y)$ corresponds to three infinitesimal intervals on the x -axis: $(x_1 - dx_1, x_1)$, $(x_2 - dx_2, x_2)$, and $(x_3 - dx_3, x_3)$. The probability that X lies in any

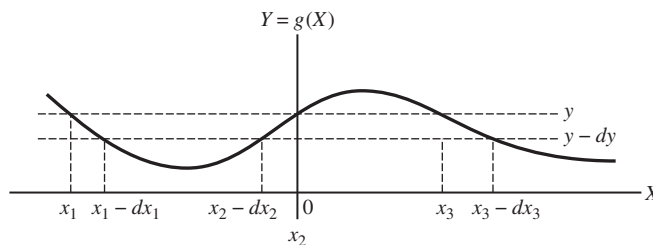


Figure 6.12
A nonmonotonic transformation of a random variable.

one of these intervals is equal to the probability that Y lies in the interval $(y - dy, y)$. This can be generalized to the case of N disjoint intervals where it follows that

$$P(y - dy, y) = \sum_{i=1}^N P(x_i - dx_i, x_i) \quad (6.85)$$

where we have generalized to N intervals on the X axis corresponding to the interval $(y - dy, y)$ on the Y axis. Since

$$P(y - dy, y) = f_Y(y) |dy| \quad (6.86)$$

and

$$P(x_i - dx_i, x_i) = f_X(x_i) |dx_i| \quad (6.87)$$

we have

$$f_Y(y) = \sum_{i=1}^N f_X(x_i) \left| \frac{dx_i}{dy} \right|_{x_i=g_i^{-1}(y)} \quad (6.88)$$

where the absolute value signs are used because a probability must be positive, and $x_i = g_i^{-1}(y)$ is the i th solution to $g(y) = x$.

EXAMPLE 6.14

Consider the transformation

$$y = x^2 \quad (6.89)$$

If $f_X(x) = 0.5 \exp(-|x|)$, find $f_Y(y)$.

Solution

There are two solutions to $x^2 = y$; these are

$$x_1 = \sqrt{y} \quad \text{for } x_1 \geq 0 \quad \text{and} \quad x_2 = -\sqrt{y} \quad \text{for } x_2 < 0, y \geq 0 \quad (6.90)$$

Their derivatives are

$$\frac{dx_1}{dy} = \frac{1}{2\sqrt{y}} \quad \text{for } x_1 \geq 0 \quad \text{and} \quad \frac{dx_2}{dy} = -\frac{1}{2\sqrt{y}} \quad \text{for } x_2 < 0, y > 0 \quad (6.91)$$

Using these results in (6.88), we obtain $f_Y(y)$ to be

$$f_Y(y) = \frac{1}{2} e^{-\sqrt{y}} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{2} e^{-\sqrt{y}} \left| \frac{1}{2\sqrt{y}} \right| = \frac{e^{-\sqrt{y}}}{2\sqrt{y}}, \quad y > 0 \quad (6.92)$$

Since Y cannot be negative, $f_Y(y) = 0$, $y < 0$. ■

For two or more random variables, we consider only one-to-one transformations and the probability of the joint occurrence of random variables lying within infinitesimal areas (or volumes for more than two random variables). Thus, suppose two new random variables U and V are defined in terms of two original, joint random variables X and Y by the relations

$$U = g_1(X, Y) \quad \text{and} \quad V = g_2(X, Y) \quad (6.93)$$

The new pdf $f_{UV}(u, v)$ is obtained from the old pdf $f_{XY}(x, y)$ by using (6.51) to write

$$P(u - du < U \leq u, v - dv < V \leq v) = P(x - dx < X \leq x, y - dy < Y \leq y)$$

or

$$f_{UV}(u, v) dA_{UV} = f_{XY}(x, y) dA_{XY} \quad (6.94)$$

where dA_{UV} is the infinitesimal area in the uv plane corresponding to the infinitesimal area dA_{XY} , in the xy plane through the transformation (6.93).

The ratio of elementary area dA_{XY} to dA_{UV} is given by the Jacobian

$$\frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \quad (6.95)$$

so that

$$f_{UV}(u, v) = f_{XY}(x, y) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|_{\substack{x=g_1^{-1}(u, v) \\ y=g_2^{-1}(u, v)}} \quad (6.96)$$

where the inverse functions $g_1^{-1}(u, v)$ and $g_2^{-1}(u, v)$ exist because the transformations defined by (6.93) are assumed to be one-to-one. An example will help clarify this discussion.

EXAMPLE 6.15

Consider the dart-throwing game discussed in connection with joint cdfs and pdfs. We assume that the joint pdf in terms of rectangular coordinates for the impact point is

$$f_{XY}(x, y) = \frac{\exp[-(x^2 + y^2)/2\sigma^2]}{2\pi\sigma^2}, \quad -\infty < x, y < \infty \quad (6.97)$$

where σ^2 is a constant. This is a special case of the *joint Gaussian pdf*, which we will discuss in more detail shortly.

Instead of cartesian coordinates, we wish to use polar coordinates R and Θ , defined by

$$R = \sqrt{X^2 + Y^2} \quad (6.98)$$

and

$$\Theta = \tan^{-1}\left(\frac{Y}{X}\right) \quad (6.99)$$

so that

$$X = R \cos \Theta = g_1^{-1}(R, \Theta) \quad (6.100)$$

and

$$Y = R \sin \Theta = g_2^{-1}(R, \Theta) \quad (6.101)$$

where

$$0 \leq \Theta < 2\pi, \quad 0 \leq R < \infty$$

so that the whole plane is covered. Under this transformation, the infinitesimal area $dx dy$ in the xy plane transforms to the area $r dr d\theta$ in the $r\theta$ plane, as determined by the Jacobian, which is

$$\frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{vmatrix} = \begin{vmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{vmatrix} = r \quad (6.102)$$

Thus, the joint pdf of R and Θ is

$$f_{R\Theta}(r, \theta) = \frac{r e^{-r^2/2\sigma^2}}{2\pi\sigma^2}, \quad \begin{matrix} 0 \leq \theta < 2\pi \\ 0 \leq r < \infty \end{matrix} \quad (6.103)$$

which follows from (6.96), which for this case takes the form

$$f_{R\Theta}(r, \theta) = r f_{XY}(x, y) \Big|_{\substack{x=r\cos\theta \\ y=r\sin\theta}} \quad (6.104)$$

If we integrate $f_{R\Theta}(r, \theta)$ over θ to get the pdf for R alone, we obtain

$$f_R(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad 0 \leq r < \infty \quad (6.105)$$

which is referred to as the *Rayleigh pdf*. The probability that the dart lands in a ring of radius r from the bull's-eye and having thickness dr is given by $f_R(r) dr$. From the sketch of the Rayleigh pdf given in Figure 6.13, we see that the most probable distance for the dart to land from the bull's-eye is $R = \sigma$. By integrating (6.105) over r , it can be shown that the pdf of Θ is uniform in $[0, 2\pi)$.

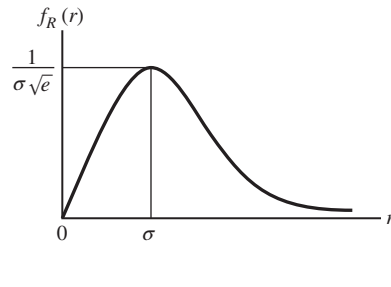


Figure 6.13
The Rayleigh pdf.

6.3 STATISTICAL AVERAGES

The probability functions (cdf and pdf) we have just discussed provide us with all the information possible about a random variable or a set of random variables. Often, such complete descriptions as provided by the pdf or cdf are not required, or in many cases, we are not able to obtain the cdf or pdf. A partial description of a random variable or set of random variables is then used and is given in terms of various statistical averages or mean values.

6.3.1 Average of a Discrete Random Variable

The statistical average, or expectation, of a discrete random variable X , which takes on the possible values x_1, x_2, \dots, x_M with the respective probabilities P_1, P_2, \dots, P_M , is defined as

$$\bar{X} = E[X] = \sum_{j=1}^M x_j P_j \quad (6.106)$$

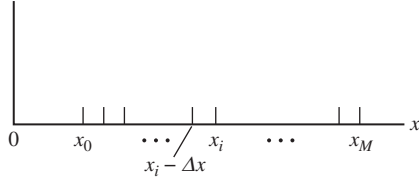


Figure 6.14
A discrete approximation for a continuous random variable X .

To show the reasonableness of this definition, we look at it in terms of relative frequency. If the underlying chance experiment is repeated a large number of times N , and $X = x_1$ is observed n_1 times and $X = x_2$ is observed n_2 times, etc., the arithmetical average of the observed values is

$$\frac{n_1 x_1 + n_2 x_2 + \cdots + n_M x_M}{N} = \sum_{j=1}^M x_j \frac{n_j}{N} \quad (6.107)$$

But, by the relative-frequency interpretation of probability, (6.2), n_j/N approaches P_j , $j = 1, 2, \dots, M$, the probability of the event $X = x_j$, as N becomes large. Thus, in the limit as $N \rightarrow \infty$, (6.107) becomes (6.106).

6.3.2 Average of a Continuous Random Variable

For the case where X is a continuous random variable with the pdf $f_X(x)$, we consider the range of values that X may take on, say x_0 to x_M , to be broken up into a large number of small subintervals of length Δx , as shown in Figure 6.14.

For example, consider a discrete approximation for finding the expectation of a continuous random variable X . The probability that X lies between $x_i - \Delta x$ and x_i is, from (6.43), given by

$$P(x_i - \Delta x < X \leq x_i) \cong f_X(x_i) \Delta x, \quad i = 1, 2, \dots, M \quad (6.108)$$

for Δx small. Thus, we have approximated X by a discrete random variable that takes on the values x_0, x_1, \dots, x_M with probabilities $f_X(x_0) \Delta x, \dots, f_X(x_M) \Delta x$, respectively. Using (6.106) the expectation of this random variable is

$$E[X] \cong \sum_{i=0}^M x_i f_X(x_i) \Delta x \quad (6.109)$$

As $\Delta x \rightarrow 0$, this becomes a better and better approximation for $E[X]$. In the limit, as $\Delta x \rightarrow dx$, the sum becomes an integral, giving

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad (6.110)$$

for the expectation of X .

6.3.3 Average of a Function of a Random Variable

We are interested not only in $E[X]$, which is referred to as the *mean* or *first moment* of X , but also in statistical averages of functions of X . Letting $Y = g(X)$, the statistical average or

expectation of the new random variable Y could be obtained as

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y) dy \quad (6.111)$$

where $f_Y(y)$ is the pdf of Y , which can be found from $f_X(x)$ by application of (6.81). However, it is often more convenient simply to find the expectation of the function $g(X)$ as given by

$$\overline{g(X)} \triangleq E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \quad (6.112)$$

which is identical to $E[Y]$ as given by (6.111). Two examples follow to illustrate the use of (6.111) and (6.112).

EXAMPLE 6.16

Suppose the random variable Θ has the pdf

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{2\pi}, & |\theta| \leq \pi \\ 0, & \text{otherwise} \end{cases} \quad (6.113)$$

Then $E[\Theta^n]$ is referred to as the n th moment of Θ and is given by

$$E[\Theta^n] = \int_{-\infty}^{\infty} \theta^n f_{\Theta}(\theta) d\theta = \int_{-\pi}^{\pi} \theta^n \frac{d\theta}{2\pi} \quad (6.114)$$

Since the integrand is odd if n is odd, $E[\Theta^n] = 0$ for n odd. For n even,

$$E[\Theta^n] = \frac{1}{\pi} \int_0^{\pi} \theta^n d\theta = \frac{1}{\pi} \frac{\theta^{n+1}}{n+1} \Big|_0^{\pi} = \frac{\pi^n}{n+1} \quad (6.115)$$

The first moment or mean of Θ , $E[\Theta]$, is a measure of the location of $f_{\Theta}(\theta)$ (that is, the ‘‘center of mass’’). Since $f_{\Theta}(\theta)$ is symmetrically located about $\theta = 0$, it is not surprising that $E[\Theta] = 0$. ■

EXAMPLE 6.17

Later we shall consider certain random waveforms that can be modeled as sinusoids with random phase angles having uniform pdf in $[-\pi, \pi)$. In this example, we consider a random variable X that is defined in terms of the uniform random variable Θ considered in Example 6.17 by

$$X = \cos \Theta \quad (6.116)$$

The density function of X , $f_X(x)$, is found as follows. First, $-1 \leq \cos \theta \leq 1$, so $f_X(x) = 0$ for $|x| > 1$. Second, the transformation is not one-to-one, there being two values of Θ for each value of X , since $\cos \theta = \cos(-\theta)$. However, we can still apply (6.81) by noting that positive and negative angles have equal probabilities and writing

$$f_X(x) = 2f_{\Theta}(\theta) \left| \frac{d\theta}{dx} \right|, \quad |x| < 1 \quad (6.117)$$

Now $\theta = \cos^{-1} x$ and $|d\theta/dx| = (1 - x^2)^{-1/2}$, which yields

$$f_X(x) = \begin{cases} \frac{1}{\pi\sqrt{1-x^2}} & |x| < 1 \\ 0, & |x| > 1 \end{cases} \quad (6.118)$$

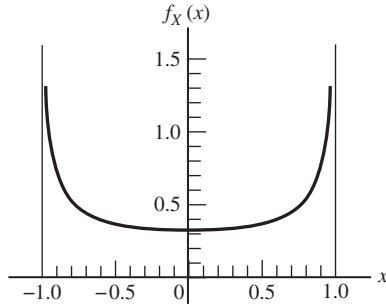


Figure 6.15
Probability-density function of a sinusoid with uniform random phase.

This pdf is illustrated in Figure 6.15. The mean and second moment of X can be calculated using either (6.111) or (6.112). Using (6.111), we obtain

$$\bar{X} = \int_{-1}^1 \frac{x}{\pi\sqrt{1-x^2}} dx = 0 \quad (6.119)$$

because the integrand is odd, and

$$\overline{X^2} = \int_{-1}^1 \frac{x^2 dx}{\pi\sqrt{1-x^2}} = \frac{1}{2} \quad (6.120)$$

by a table of integrals. Using (6.112), we find that

$$\bar{X} = \int_{-\pi}^{\pi} \cos\theta \frac{d\theta}{2\pi} = 0 \quad (6.121)$$

and

$$\overline{X^2} = \int_{-\pi}^{\pi} \cos^2\theta \frac{d\theta}{2\pi} = \int_{-\pi}^{\pi} \frac{1}{2} (1 + \cos 2\theta) \frac{d\theta}{2\pi} = \frac{1}{2} \quad (6.122)$$

as obtained by finding $E[X]$ and $E[X^2]$ directly. ■

6.3.4 Average of a Function of More Than One Random Variable

The expectation of a function $g(X, Y)$ of two random variables X and Y is defined in a manner analogous to the case of a single random variable. If $f_{XY}(x, y)$ is the joint pdf of X and Y , the expectation of $g(X, Y)$ is

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy \quad (6.123)$$

The generalization to more than two random variables should be obvious.

Equation (6.123) and its generalization to more than two random variables include the single-random-variable case, for suppose $g(X, Y)$ is replaced by a function of X alone, say $h(X)$. Then using (6.57) we obtain the following from (6.123):

$$\begin{aligned} E[h(X)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x) f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} h(x) f_X(x) dx \end{aligned} \quad (6.124)$$

where the fact that $\int_{-\infty}^{\infty} f_{XY}(x, y) dy = f_X(x)$ has been used.

EXAMPLE 6.18

Consider the joint pdf of Example 6.11 and the expectation of $g(X, Y) = XY$. From (6.123), this expectation is

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy \\ &= \int_0^{\infty} \int_0^{\infty} 2xy e^{-(2x+y)} dx dy \\ &= 2 \int_0^{\infty} x e^{-2x} dx \int_0^{\infty} y e^{-y} dy = \frac{1}{2} \end{aligned} \quad (6.125)$$

We recall from Example 6.11 that X and Y are statistically independent. From the last line of the preceding equation for $E[XY]$, we see that

$$E[XY] = E[X]E[Y] \quad (6.126)$$

a result that holds in general for statistically independent random variables. In fact, for statistically independent random variables, it readily follows that

$$E[h(X)g(Y)] = E[h(X)]E[g(Y)], \quad X \text{ and } Y \text{ statistically independent} \quad (6.127)$$

where $h(X)$ and $g(Y)$ are two functions of X and Y , respectively.

In the special case where $h(X) = X^m$ and $g(Y) = Y^n$, and X and Y are not statistically independent in general, the expectations $E[X^m Y^n]$ are referred to as the *joint moments* of order $m + n$ of X and Y . According to (6.127), the *joint moments of statistically independent random variables factor into the products of the corresponding marginal moments*.

When finding the expectation of a function of more than one random variable, it may be easier to use the concept of conditional expectation. Consider, for example, a function $g(X, Y)$ of two random variables X and Y , with the joint pdf $f_{XY}(x, y)$. The expectation of $g(X, Y)$ is

$$\begin{aligned} E[g(X, Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx \right] f_Y(y) dy \\ &= E \{ E [g(X, Y) | Y] \} \end{aligned} \quad (6.128)$$

where $f_{X|Y}(x|y)$ is the conditional pdf of X given Y , and $E[g(X, Y)|Y] = \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x|y) dx$ is called the *conditional expectation of $g(X, Y)$ given $Y = y$* .

EXAMPLE 6.19

As a specific application of conditional expectation, consider the firing of projectiles at a target. Projectiles are fired until the target is hit for the first time, after which firing ceases. Assume that the probability of a projectile's hitting the target is p and that the firings are independent of one another. Find the average number of projectiles fired at the target.

Solution

To solve this problem, let N be a random variable denoting the number of projectiles fired at the target. Let the random variable H be 1 if the first projectile hits the target and 0 if it does not. Using the concept of conditional expectation, we find the average value of N is given by

$$\begin{aligned} E[N] &= E\{E[N|H]\} = pE[N|H = 1] + (1 - p)E[N|H = 0] \\ &= p \times 1 + (1 - p)(1 + E[N]) \end{aligned} \quad (6.129)$$

where $E[N|H = 0] = 1 + E[N]$ because $N \geq 1$ if a miss occurs on the first firing. By solving the last expression for $E[N]$, we obtain

$$E[N] = \frac{1}{p} \quad (6.130)$$

If $E[N]$ is evaluated directly, it is necessary to sum the series:

$$E[N] = 1 \times p + 2 \times (1 - p)p + 3 \times (1 - p)^2 p + \dots \quad (6.131)$$

which is not too difficult in this instance.³ However, the conditional-expectation method clearly makes it easier to keep track of the bookkeeping. ■

6.3.5 Variance of a Random Variable

The statistical average

$$\sigma_x^2 \triangleq E\{[X - E(X)]^2\} \quad (6.132)$$

is called the *variance* of the random variable X ; σ_x is called the *standard deviation* of X and is a measure of the concentration of the pdf of X , or $f_X(x)$, about the mean. The notation $\text{var}\{X\}$ for σ_x^2 is sometimes used. A useful relation for obtaining σ_x^2 is

$$\sigma_x^2 = E[X^2] - E^2[X] \quad (6.133)$$

³Consider $E[N] = p(1 + 2q + 3q^2 + 4q^3 + \dots)$ where $q = 1 - p$. The sum $S = 1 + q + q^2 + q^3 + \dots = \frac{1}{1-q}$ can be used to derive the sum of $1 + 2q + 3q^2 + 4q^3 + \dots$ by differentiation with respect to q : $\frac{dS}{dq} = 1 + 2q + 3q^2 + \dots = \frac{d}{dq} \frac{1}{1-q} = \frac{1}{(1-q)^2}$ so that $E[N] = p \frac{1}{(1-q)^2} = \frac{1}{p}$.

which, in words, says that the variance of X is simply its second moment minus its mean, squared. To prove (6.133), let $E[X] = m_x$. Then

$$\begin{aligned}\sigma_x^2 &= \int_{-\infty}^{\infty} (x - m_x)^2 f_X(x) dx = \int_{-\infty}^{\infty} (x^2 - 2xm_x + m_x^2) f_X(x) dx \\ &= E[X^2] - 2m_x^2 + m_x^2 = E[X^2] - E^2[X]\end{aligned}\quad (6.134)$$

which follows because $\int_{-\infty}^{\infty} x f_X(x) dx = m_x$.

EXAMPLE 6.20

Let X have the uniform pdf

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}\quad (6.135)$$

Then

$$E[X] = \int_a^b x \frac{dx}{b-a} = \frac{1}{2}(a+b)\quad (6.136)$$

and

$$E[X^2] = \int_a^b x^2 \frac{dx}{b-a} = \frac{1}{3}(b^2 + ab + a^2)\quad (6.137)$$

which follows after a little work. Thus,

$$\sigma_x^2 = \frac{1}{3}(b^2 + ab + a^2) - \frac{1}{4}(a^2 + 2ab + b^2) = \frac{1}{12}(a-b)^2\quad (6.138)$$

Consider the following special cases:

1. $a = 1$ and $b = 2$, for which $\sigma_x^2 = \frac{1}{12}$.
2. $a = 0$ and $b = 1$, for which $\sigma_x^2 = \frac{1}{12}$.
3. $a = 0$ and $b = 2$, for which $\sigma_x^2 = \frac{1}{3}$.

For cases 1 and 2, the pdf of X has the same width but is centered about different means; the variance is the same for both cases. In case 3, the pdf is wider than it is for cases 1 and 2, which is manifested by the larger variance. ■

6.3.6 Average of a Linear Combination of N Random Variables

It is easily shown that the expected value, or average, of an arbitrary linear combination of random variables is the same as the linear combination of their respective means. That is,

$$E\left[\sum_{i=1}^N a_i X_i\right] = \sum_{i=1}^N a_i E[X_i]\quad (6.139)$$

where X_1, X_2, \dots, X_N are random variables and a_1, a_2, \dots, a_N are arbitrary constants. Equation (6.139) will be demonstrated for the special case $N = 2$; generalization to the case $N > 2$ is not difficult, but results in unwieldy notation (proof by induction can also be used).

Let $f_{X_1 X_2}(x_1, x_2)$ be the joint pdf of X_1 and X_2 . Then, using the definition of the expectation of a function of two random variables in (6.123), it follows that

$$\begin{aligned} E[a_1 X_1 + a_2 X_2] &\triangleq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (a_1 x_1 + a_2 x_2) f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \\ &= a_1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \\ &\quad + a_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_2 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 \end{aligned} \quad (6.140)$$

Considering the first double integral and using (6.57) (with $x_1 = x$ and $x_2 = y$) and (6.110), we find that

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1 X_2}(x_1, x_2) dx_1 dx_2 &= \int_{-\infty}^{\infty} x_1 \left\{ \int_{-\infty}^{\infty} f_{X_1 X_2}(x_1, x_2) dx_2 \right\} dx_1 \\ &= \int_{-\infty}^{\infty} x_1 f_X(x_1) dx_1 \\ &= E\{X_1\} \end{aligned} \quad (6.141)$$

Similarly, it can be shown that the second double integral reduces to $E[X_2]$. Thus, (6.139) has been proved for the case $N = 2$. Note that (6.139) holds regardless of whether or not the X_i terms are independent. Also, it should be noted that a similar result holds for a linear combination of functions of N random variables.

6.3.7 Variance of a Linear Combination of Independent Random Variables

If X_1, X_2, \dots, X_N are statistically independent random variables, then

$$\text{var} \left[\sum_{i=1}^N a_i X_i \right] = \sum_{i=1}^N a_i^2 \text{var} \{ X_i \} \quad (6.142)$$

where a_1, a_2, \dots, a_N are arbitrary constants and $\text{var}[X_i] \triangleq E[(X_i - \bar{X}_i)^2]$. This relation will be demonstrated for the case $N = 2$. Let $Z = a_1 X_1 + a_2 X_2$ and let $f_{X_i}(x_i)$ be the marginal pdf of X_i . Then the joint pdf of X_1 and X_2 is $f_{X_1}(x_1) f_{X_2}(x_2)$ by the assumption of statistical independence. Also, $E[Z] = a_1 E[X_1] + a_2 E[X_2] \triangleq a_1 \bar{X}_1 + a_2 \bar{X}_2$ by (6.139), and $\text{var}[Z] = E[(Z - \bar{Z})^2]$. But, since $Z = a_1 X_1 + a_2 X_2$, we may write $\text{var}[Z]$ as

$$\begin{aligned} \text{var}[Z] &= E \left\{ [(a_1 X_1 + a_2 X_2) - (a_1 \bar{X}_1 + a_2 \bar{X}_2)]^2 \right\} \\ &= E \left\{ [a_1(X_1 - \bar{X}_1) + a_2(X_2 - \bar{X}_2)]^2 \right\} \\ &= a_1^2 E[(X_1 - \bar{X}_1)^2] + 2a_1 a_2 E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)] \\ &\quad + a_2^2 E[(X_2 - \bar{X}_2)^2] \end{aligned} \quad (6.143)$$

The first and last terms in the preceding equation are $a_1^2 \text{var}[X_1]$ and $a_2^2 \text{var}[X_2]$, respectively. The middle term is zero, since

$$\begin{aligned} & E[(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - \bar{X}_1)(x_2 - \bar{X}_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} (x_1 - \bar{X}_1) f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty} (x_2 - \bar{X}_2) f_{X_2}(x_2) dx_2 \\ &= (\bar{X}_1 - \bar{X}_1) (\bar{X}_2 - \bar{X}_2) = 0 \end{aligned} \quad (6.144)$$

Note that the assumption of *statistical independence* was used to show that the middle term above is zero (it is a sufficient, but not necessary, condition).

6.3.8 Another Special Average—The Characteristic Function

Letting $g(X) = e^{jvX}$ in (6.112), we obtain an average known as the *characteristic function* of X , or $M_X(jv)$, defined as

$$M_X(jv) \triangleq E[e^{jvX}] = \int_{-\infty}^{\infty} f_X(x) e^{jvx} dx \quad (6.145)$$

It is seen that $M_X(jv)$ would be the *Fourier transform* of $f_X(x)$, as we have defined the Fourier transform in Chapter 2, provided a minus sign has been used in the exponent instead of a plus sign. Thus, if $j\omega$ is replaced by $-jv$ in Fourier transform tables, they can be used to obtain characteristic functions from pdfs (sometimes it is convenient to use the variable s in place of jv ; the resulting function is called the *moment generating function*).

A pdf is obtained from the corresponding characteristic function by the inverse transform relationship

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} M_X(jv) e^{-jvx} dv \quad (6.146)$$

This illustrates one possible use of the characteristic function. It is sometimes easier to obtain the characteristic function than the pdf, and the latter is then obtained by inverse Fourier transformation, either analytically or numerically.

Another use for the characteristic function is to obtain the moments of a random variable. Consider the differentiation of (6.145) with respect to v . This gives

$$\frac{\partial M_X(jv)}{\partial v} = j \int_{-\infty}^{\infty} x f_X(x) e^{jvx} dx \quad (6.147)$$

Setting $v = 0$ after differentiation and dividing by j , we obtain

$$E[X] = (-j) \left. \frac{\partial M_X(jv)}{\partial v} \right|_{v=0} \quad (6.148)$$

For the n th moment, the relation

$$E[X^n] = (-j)^n \left. \frac{\partial^n M_X(jv)}{\partial v^n} \right|_{v=0} \quad (6.149)$$

can be proved by repeated differentiation.

EXAMPLE 6.21

By use a table of Fourier transforms, the one-sided exponential pdf

$$f_X(x) = \exp(-x)u(x) \quad (6.150)$$

is found to have the characteristic function

$$M_X(jv) = \int_0^{\infty} e^{-x} e^{jvx} dx = \frac{1}{1 - jv} \quad (6.151)$$

By repeated differentiation or expansion of the characteristic function in a power series in jv , it follows from (6.149) that $E\{X^n\} = n!$ for this random variable. ■

6.3.9 The pdf of the Sum of Two Independent Random Variables

Given two *statistically independent* random variables X and Y with known pdfs $f_X(x)$ and $f_Y(y)$, respectively, the pdf of their sum $Z = X + Y$ is often of interest. The characteristic function will be used to find the pdf of Z , or $f_Z(z)$, even though we could find the pdf of Z directly.

From the definition of the characteristic function of Z , we write

$$\begin{aligned} M_Z(jv) &= E[e^{jvZ}] = E[e^{jv(X+Y)}] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{jv(x+y)} f_X(x) f_Y(y) dx dy \end{aligned} \quad (6.152)$$

since the joint pdf of X and Y is $f_X(x)f_Y(y)$ by the assumption of statistical independence of X and Y . We can write (6.152) as the product of two integrals, since $e^{jv(x+y)} = e^{jvx}e^{jvy}$. This results in

$$\begin{aligned} M_Z(jv) &= \int_{-\infty}^{\infty} f_X(x) e^{jvx} dx \int_{-\infty}^{\infty} f_Y(y) e^{jvy} dy \\ &= E[e^{jvX}] E[e^{jvY}] \end{aligned} \quad (6.153)$$

From the definition of the characteristic function, given by (6.145), we see that

$$M_Z(jv) = M_X(jv) M_Y(jv) \quad (6.154)$$

where $M_X(jv)$ and $M_Y(jv)$ are the characteristic functions of X and Y , respectively. Remembering that the characteristic function is the Fourier transform of the corresponding pdf and that a product in the frequency domain corresponds to convolution in the time domain, it follows that

$$f_Z(z) = f_X(x) * f_Y(y) = \int_{-\infty}^{\infty} f_X(z-u) f_Y(u) du \quad (6.155)$$

The following example illustrates the use of (6.155).

EXAMPLE 6.22

Consider the sum of four identically distributed, independent random variables,

$$Z = X_1 + X_2 + X_3 + X_4 \quad (6.156)$$

where the pdf of each X_i is

$$f_{X_i}(x_i) = \Pi(x_i) = \begin{cases} 1, & |x_i| \leq \frac{1}{2} \\ 0, & \text{otherwise, } i = 1, 2, 3, 4 \end{cases} \quad (6.157)$$

where $\Pi(x_i)$ is the unit rectangular pulse function defined in Chapter 2. We find $f_Z(z)$ by applying (6.155) twice. Thus, let

$$Z_1 = X_1 + X_2 \quad \text{and} \quad Z_2 = X_3 + X_4 \quad (6.158)$$

The pdfs of Z_1 and Z_2 are identical, both being the convolution of a uniform density with itself. From Table 2.2, we can immediately write down the result:

$$f_{Z_i}(z_i) = \Lambda(z_i) = \begin{cases} 1 - |z_i|, & |z_i| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (6.159)$$

where $f_{Z_i}(z_i)$ is the pdf of $Z_i, i = 1, 2$. To find $f_Z(z)$, we simply convolve $f_{Z_i}(z_i)$ with itself. Thus,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z_i}(z-u)f_{Z_i}(u) du \quad (6.160)$$

The factors in the integrand are sketched in Figure 6.16(a). Clearly, $f_Z(z) = 0$ for $z < -2$ or $z > 2$. Since $f_{Z_i}(z_i)$ is even, $f_Z(z)$ is also even. Thus, we need not consider $f_Z(z)$ for $z < 0$. From Figure 6.16(a) it follows that for $1 \leq z \leq 2$,

$$f_Z(z) = \int_{z-1}^1 (1-u)(1+u-z) du = \frac{1}{6}(2-z)^3 \quad (6.161)$$

and for $0 \leq z \leq 1$, we obtain

$$\begin{aligned} f_Z(z) &= \int_{z-1}^0 (1+u)(1+u-z) du + \int_0^z (1-u)(1+u-z) du \\ &\quad + \int_z^1 (1-u)(1-u+z) du \\ &= (1-z) - \frac{1}{3}(1-z)^3 + \frac{1}{6}z^3 \end{aligned} \quad (6.162)$$

A graph of $f_Z(z)$ is shown in Figure 6.16(b) along with the graph of the function

$$\frac{\exp\left(-\frac{3}{2}z^2\right)}{\sqrt{\frac{2}{3}\pi}} \quad (6.163)$$

which represents a marginal Gaussian pdf of mean 0 and variance $\frac{1}{3}$, the same variance as $Z = X_1 + X_2 + X_3 + X_4$ [the results of Example (6.20) and Equation (6.142) can be used to obtain the variance of Z]. We will describe the Gaussian pdf more fully later.

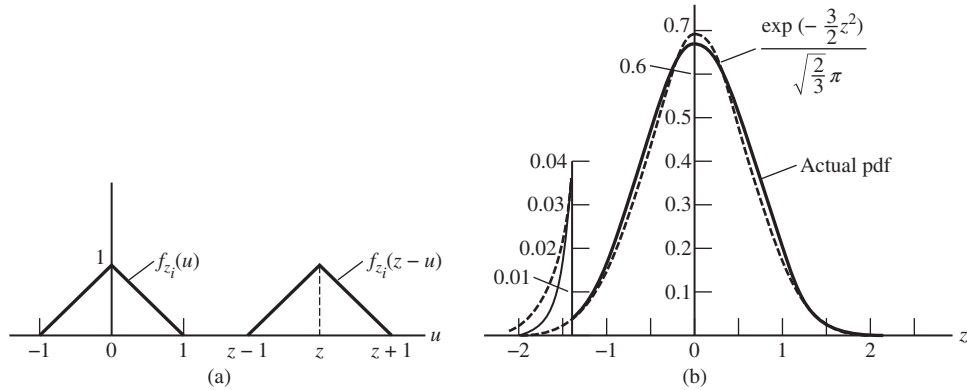


Figure 6.16 The pdf for the sum of four independent uniformly distributed random variables. (a) Convolution of two triangular pdfs. (b) Comparison of actual and Gaussian pdfs.

The reason for the striking similarity of the two pdfs shown in Figure 6.16(b) will become apparent when the central-limit theorem is discussed in Section 6.4.5.

6.3.10 Covariance and the Correlation Coefficient

Two useful joint averages of a pair of random variables X and Y are their covariance μ_{XY} , defined as

$$\mu_{XY} = E[(X - \bar{X})(Y - \bar{Y})] = E[XY] - E[X]E[Y] \tag{6.164}$$

and their correlation coefficient ρ_{XY} , which is written in terms of the covariance as

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y} \tag{6.165}$$

From the preceding two expressions we have the relationship

$$E[XY] = \sigma_X \sigma_Y \rho_{XY} + E[X]E[Y] \tag{6.166}$$

Both μ_{XY} and ρ_{XY} are measures of the interdependence of X and Y . The correlation coefficient is more convenient because it is normalized such that $-1 \leq \rho_{XY} \leq 1$. If $\rho_{XY} = 0$, X and Y are said to be *uncorrelated*.

It is easily shown that $\rho_{XY} = 0$ for statistically independent random variables. If X and Y are independent, their joint pdf $f_{XY}(x, y)$ is the product of the respective marginal pdfs; that is, $f_{XY}(x, y) = f_X(x)f_Y(y)$. Thus,

$$\begin{aligned} \mu_{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{X})(y - \bar{Y}) f_X(x)f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} (x - \bar{X}) f_X(x) dx \int_{-\infty}^{\infty} (y - \bar{Y}) f_Y(y) dy \\ &= (\bar{X} - \bar{X})(\bar{Y} - \bar{Y}) = 0 \end{aligned} \tag{6.167}$$

Considering next the cases $X = \pm\alpha Y$, so that $\bar{X} = \pm\alpha\bar{Y}$, where α is a positive constant, we obtain

$$\begin{aligned}\mu_{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\pm\alpha y \mp \alpha\bar{Y})(y - \bar{Y})f_{XY}(x, y) dx dy \\ &= \pm\alpha \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \bar{Y})^2 f_{XY}(x, y) dx dy \\ &= \pm\alpha\sigma_Y^2\end{aligned}\tag{6.168}$$

Using (6.142) with $N = 1$, we can write the variance of X as $\sigma_X^2 = \alpha^2\sigma_Y^2$. Thus, the correlation coefficient is

$$\rho_{XY} = +1 \text{ for } X = +\alpha Y, \quad \text{and} \quad \rho_{XY} = -1 \text{ for } X = -\alpha Y$$

To summarize, the correlation coefficient of two independent random variables is zero. When two random variables are linearly related, their correlation is $+1$ or -1 depending on whether one is a positive or a negative constant times the other.

6.4 SOME USEFUL PDFS

We have already considered several often used probability distributions in the examples.⁴ These have included the Rayleigh pdf (Example 6.15), the pdf of a sinewave of random phase (Example 6.17), and the uniform pdf (Example 6.20). Some others, which will be useful in our future considerations, are given below.

6.4.1 Binomial Distribution

One of the most common discrete distributions in the application of probability to systems analysis is the binomial distribution. We consider a chance experiment with two mutually exclusive, exhaustive outcomes A and \bar{A} , where \bar{A} denotes the compliment of A , with probabilities $P(A) = p$ and $P(\bar{A}) = q = 1 - p$, respectively. Assigning the discrete random variable K to be numerically equal to the number of times event A occurs in n trials of our chance experiment, we seek the probability that exactly $k \leq n$ occurrences of the event A occur in n repetitions of the experiment. (Thus, our actual chance experiment is the replication of the basic experiment n times.) The resulting distribution is called the *binomial distribution*.

Specific examples in which the binomial distribution is the result are the following: In n tosses of a coin, what is the probability of $k \leq n$ heads? In the transmission of n messages through a channel, what is the probability of $k \leq n$ errors? Note that in all cases we are interested in exactly k occurrences of the event, not, for example, at least k of them, although we may find the latter probability if we have the former.

Although the problem being considered is very general, we solve it by visualizing the coin-tossing experiment. We wish to obtain the probability of k heads in n tosses of the coin if the probability of a head on a single toss is p and the probability of a tail is $1 - p = q$. One

⁴Useful probability distributions are summarized in Table 6.4 at the end of this chapter.

of the possible sequences of k heads in n tosses is

$$\underbrace{H H \cdots H}_{k \text{ heads}} \underbrace{T T \cdots T}_{n-k \text{ tails}}$$

Since the trials are independent, the probability of this particular sequence is

$$\underbrace{p \cdot p \cdot p \cdots p}_{k \text{ factors}} \cdot \underbrace{q \cdot q \cdot q \cdots q}_{n-k \text{ factors}} = p^k q^{n-k} \quad (6.169)$$

But the preceding sequence of k heads in n trials is only one of

$$\binom{n}{k} \triangleq \frac{n!}{k!(n-k)!} \quad (6.170)$$

possible sequences, where $\binom{n}{k}$ is the binomial coefficient. To see this, we consider the number of ways k identifiable heads can be arranged in n slots. The first head can fall in any of the n slots, the second in any of $n-1$ slots (the first head already occupies one slot), the third in any of $n-2$ slots, and so on for a total of

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} \quad (6.171)$$

possible arrangements in which each head is identified. However, we are not concerned about which head occupies which slot. For each possible identifiable arrangement, there are $k!$ arrangements for which the heads can be switched around and with the same slots occupied. Thus, the total number of arrangements, if we do not identify the particular head occupying each slot, is

$$\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} \quad (6.172)$$

Since the occurrence of any of these $\binom{n}{k}$ possible arrangements precludes the occurrence of any other [that is, the $\binom{n}{k}$ outcomes of the experiment are mutually exclusive], and since each occurs with probability $p^k q^{n-k}$, the probability of exactly k heads in n trials in any order is

$$P(K = k) \triangleq P_n(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \dots, n \quad (6.173)$$

Equation (6.173), known as the *binomial probability distribution* (note that it is not a pdf or a cdf), is plotted in Figure 6.17(a)–(d) for four different values of p and n .

The mean of a binomially distributed random variable K , by Equation (6.109), is given by

$$E[K] = \sum_{k=0}^n k \frac{n!}{k!(n-k)!} p^k q^{n-k} \quad (6.174)$$

Noting that the sum can be started at $k = 1$ since the first term is zero, we can write

$$E[K] = \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \quad (6.175)$$

where the relation $k! = k(k-1)!$ has been used. Letting $m = k - 1$, we get the sum

$$\begin{aligned} E[K] &= \sum_{m=0}^{n-1} \frac{n!}{m!(n-m-1)!} p^{m+1} q^{n-m-1} \\ &= np \sum_{m=0}^{n-1} \frac{(n-1)!}{m!(n-m-1)!} p^m q^{n-m-1} \end{aligned} \quad (6.176)$$

Finally, letting $\ell = n - 1$ and recalling that, by the binomial theorem,

$$(x + y)^\ell = \sum_{m=0}^{\ell} \binom{\ell}{m} x^m y^{\ell-m} \quad (6.177)$$

we obtain

$$\bar{K} = E[K] = np(p+q)^\ell = np \quad (6.178)$$

since $p + q = 1$. The result is reasonable; in a long sequence of n tosses of a fair coin ($p = q = \frac{1}{2}$), we would expect about $np = \frac{1}{2}n$ heads.

We can go through a similar series of manipulations to show that $E[K^2] = np(np+q)$. Using this result, it follows that the variance of a binomially distributed random variable is

$$\sigma_K^2 = E[K^2] - E^2[K] = npq = \bar{K}(1-p) \quad (6.179)$$

EXAMPLE 6.23

The probability of having two girls in a four-child family, assuming single births and equal probabilities of male and female births, from (6.173), is

$$P_4(2) = \binom{4}{2} \left(\frac{1}{2}\right)^4 = \frac{3}{8} \quad (6.180)$$

Similarly, it can be shown that the probability of 0, 1, 3, and 4 girls is $\frac{1}{16}$, $\frac{1}{4}$, $\frac{1}{4}$, and $\frac{1}{16}$, respectively. Note that the sum of the probabilities for 0, 1, 2, 3, and 4 girls (or boys) is 1, as it should be. ■

6.4.2 Laplace Approximation to the Binomial Distribution

When n becomes large, computations using (6.173) become unmanageable. In the limit as $n \rightarrow \infty$, it can be shown that for $|k - np| \leq \sqrt{npq}$

$$P_n(k) \cong \frac{1}{\sqrt{2\pi npq}} \exp \left[-\frac{(k - np)^2}{2npq} \right] \quad (6.181)$$

which is called the Laplace approximation to the binomial distribution. A comparison of the Laplace approximation with the actual binomial distribution is given in Figure 6.17(e).

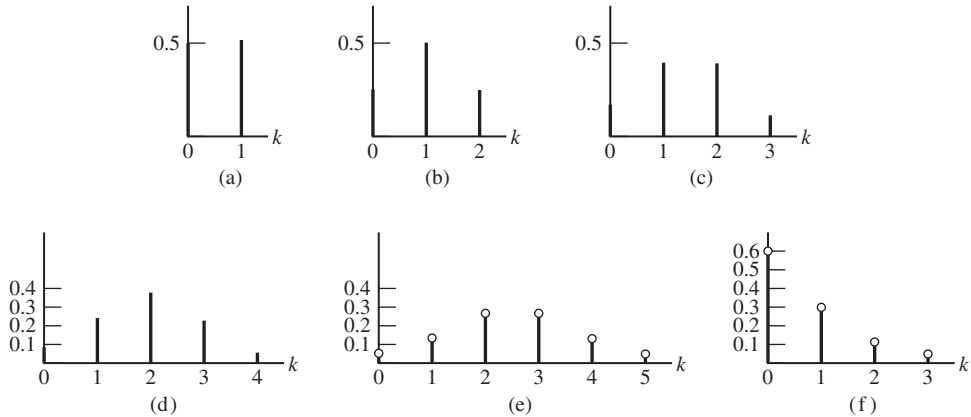


Figure 6.17

The binomial distribution with comparison to Laplace and Poisson approximations. (a) $n = 1$, $p = 0.5$. (b) $n = 2$, $p = 0.5$. (c) $n = 3$, $p = 0.5$. (d) $n = 4$, $p = 0.5$. (e) $n = 5$, $p = 0.5$. Circles are Laplace approximations. (f) $n = 5$, $p = \frac{1}{10}$. Circles are Poisson approximations.

6.4.3 Poisson Distribution and Poisson Approximation to the Binomial Distribution

Consider a chance experiment in which an event whose probability of occurrence in a very small time interval ΔT is $P = \alpha \Delta T$, where α is a constant of proportionality. If successive occurrences are statistically independent, then the probability of k events in time T is

$$P_T(k) = \frac{(\alpha T)^k}{k!} e^{-\alpha T} \quad (6.182)$$

For example, the emission of electrons from a hot metal surface obeys this law, which is called the *Poisson distribution*.

The Poisson distribution can be used to approximate the binomial distribution when the number of trials n is large, the probability of each event p is small, and the product $np \cong npq$. The approximation is

$$P_n(k) \cong \frac{(\bar{K})^k}{k!} e^{-\bar{K}} \quad (6.183)$$

where, as calculated previously, $\bar{K} = E[K] = np$ and $\sigma_k^2 = E[K]q = npq \cong E[K]$ for $q = 1 - p \cong 1$. This approximation is compared with the binomial distribution in Figure 6.17(f).

EXAMPLE 6.24

The probability of error on a single transmission in a digital communication system is $P_E = 10^{-4}$. What is the probability of more than three errors in 1000 transmissions?

Solution

We find the probability of three errors or less from (6.183):

$$P(K \leq 3) = \sum_{k=0}^3 \frac{(\bar{K})^k}{k!} e^{-\bar{K}} \quad (6.184)$$

where $\bar{K} = (10^{-4})(1000) = 0.1$. Hence,

$$P(K \leq 3) = e^{-0.1} \left[\frac{(0.1)^0}{0!} + \frac{(0.1)^1}{1!} + \frac{(0.1)^2}{2!} + \frac{(0.1)^3}{3!} \right] \cong 0.999996 \quad (6.185)$$

Therefore, $P(K > 3) = 1 - P(K \leq 3) \cong 4 \times 10^{-6}$. ■

COMPUTER EXAMPLE 6.1

The MATLAB program given below does a Monte Carlo simulation of the digital communication system described in the above example.

```
% file: c6ce1
% Simulation of errors in a digital communication system
%
N.sim = input('Enter number of trials ');
N = input('Bit block size for simulation ');
N.errors = input('Simulate the probability of more than .. errors
occurring ');
PE = input('Error probability on each bit ');
count = 0;
for n = 1:N.sim
    U = rand(1, N);
    Error = (-sign(U-PE)+1)/2; % Error array - elements are 1 where
errors occur
    if sum(Error) > N.errors
        count = count + 1;
    end
end
P.greater = count/N.sim
% End of script file
```

A typical run follows. To cut down on the simulation time, blocks of 1000 bits are simulated with a probability of error on each bit of 10^{-3} . Note that the Poisson approximation does not hold in this case because $\bar{K} = (10^{-3})(1000) = 1$ is not much less than 1. Thus, to check the results analytically, we must use the binomial distribution. Calculation gives $P(0 \text{ errors}) = 0.3677$, $P(1 \text{ error}) = 0.3681$, $P(2 \text{ errors}) = 0.1840$, and $P(3 \text{ errors}) = 0.0613$ so that $P(> 3 \text{ errors}) = 1 - 0.3677 - 0.3681 - 0.1840 - 0.0613 = 0.0189$. This matches with the simulated result if both are rounded to two decimal places.

```
error.sim
Enter number of trials 10000
Bit block size for simulation 1000
Simulate the probability of more than .. errors occurring 3
Error probability on each bit .001
P.greater = 0.0199
```

6.4.4 Geometric Distribution

Suppose we are interested in the probability of the first head in a series of coin tossings, or the first error in a long string of digital signal transmissions occurring on the k th trial. The distribution describing such experiments is called the geometric distribution and is

$$P(k) = pq^{k-1}, \quad 1 \leq k < \infty \quad (6.186)$$

where p is the probability of the event of interest occurring (i.e., head, error, etc.) and q is the probability of it not occurring.

EXAMPLE 6.25

The probability of the first error occurring at the 1000th transmission in a digital data transmission system where the probability of error is $p = 10^{-6}$ is

$$P(1000) = 10^{-6}(1 - 10^{-6})^{999} = 9.99 \times 10^{-7} \cong 10^{-6}$$

6.4.5 Gaussian Distribution

In our future considerations, the Gaussian pdf will be used repeatedly. There are at least two reasons for this. One is that the assumption of Gaussian statistics for random phenomena often makes an intractable problem tractable. The other, and more fundamental reason, is that because of a remarkable phenomenon summarized by a theorem called the *central-limit theorem*, many naturally occurring random quantities, such as noise or measurement errors, are Gaussianly distributed. The following is a statement of the central-limit theorem.

THE CENTRAL-LIMIT THEOREM

Let X_1, X_2, \dots be independent, identically distributed random variables, each with finite mean m and finite variance σ^2 . Let Z_n be a sequence of unit-variance, zero-mean random variables, defined as

$$Z_n \triangleq \frac{\sum_{i=1}^n X_i - nm}{\sigma\sqrt{n}} \quad (6.187)$$

Then

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = \int_{-\infty}^z \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt \quad (6.188)$$

In other words, the cdf of the normalized sum (6.187) approaches a Gaussian cdf, no matter what the distribution of the component random variables. The only restriction is that they be independent and identically distributed and that their means and variances be finite. In some cases the independence and identically distributed assumptions can be relaxed. It is important, however, that no one of the component random variables or a finite combination of them dominate the sum.

We will not prove the central-limit theorem or use it in later work. We state it here simply to give partial justification for our almost exclusive assumption of Gaussian statistics from now on. For example, electrical noise is often the result of a superposition of voltages due to a large number of charge carriers. Turbulent boundary-layer pressure fluctuations on an aircraft skin are the superposition of minute pressures due to numerous eddies. Random errors in experimental measurements are due to many irregular fluctuating causes. In all these cases, the Gaussian approximation for the fluctuating quantity is useful and valid. Example 6.23

illustrates that surprisingly few terms in the sum are required to give a Gaussian-appearing pdf, even where the component pdfs are far from Gaussian.

The generalization of the joint Gaussian pdf first introduced in Example 6.15 is

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left\{ -\frac{\left(\frac{x-m_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-m_x}{\sigma_x}\right)\left(\frac{y-m_y}{\sigma_y}\right) + \left(\frac{y-m_y}{\sigma_y}\right)^2}{2(1-\rho^2)} \right\} \quad (6.189)$$

where, through straightforward but tedious integrations, it can be shown that

$$m_x = E[X] \quad \text{and} \quad m_y = E[Y] \quad (6.190)$$

$$\sigma_x^2 = \text{var}\{X\} \quad (6.191)$$

$$\sigma_y^2 = \text{var}\{Y\} \quad (6.192)$$

and

$$\rho = \frac{E[(X - m_x)(Y - m_y)]}{\sigma_x\sigma_y} \quad (6.193)$$

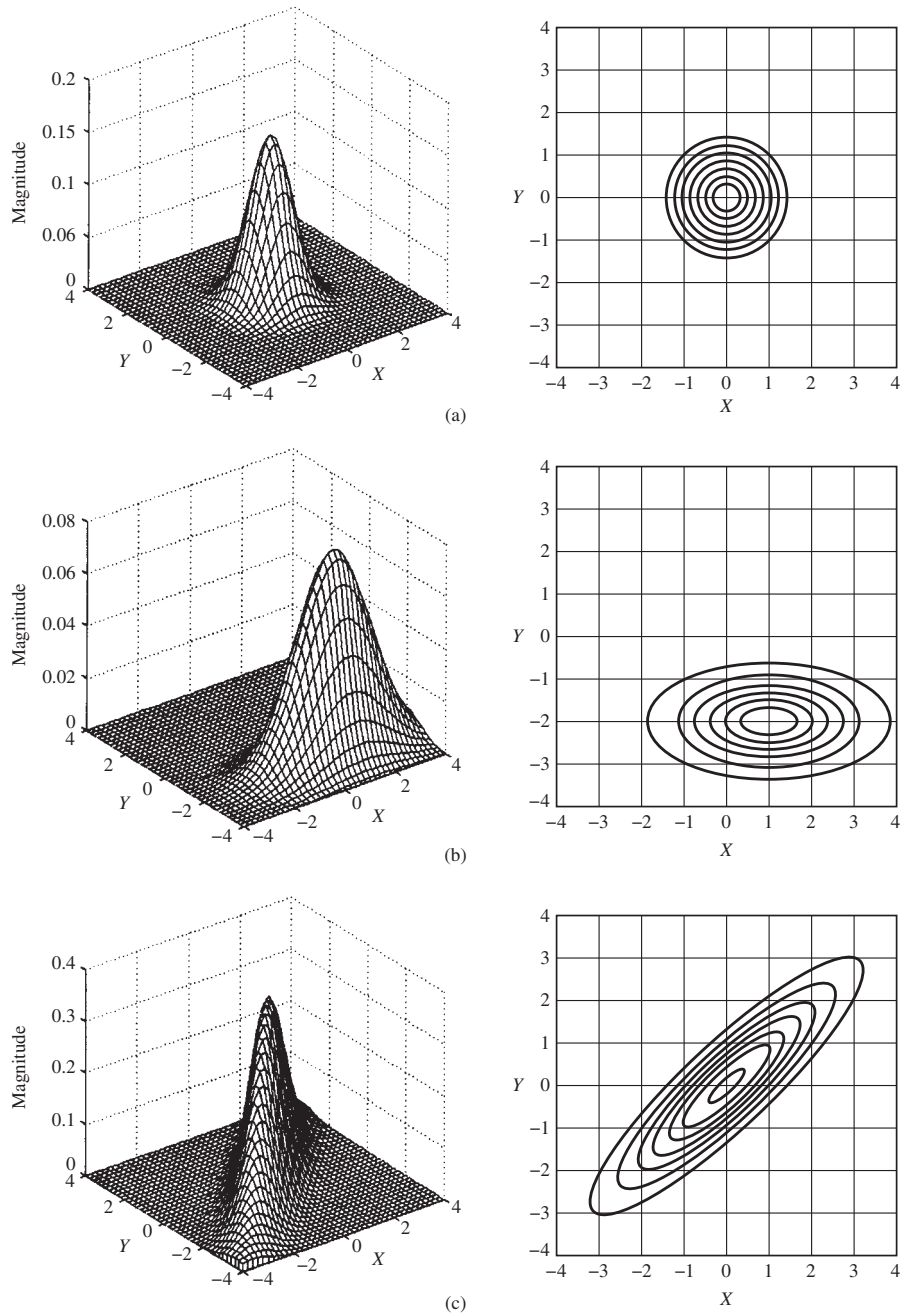
The joint pdf for $N > 2$ Gaussian random variables may be written in a compact fashion through the use of matrix notation. The general form is given in Appendix B.

Figure 6.18 illustrates the bivariate Gaussian density function, and the associated contour plots, as the five parameters m_x , m_y , σ_x^2 , σ_y^2 , and ρ are varied. The contour plots provide information on the shape and orientation of the pdf that is not always apparent in a three-dimensional illustration of the pdf from a single viewing point. Figure 6.18(a) illustrates the bivariate Gaussian pdf for which X and Y are zero mean, unit variance and uncorrelated. Since the variances of X and Y are equal, and since X and Y are uncorrelated, the contour plots are circles in the X - Y plane. We can see why two-dimensional Gaussian noise, in which the two components have equal variance and are uncorrelated, is said to exhibit circular symmetry. Figure 6.18(b) shows the case in which X and Y are uncorrelated but $m_x = 1$, $m_y = -2$, $\sigma_x^2 = 2$, and $\sigma_y^2 = 1$. The means are clear from observation of the contour plot. In addition, the spread of the pdf is greater in the X direction than in the Y direction because $\sigma_x^2 > \sigma_y^2$. In Figure 6.18(c) the means of X and Y are both zero but the correlation coefficient is set equal to 0.9. We see that the contour lines denoting a constant value of the density function are symmetrical about the line $X = Y$ in the X - Y plane. This results, of course, because the correlation coefficient is a measure of the linear relationship between X and Y . In addition, note that the pdfs described in Figures 5.18(a) and 5.18(b) can be factored into the product of two marginal pdfs since, for these two cases, X and Y are uncorrelated.

The marginal pdf for X (or Y) can be obtained by integrating (6.189) over y (or x). Again, the integration is tedious. The marginal pdf for X is

$$n(m_x, \sigma_x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp \left[-(x - m_x)^2 / 2\sigma_x^2 \right] \quad (6.194)$$

where the notation $n(m_x, \sigma_x)$ has been introduced to denote a Gaussian pdf of mean m_x and standard deviation σ_x . A similar expression holds for the pdf of Y with appropriate parameter changes. This function is shown in Figure 6.19.

**Figure 6.18**

Bivariate Gaussian pdfs and corresponding contour plots. (a) $m_x = 0$, $m_y = 0$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, and $\rho = 0$; (b) $m_x = 1$, $m_y = -2$, $\sigma_x^2 = 2$, $\sigma_y^2 = 1$, and $\rho = 0$; (c) $m_x = 0$, $m_y = 0$, $\sigma_x^2 = 1$, $\sigma_y^2 = 1$, and $\rho = 0.9$.

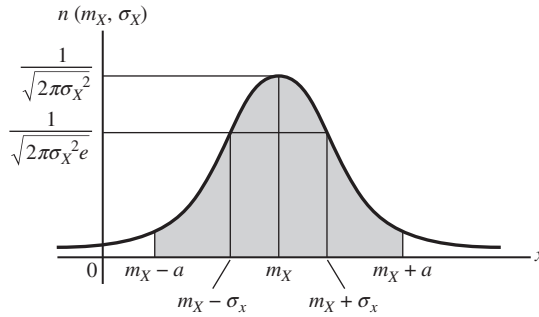


Figure 6.19
The Gaussian pdf with mean m_x and variance σ_x^2 .

We will sometimes assume in the discussions to follow that $m_x = m_y = 0$ in (6.189) and (6.194), for if they are not zero, we can consider new random variables X' and Y' defined as $X' = X - m_x$ and $Y' = Y - m_y$, which do have zero means. Thus, no generality is lost in assuming zero means.

For $\rho = 0$, that is, X and Y uncorrelated, the cross term in the exponent of (6.189) is zero, and $f_{XY}(x, y)$, with $m_x = m_y = 0$, can be written as

$$f_{XY}(x, y) = \frac{\exp(-x^2/2\sigma_x^2)}{\sqrt{2\pi\sigma_x^2}} \frac{\exp(-y^2/2\sigma_y^2)}{\sqrt{2\pi\sigma_y^2}} = f_X(x)f_Y(y) \quad (6.195)$$

Thus, *uncorrelated Gaussian random variables are also statistically independent*. We emphasize that this does not hold for all pdfs, however.

It can be shown that the sum of any number of Gaussian random variables, independent or not, is Gaussian. The sum of two independent Gaussian random variables is easily shown to be Gaussian. Let $Z = X_1 + X_2$, where the pdf of X_i is $n(m_i, \sigma_i)$. Using a table of Fourier transforms or completing the square and integrating, we find that the characteristic function of X_i is

$$\begin{aligned} M_{X_i}(jv) &= \int_{-\infty}^{\infty} (2\pi\sigma_i^2)^{-1/2} \exp\left[-\frac{(x_i - m_i)^2}{2\sigma_i^2}\right] \exp(jvx_i) dx_i \\ &= \exp\left(jm_i v - \frac{\sigma_i^2 v^2}{2}\right) \end{aligned} \quad (6.196)$$

Thus, the characteristic function of Z is

$$M_Z(jv) = M_{X_1}(jv)M_{X_2}(jv) = \exp\left[j(m_1 + m_2)v - \frac{(\sigma_1^2 + \sigma_2^2)v^2}{2}\right] \quad (6.197)$$

which is the characteristic function (6.196) of a Gaussian random variable of mean $m_1 + m_2$ and variance $\sigma_1^2 + \sigma_2^2$.

6.4.6 Gaussian Q-Function

As Figure 6.19 shows, $n(m_x, \sigma_x)$ describes a continuous random variable that may take on any value in $(-\infty, \infty)$ but is most likely to be found near $X = m_x$. The even symmetry of $n(m_x, \sigma_x)$ about $x = m_x$ leads to the conclusion that $P(X \leq m_x) = P(X \geq m_x) = \frac{1}{2}$.

Suppose we wish to find the probability that X lies in the interval $[m_x - a, m_x + a]$. Using (6.42), we can write this probability as

$$P[m_x - a \leq X \leq m_x + a] = \int_{m_x - a}^{m_x + a} \frac{\exp\left[-(x - m_x)^2 / 2\sigma_x^2\right]}{\sqrt{2\pi\sigma_x^2}} dx \quad (6.198)$$

which is the shaded area in Figure 6.19. With the change of variables $y = (x - m_x)/\sigma_x$, this gives

$$\begin{aligned} P[m_x - a \leq X \leq m_x + a] &= \int_{-a/\sigma_x}^{a/\sigma_x} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \\ &= 2 \int_0^{a/\sigma_x} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \end{aligned} \quad (6.199)$$

where the last integral follows by virtue of the integrand being even. Unfortunately, this integral cannot be evaluated in closed form.

The Gaussian Q -function, or simply Q -function, is defined as⁵

$$Q(u) = \int_u^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \quad (6.200)$$

This function has been evaluated numerically, and rational and asymptotic approximations are available to evaluate it for moderate and large arguments, respectively.⁶ Using this transcendental function definition, we may rewrite (6.199) as

$$\begin{aligned} P[m_x - a \leq X \leq m_x + a] &= 2 \left[\frac{1}{2} - \int_{a/\sigma_x}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \right] \\ &= 1 - 2Q\left(\frac{a}{\sigma_x}\right) \end{aligned} \quad (6.201)$$

A useful approximation for the Q -function for large arguments is

$$Q(u) \cong \frac{e^{-u^2/2}}{u\sqrt{2\pi}}, \quad u \gg 1 \quad (6.202)$$

⁵An integral representation with finite limits for the Q -function is $Q(x) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{x^2}{2\sin^2\phi}\right) d\phi$.

⁶These are provided in M. Abramowitz and I. Stegun (eds.), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Applied Mathematics Series No. 55, Issued June 1964 (pp. 931ff). Also New York: Dover, 1972.

Numerical comparison of (6.200) and (6.202) shows that less than a 6% error results for $u \geq 3$ in using this approximation. This, and other results for the Q -function, are given in Appendix F (see Part F.1).

Related integrals are the error function and the complementary error function, defined as

$$\begin{aligned} \operatorname{erf}(u) &= \frac{2}{\sqrt{\pi}} \int_0^u e^{-y^2} dy \\ \operatorname{erfc}(u) &= 1 - \operatorname{erf}(u) = \frac{2}{\sqrt{\pi}} \int_u^\infty e^{-y^2} dy \end{aligned} \quad (6.203)$$

respectively. They can be shown to be related to the Q -function by

$$Q(u) = \frac{1}{2} \operatorname{erfc}\left(\frac{u}{\sqrt{2}}\right) \quad \text{or} \quad \operatorname{erfc}(v) = 2Q(\sqrt{2}v) \quad (6.204)$$

MATLAB includes function programs for `erf` and `erfc`, and the inverse error and complementary error functions, `erfinv` and `erfcinv`, respectively.

6.4.7 Chebyshev's Inequality

The difficulties encountered above in evaluating (6.198) and probabilities like it make an approximation to such probabilities desirable. Chebyshev's inequality gives us a lower bound, regardless of the specific form of the pdf involved, provided its second moment is finite. The probability of finding a random variable X within $\pm k$ standard deviations of its mean is at least $1 - 1/k^2$, according to Chebyshev's inequality. That is,

$$P[|X - m_x| \leq k\sigma_x] \geq 1 - \frac{1}{k^2}, \quad k > 0 \quad (6.205)$$

Considering $k = 3$, we obtain

$$P[|X - m_x| \leq 3\sigma_x] \geq \frac{8}{9} \cong 0.889 \quad (6.206)$$

Assuming X is Gaussian and using the Q -function, this probability can be computed to be 0.9973. In words, according to Chebyshev's inequality, the probability that a random variable deviates from its mean by more than ± 3 standard deviations is not greater than 0.111, regardless of its pdf. (There is the restriction that its second moment must be finite.) Note that the bound for this example is not very tight.

6.4.8 Collection of Probability Functions and Their Means and Variances

The probability functions (pdfs and probability distributions) discussed above are collected in Table 6.4 along with some additional functions that come up from time to time. Also given are the means and variances of the corresponding random variables.

Table 6.4 Probability Distributions of Some Random Variables with Means and Variances

Probability-density or mass function	Mean	Variance
Uniform: $f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases}$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
Gaussian: $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-(x-m)^2/2\sigma^2]$	m	σ^2
Rayleigh: $f_R(r) = \frac{r}{\sigma^2} \exp(-r^2/2\sigma^2), \quad r \geq 0$	$\sqrt{\frac{\pi}{2}}\sigma$	$\frac{1}{2}(4-\pi)\sigma^2$
Laplacian: $f_X(x) = \frac{\alpha}{2} \exp(-\alpha x), \quad \alpha > 0$	0	$2/\alpha^2$
One-sided exponential: $f_X(x) = \alpha \exp(-\alpha x) u(x)$	$1/\alpha$	$1/\alpha^2$
Hyperbolic: $f_X(x) = \frac{(m-1)h^{m-1}}{2(x +h)^m}, \quad m > 3, h > 0$	0	$\frac{2h^2}{(m-3)(m-2)}$
Nakagami- m : $f_X(x) = \frac{2m^m}{\Gamma(m)} x^{2m-1} \exp(-mx^2), \quad x \geq 0$	$\frac{1 \times 3 \times \dots \times (2m-1)}{2^m \Gamma(m)}$	$\frac{\Gamma(m+1)}{\Gamma(m)\sqrt{m}}$
Central Chi-square ($n = \text{degrees of freedom}$) ¹ : $f_X(x) = \frac{x^{n/2-1}}{\sigma^n 2^{n/2} \Gamma(n/2)} \exp(-x/2\sigma^2)$	$n\sigma^2$	$2n\sigma^4$
Lognormal ² : $f_X(x) = \frac{1}{x\sqrt{2\pi\sigma_y^2}} \exp\left[-(\ln x - m_y)^2/2\sigma_y^2\right]$	$\exp\left(m_y + 2\sigma_y^2\right)$	$\exp\left(2m_y + \sigma_y^2\right) \times \left[\exp\left(\sigma_y^2\right) - 1\right]$
Binomial: $P_n(k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \dots, n, \quad p+q=1$	np	npq
Poisson: $P(k) = \frac{\lambda^k}{k!} \exp(-\lambda), \quad k = 0, 1, 2, \dots$	λ	λ
Geometric: $P(k) = pq^{k-1}, \quad k = 1, 2, \dots$	$1/p$	q/p^2

¹ $\Gamma(m)$ is the gamma function and equals $(m-1)!$ for m an integer.²The lognormal random variable results from the transformation $Y = \ln X$ where Y is a Gaussian random variable with mean m_y and variance σ_y^2 .

Further Reading

Several books are available that deal with probability theory for engineers. Among these are Leon-Garcia (1994), Ross (2002), and Walpole, Meyers, Meyers, and Ye (2007). A good overview with many examples is Ash (1992). Simon (2002) provides a compendium of relations involving the Gaussian distribution.

Summary

1. The objective of probability theory is to attach real numbers between 0 and 1, called *probabilities*, to the *outcomes* of chance experiments—that is, experiments in which the outcomes are not uniquely determined by the causes but depend on chance—and to interrelate probabilities of events, which are defined to be combinations of outcomes.

2. Two events are *mutually exclusive* if the occurrence of one of them precludes the occurrence of the other. A set of events is said to be *exhaustive* if one of them must occur in the performance of a chance experiment. The null event happens with probability zero, and the certain event happens with probability one in the performance of a chance experiment.

3. The equally likely definition of the probability $P(A)$ of an event A states that if a chance experiment can result in a number N of mutually exclusive, equally likely outcomes, then $P(A)$ is the ratio of the number of outcomes favorable to A , or N_A , to the total number. It is a circular definition in that probability is used to define probability, but it is nevertheless useful in many situations such as drawing cards from well-shuffled decks.

4. The relative-frequency definition of the probability of an event A assumes that the chance experiment is replicated a large number of times N and

$$P(A) = \lim_{N \rightarrow \infty} \frac{N_A}{N}$$

where N_A is the number of replications resulting in the occurrence of A .

5. The axiomatic approach defines the probability $P(A)$ of an event A as a real number satisfying the following axioms:

- (a) $P(A) \geq 0$.
- (b) $P(\text{certain event}) = 1$.
- (c) If A and B are mutually exclusive events, $P(A \cup B) = P(A) + P(B)$.

The axiomatic approach encompasses the equally likely and relative-frequency definitions.

6. Given two events A and B , the compound event “ A or B or both,” is denoted as $A \cup B$, the compound event “both A and B ” is denoted as $\overline{(A \cap B)}$ or as (AB) , and the event “not A ” is denoted as \overline{A} . If A and B are not necessarily mutually exclusive, the axioms of probability may be used to show that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Letting $P(A|B)$ denote the probability of A occurring given that B occurred and $P(B|A)$ denote the probability of B given A , these probabilities are defined, respectively, as

$$P(A|B) = \frac{P(AB)}{P(B)} \quad \text{and} \quad P(B|A) = \frac{P(AB)}{P(A)}$$

A special case of Bayes’ rule results by putting these two definitions together:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Statistically independent events are events for which $P(AB) = P(A)P(B)$.

7. A random variable is a rule that assigns real numbers to the outcomes of a chance experiment. For example, in flipping a coin, assigning $X = +1$ to the occurrence of a head and $X = -1$ to the occurrence of a tail constitutes the assignment of a discrete-valued random variable.

8. The cumulative-distribution function (cdf) $F_X(x)$ of a random variable X is defined as the probability that $X \leq x$ where x is a running variable. $F_X(x)$ lies between 0 and 1 with $F_X(-\infty) = 0$ and $F_X(\infty) = 1$, is continuous from the right, and is a nondecreasing function of its argument. Discrete random variables have step-discontinuous cdfs, and continuous random variables have continuous cdfs.

9. The probability-density function (pdf) $f_X(x)$ of a random variable X is defined to be the derivative of the cdf. Thus,

$$F_X(x) = \int_{-\infty}^x f_X(\eta) d\eta$$

The pdf is nonnegative and integrates over all x to unity. A useful interpretation of the pdf is that $f_X(x) dx$

is the probability of the random variable X lying in an infinitesimal range dx about x .

10. The joint cdf $F_{XY}(x, y)$ of two random variables X and Y is defined as the probability that $X \leq x$ and $Y \leq y$ where x and y are particular values of X and Y . Their joint pdf $f_{XY}(x, y)$ is the second partial derivative of the cdf first with respect to x and then with respect to y . The cdf of X alone (that is, the marginal cdf) is found by setting $y(x)$ to infinity in the argument of F_{XY} . The pdf of X (Y) alone (that is, the marginal pdf) is found by integrating f_{XY} over all y (x).

11. Two statistically independent random variables have joint cdfs and pdfs that factor into the respective marginal cdfs or pdfs.

12. The conditional pdf of X given Y is defined as

$$f_{X|Y}(x|y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

with a similar definition for $f_{Y|X}(y|x)$. The expression $f_{X|Y}(x|y) dx$ can be interpreted as the probability that $x - dx < X \leq x$ given $Y = y$.

13. Given $Y = g(X)$ where $g(X)$ is a monotonic function,

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|_{x=g^{-1}(y)}$$

where $g^{-1}(y)$ is the inverse of $y = g(x)$. Joint pdfs of functions of more than one random variable can be transformed similarly.

14. Important probability functions defined in Chapter 5 are the Rayleigh pdf [Equation (6.105)], the pdf of a random-phased sinusoid (Example 6.17), the uniform pdf [Example 6.20, Equation (6.135)], the binomial probability distribution [Equation (6.174)], the Laplace and Poisson approximations to the binomial distribution [Equations (6.181) and (6.183)], and the Gaussian pdf [Equations (6.189) and (6.192)].

15. The statistical average, or expectation, of a function $g(X)$ of a random variable X with pdf $f_X(x)$ is defined as

$$E[g(X)] = \overline{g(X)} = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

The average of $g(X) = X^n$ is called the n th moment of X . The first moment is known as the *mean* of X . Averages of functions of more than one random variable are found by integrating the function times the joint pdf over the ranges of its arguments. The averages $g(X, Y) = \overline{X^n Y^m} \triangleq E\{X^n Y^m\}$ are called the *joint moments* of the

order $m + n$. The variance of a random variable X is the average $\overline{(X - \bar{X})^2} = \overline{X^2} - \bar{X}^2$.

16. The average $E[\sum a_i X_i]$ is $\sum a_i E[X_i]$; that is, the operations of summing and averaging can be interchanged. The variance of a sum of random variables is the sum of the respective variances *if the random variables are statistically independent*.

17. The characteristic function $m_X(jv)$ of a random variable X that has the pdf $f_X(x)$ is the expectation of $\exp(jvX)$ or, equivalently, the Fourier transform of $f_X(x)$ with a plus sign in the exponential of the Fourier-transform integral. Thus, the pdf is the inverse Fourier transform (with the sign in the exponent changed from minus to plus) of the characteristic function.

18. The n th moment of X can be found from $M_X(jv)$ by differentiating with respect to v for n times, multiplying by $(-j)^n$, and setting $v = 0$. The characteristic function of $Z = X + Y$, where X and Y are independent, is the product of the respective characteristic functions of X and Y . Thus, by the convolution theorem of Fourier transforms, the pdf of Z is the convolution of the pdfs of X and Y .

19. The covariance μ_{XY} of two random variables X and Y is the average

$$\mu_{XY} = E[(X - \bar{X})(Y - \bar{Y})] = E[XY] - E[X]E[Y]$$

The correlation coefficient ρ_{XY} is

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y}$$

Both give a measure of the linear interdependence of X and Y , but ρ_{XY} is handier because it is bounded by ± 1 . If $\rho_{XY} = 0$, the random variables are said to be uncorrelated.

20. The central-limit theorem states that, under suitable restrictions, the sum of a large number N of independent random variables with finite variances (not necessarily with the same pdfs) tends to a Gaussian pdf as N becomes large.

21. The Q -function can be used to compute probabilities of Gaussian random variables being in certain ranges. The Q -function is tabulated in Appendix F.1, and rational and asymptotic approximations are given for computing it. It can be related to the error function through (6.204).

22. Chebyshev's inequality gives the lower bound of the probability that a random variable is within k standard deviations of its mean as $1 - \frac{1}{k^2}$, regardless of the pdf of the random variable (its second moment must be finite).

23. Table 6.4 summarizes a number of useful probability distributions with their means and variances.

Drill Problems

6.1 A fair coin and a fair die (six sides) are tossed simultaneously with neither affecting the outcome of the other. Give probabilities for the following events using the principle of equal likelihood:

- (a) A head and a six;
- (b) A tail and a one or a two;
- (c) A tail or a head and a four;
- (d) A head and a number less than a five;
- (e) A tail or a head and a number greater than a four;
- (f) A tail and a number greater than a six.

6.2 In tossing a six-sided fair die, we define event $A = \{2 \text{ or } 4 \text{ or } 6\}$ and event $B = \{1 \text{ or } 3 \text{ or } 5 \text{ or } 6\}$. Using equal likelihood and the axioms of probability, find the following:

- (a) $P(A)$;
- (b) $P(B)$;
- (c) $P(A \cup B)$;
- (d) $P(A \cap B)$;
- (e) $P(A|B)$;
- (f) $P(B|A)$.

6.3 In tossing a single six-sided fair die, event $A = \{1 \text{ or } 3\}$, event $B = \{2 \text{ or } 3 \text{ or } 4\}$, and event $C = \{4 \text{ or } 5 \text{ or } 6\}$. Find the following probabilities:

- (a) $P(A)$;
- (b) $P(B)$;
- (c) $P(C)$;
- (d) $P(A \cup B)$;
- (e) $P(A \cup C)$;
- (f) $P(B \cup C)$;
- (g) $P(A \cap B)$;
- (h) $P(A \cap C)$;
- (i) $P(B \cap C)$;
- (j) $P(A \cap (B \cap C))$;
- (k) $P(A \cup (B \cup C))$.

6.4 Referring to Drill Problem 6.2, find the following:

- (a) $P(A|B)$;
- (b) $P(B|A)$.

6.5 Referring to Drill Problem 6.3, find the following:

- (a) $P(A|B)$;
- (b) $P(B|A)$;
- (c) $P(A|C)$;
- (d) $P(C|A)$;
- (e) $P(B|C)$;
- (f) $P(C|B)$.

6.6

- (a) What is the probability drawing an ace from a 52-card deck with a single draw?
- (b) What is the probability drawing the ace of spades from a 52-card deck with a single draw?
- (c) What is the probability of drawing the ace of spades from a 52-card deck with a single draw given that the card drawn was black?

6.7 Given a pdf of the form $f_X(x) = A \exp(-\alpha x)u(x-1)$, where $u(x)$ is the unit step and A and α are positive constants, do the following:

- (a) Give the relationship between A and α .
- (b) Find the cdf.
- (c) Find the probability that $2 < X \leq 4$.
- (d) Find the mean of X .
- (e) Find the mean square of X .
- (f) Find the variance of X .

6.8 Given a joint pdf defined as

$$f_{XY}(x, y) = \begin{cases} 1, & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Find the following:

- (a) $f_X(x)$;
- (b) $f_Y(y)$;
- (c) $E[X]$, $E[Y]$, $E[X^2]$, $E[Y^2]$, σ_X^2 , σ_Y^2 ;
- (d) $E[XY]$;
- (e) μ_{XY} .

6.9

- (a) What is the probability of getting two or fewer heads in tossing a fair coin 10 times?
- (b) What is the probability of getting exactly five heads in tossing a fair coin 10 times?

6.10 A random variable Z is defined as $Z = X + Y$ where X and Y are Gaussian with the following statistics:

1. $E[X] = 2, E[Y] = -3$
2. $\sigma_X = 2, \sigma_Y = 3$
3. $\mu_{XY} = 0.5$

Find the pdf of Z .

6.11 Let a random variable Z be defined in terms of three independent random variables as $Z = 2X_1 + 4X_2 + 3X_3$, where the means of X_1, X_2 , and X_3 are $-1, 5$, and -2 , respectively, and their respective variances are $4, 7$, and 1 . Find the following:

- (a) The mean of Z ;
- (b) The variance of Z ;
- (c) The standard deviation of Z ;
- (d) The pdf of Z if X_1, X_2 , and X_3 are Gaussian.

6.12 The characteristic function of a random variable, X , is $M_X(jv) = (1 + v^2)^{-1}$. Find the following:

- (a) The mean of X ;
- (b) The variance of X ;
- (c) The pdf of X .

6.13 A random variable is defined as the sum of ten independent random variables, which are all uniformly distributed in $[-0.5, 0.5]$.

- (a) According to the central-limit theorem, write down an approximate expression for the pdf of the sum, $Z = \sum_{i=1}^{10} X_i$
- (b) What is the value of the approximating pdf for $z = \pm 5.1$? What is the value of the pdf for the actual sum random variable for this value of z ?

6.14 A fair coin is tossed 100 times. According to the Laplace approximation, what is the probability that exactly (a) 50 heads are obtained? (b) 51 heads? (c) 52 heads? (d) Is the Laplace approximation valid in these computations?

6.15 The probability of error on a single transmission in a digital communication system is $P_E = 10^{-3}$. (a) What is the probability of 0 errors in 100 transmissions? (b) 1 error in 100? (c) 2 errors in 100? (d) 2 or fewer errors in 100?

Problems

Section 6.1

6.1 A circle is divided into 21 equal parts. A pointer is spun until it stops on one of the parts, which are numbered from 1 through 21. Describe the sample space and, assuming equally likely outcomes, find

- (a) $P(\text{an even number})$;
- (b) $P(\text{the number } 21)$;
- (c) $P(\text{the numbers } 4, 5, \text{ or } 9)$;
- (d) $P(\text{a number greater than } 10)$.

6.2 If five cards are drawn without replacement from an ordinary deck of cards, what is the probability that

- (a) three kings and two aces result;
- (b) four of a kind result;
- (c) all are of the same suit;
- (d) an ace, king, queen, jack, and ten of the same suit result;
- (e) given that an ace, king, jack, and ten have been drawn, what is the probability that the next card drawn will be a queen (not all of the same suit)?

6.3 What equations must be satisfied in order for three events A, B , and C to be independent? (*Hint*: They must be independent by pairs, but this is not sufficient.)

6.4 Two events, A and B , have respective marginal probabilities $P(A) = 0.2$ and $P(B) = 0.5$, respectively. Their joint probability is $P(A \cap B) = 0.4$.

- (a) Are they statistically independent? Why or why not?
- (b) What is the probability of A or B or both occurring?
- (c) In general, what must be true for two events to be both statistically independent and mutually exclusive?

6.5 Figure 6.20 is a graph that represents a communication network, where the nodes are receiver/repeater boxes and the edges (or links) represent communication channels, which, if connected, convey the message perfectly. However, there is the probability p that a link will be broken and the probability $q = 1 - p$ that it will be whole. *Hint*: Use a tree diagram like Figure 6.2.

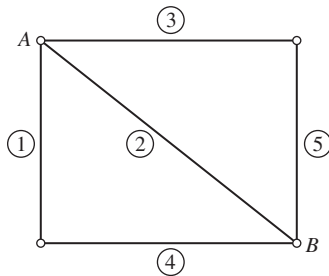


Figure 6.20

Table 6.5 Table of Probabilities for Problem 6.7.

	B_1	B_2	B_3	$P(A_i)$
A_1	0.05		0.45	0.55
A_2		0.15	0.10	
A_3	0.05	0.05		0.15
$P(B_j)$				1.0

- (a) What is the probability that at least one working path is available between the nodes labeled A and B ?
- (b) Remove link 4. Now what is the probability that at least one working path is available between nodes A and B ?
- (c) Remove link 2. What is the probability that at least one working path is available between nodes A and B ?
- (d) Which is the more serious situation, the removal of link 4 or link 2? Why?

6.6 Given a binary communication channel where A = input and B = output, let $P(A) = 0.45$, $P(B|A) = 0.95$, and $P(\overline{B}|A) = 0.65$. Find $P(A|B)$ and $P(A|\overline{B})$.

6.7 Given the table of joint probabilities of Table 6.5,

- (a) Find the probabilities omitted from Table 6.5,
- (b) Find the probabilities $P(A_3|B_3)$, $P(B_2|A_1)$, and $P(B_3|A_2)$.

Section 6.2

6.8 Two dice are tossed.

- (a) Let X_1 be a random variable that is numerically equal to the total number of spots on the up faces of the dice. Construct a table that defines this random variable.

- (b) Let X_2 be a random variable that has the value of 1 if the sum of the number of spots up on both dice is even and the value zero if it is odd. Repeat part (a) for this case.

6.9 Three fair coins are tossed simultaneously such that they don't interact. Define a random variable $X = 1$ if an even number of heads is up and $X = 0$ otherwise. Plot the cumulative-distribution function and the probability-density function corresponding to this random variable.

6.10 A certain *continuous* random variable has the cumulative-distribution function

$$F_X(x) = \begin{cases} 0, & x < 0 \\ Ax^4, & 0 \leq x \leq 12 \\ B, & x > 12 \end{cases}$$

- (a) Find the proper values for A and B .
- (b) Obtain and plot the pdf $f_X(x)$.
- (c) Compute $P(X > 5)$.
- (d) Compute $P(4 \leq X < 6)$.

6.11 The following functions can be pdfs if the constants are chosen properly. Find the proper conditions on the constants so that they are. [$A, B, C, D, \alpha, \beta, \gamma$, and τ are positive constants and $u(x)$ is the unit step function.]

- (a) $f(x) = Ae^{-\alpha x}u(x)$, where $u(x)$ is the unit step
- (b) $f(x) = Be^{\beta x}u(-x)$

- (c) $f(x) = Ce^{-\gamma x}u(x-1)$
 (d) $f(x) = C[u(x) - u(x-\tau)]$

6.12 Test X and Y for independence if

- (a) $f_{XY}(x, y) = Ae^{-|x|-2|y|}$
 (b) $f_{XY}(x, y) = C(1-x-y), 0 \leq x \leq 1-y$ and $0 \leq y \leq 1$

Prove your answers.

6.13 The joint pdf of two random variables is

$$f_{XY}(x, y) = \begin{cases} C(1+xy), & 0 \leq x \leq 4, 0 \leq y \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Find the following:

- (a) The constant C
 (b) $f_{XY}(1, 1.5)$
 (c) $f_{XY}(x, 3)$
 (d) $f_{X|Y}(x | 3)$

6.14 The joint pdf of the random variables X and Y is

$$f_{XY}(x, y) = Axye^{-(x+y)}, x \geq 0 \text{ and } y \geq 0$$

- (a) Find the constant A .
 (b) Find the marginal pdfs of X and Y , $f_X(x)$ and $f_Y(y)$.
 (c) Are X and Y statistically independent? Justify your answer.

6.15

- (a) For what value of $\alpha > 0$ is the function

$$f(x) = \alpha x^{-2}u(x-\alpha)$$

a probability-density function? Use a sketch to illustrate your reasoning and recall that a pdf has to integrate to one. [$u(x)$ is the unit step function.]

- (b) Find the corresponding cumulative-distribution function.
 (c) Compute $P(X \geq 10)$.

6.16 Given the Gaussian random variable with the pdf

$$f_X(x) = \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$$

where $\sigma > 0$ is the standard deviation. If $Y = X^2$, find the pdf of Y . *Hint:* Note that $Y = X^2$ is symmetrical about $X = 0$ and that it is impossible for Y to be less than zero.

6.17 A nonlinear system has input X and output Y . The pdf for the input is Gaussian as given in Problem 6.16.

Determine the pdf of the output, assuming that the nonlinear system has the following input/output relationship:

$$(a) Y = \begin{cases} aX, & X \geq 0 \\ 0, & X < 0 \end{cases}$$

Hint: When $X < 0$, what is Y ? How is this manifested in the pdf for Y ?

- (b) $Y = |X|$;
 (c) $Y = X - X^3/3$.

Section 6.3

6.18 Let $f_X(x) = A \exp(-bx)u(x-2)$ for all x where A and b are positive constants.

- (a) Find the relationship between A and b such that this function is a pdf.
 (b) Calculate $E(X)$ for this random variable.
 (c) Calculate $E(X^2)$ for this random variable.
 (d) What is the variance of this random variable?

6.19

- (a) Consider a random variable uniformly distributed between 0 and 2. Show that $E(X^2) > E^2(X)$.
 (b) Consider a random variable uniformly distributed between 0 and 4. Show that $E(X^2) > E^2(X)$.
 (c) Can you show in general that for any random variable it is true that $E(X^2) > E^2(X)$ unless the random variable is zero almost always? (*Hint:* Expand $E\{[X - E(X)]^2 \geq 0\}$ and note that it is 0 only if $X = 0$ with probability 1.)

6.20 Verify the entries in Table 6.5 for the mean and variance of the following probability distributions:

- (a) Rayleigh;
 (b) One-sided exponential;
 (c) Hyperbolic;
 (d) Poisson;
 (e) Geometric.

6.21 A random variable X has the pdf

$$f_X(x) = Ae^{-bx}[u(x) - u(x-B)]$$

where $u(x)$ is the unit step function and A , B , and b are positive constants.

304 Chapter 6 • Overview of Probability and Random Variables

- (a) Find the proper relationship between the constants A , b , and B . Express b in terms of A and B .
- (b) Determine and plot the cdf.
- (c) Compute $E(X)$.
- (d) Determine $E(X^2)$.
- (e) What is the variance of X ?

6.22 If

$$f_X(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

show that

- (a) $E[X^{2n}] = 1 \cdot 3 \cdot 5 \cdots (2n-1)\sigma^{2n}$, for $n = 1, 2, \dots$
- (b) $E[X^{2n-1}] = 0$ for $n = 1, 2, \dots$

6.23 The random variable has pdf

$$f_X(x) = \frac{1}{2}\delta(x-5) + \frac{1}{8}[u(x-4) - u(x-8)]$$

where $u(x)$ is the unit step. Determine the mean and the variance of the random variable thus defined.

6.24 Two random variables X and Y have means and variances given below:

$$m_x = 1 \quad \sigma_x^2 = 4 \quad m_y = 3 \quad \sigma_y^2 = 7$$

A new random variable Z is defined as

$$Z = 3X - 4Y$$

Determine the mean and variance of Z for each of the following cases of correlation between the random variables X and Y :

- (a) $\rho_{XY} = 0$
- (b) $\rho_{XY} = 0.2$
- (c) $\rho_{XY} = 0.7$
- (d) $\rho_{XY} = 1.0$

6.25 Two Gaussian random variables X and Y , with zero means and variances σ^2 , between which there is a correlation coefficient ρ , have a joint probability-density function given by

$$f(x, y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2\sigma^2(1-\rho^2)}\right]$$

The marginal pdf of Y can be shown to be

$$f_Y(y) = \frac{\exp(-y^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}}$$

Find the conditional pdf $f_{X|Y}(x|y)$.

6.26 Using the definition of a conditional pdf given by Equation (6.62) and the expressions for the marginal and joint Gaussian pdfs, show that for two jointly Gaussian random variables X and Y , the conditional density function of X given Y has the form of a Gaussian density with conditional mean and the conditional variance given by

$$E(X|Y) = m_x + \frac{\rho\sigma_x}{\sigma_y}(Y - m_y)$$

and

$$\text{var}(X|Y) = \sigma_x^2(1 - \rho^2)$$

respectively.

6.27 The random variable X has a probability-density function uniform in the range $0 \leq x \leq 2$ and zero elsewhere. The independent variable Y has a density uniform in the range $1 \leq y \leq 5$ and zero elsewhere. Find and plot the density of $Z = X + Y$.

6.28 A random variable X is defined by

$$f_X(x) = 4e^{-8|x|}$$

The random variable Y is related to X by $Y = 4 + 5X$.

- (a) Determine $E[X]$, $E[X^2]$, and σ_x^2 .
- (b) Determine $f_Y(y)$.
- (c) Determine $E[Y]$, $E[Y^2]$, and σ_y^2 . (*Hint:* The result of part (b) is not necessary to do this part, although it may be used.)
- (d) If you used $f_Y(y)$ in part (c), repeat that part using only $f_X(x)$.

6.29 A random variable X has the probability-density function

$$f_X(x) = \begin{cases} ae^{-ax}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

where a is an arbitrary positive constant.

- (a) Determine the characteristic function $M_X(jv)$.
- (b) Use the characteristic function to determine $E[X]$ and $E[X^2]$.
- (c) Check your results by computing

$$\int_{-\infty}^{\infty} x^n f_X(x) dx$$

for $n = 1$ and 2 .

- (d) Compute σ_x^2 .

Section 6.4

6.30 Compare the binomial, Laplace, and Poisson distributions for

- (a) $n = 3$ and $p = \frac{1}{5}$
- (b) $n = 3$ and $p = \frac{1}{10}$
- (c) $n = 10$ and $p = \frac{1}{5}$
- (d) $n = 10$ and $p = \frac{1}{10}$

6.31 An honest coin is flipped 10 times.

- (a) Determine the probability of the occurrence of either 5 or 6 heads.
- (b) Determine the probability of the first head occurring at toss number 5.
- (c) Repeat parts (a) and (b) for flipping 100 times and the probability of the occurrence of 50 to 60 heads inclusive and the probability of the first head occurring at toss number 50.

6.32 Passwords in a computer installation take the form $X_1X_2X_3X_4$, where each character X_i is one of the 26 letters of the alphabet. Determine the maximum possible number of different passwords available for assignment for each of the two following conditions:

- (a) A given letter of the alphabet can be used only once in a password.
- (b) Letters can be repeated if desired, so that each X_i is completely arbitrary.
- (c) If selection of letters for a given password is completely random, what is the probability that your competitor could access, on a single try, your computer in part (a)? part (b)?

6.33 Assume that 20 honest coins are tossed.

- (a) By applying the binomial distribution, find the probability that there will be fewer than 3 heads.
- (b) Do the same computation using the Laplace approximation.
- (c) Compare the results of parts (a) and (b) by computing the percent error of the Laplace approximation.

6.34 A digital data transmission system has an error probability of 10^{-5} per digit.

- (a) Find the probability of exactly 1 error in 10^5 digits.

- (b) Find the probability of exactly 2 errors in 10^5 digits.

- (c) Find the probability of more than 5 errors in 10^5 digits.

6.35 Assume that two random variables X and Y are jointly Gaussian with $m_x = m_y = 1$, $\sigma_x^2 = \sigma_y^2 = 4$.

- (a) Making use of (6.194), write down an expression for the marginal pdfs of X and of Y .
- (b) Write down an expression for the conditional pdf $f_{X|Y}(x|y)$ by using the result of (a) and an expression for $f_{XY}(x, y)$ written down from (6.189). Deduce that $f_{Y|X}(y|x)$ has the same form with y replacing x .
- (c) Put $f_{X|Y}(x|y)$ into the form of a marginal Gaussian pdf. What is its mean and variance? (The mean will be a function of y .)

6.36 Consider the Cauchy density function

$$f_X(x) = \frac{K}{1+x^2}, \quad -\infty \leq x \leq \infty$$

- (a) Find K .
- (b) Show that $\text{var}\{X\}$ is not finite.
- (c) Show that the characteristic function of a Cauchy random variable is $M_X(jv) = \pi K e^{-|v|}$.
- (d) Now consider $Z = X_1 + \dots + X_N$ where the X_i 's are independent Cauchy random variables. Thus, their characteristic function is

$$M_Z(jv) = (\pi K)^N \exp(-N|v|)$$

Show that $f_Z(z)$ is Cauchy. (Comment: $f_Z(z)$ is not Gaussian as $N \rightarrow \infty$ because $\text{var}\{X_i\}$ is not finite and the conditions of the central-limit theorem are violated.)

6.37 (Chi-squared pdf) Consider the random variable $Y = \sum_{i=1}^N X_i^2$ where the X_i 's are independent Gaussian random variables with pdfs $n(0, \sigma)$.

- (a) Show that the characteristic function of X_i^2 is

$$M_{X_i^2}(jv) = (1 - 2jv\sigma^2)^{-1/2}$$

- (b) Show that the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{y^{N/2-1} e^{-y/2\sigma^2}}{2^{N/2} \sigma^N \Gamma(N/2)}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

where $\Gamma(x)$ is the gamma function, which, for $x = n$, an integer is $\Gamma(n) = (n-1)!$. This pdf is

known as the χ^2 (chi-squared) pdf with N degrees of freedom. *Hint:* Use the Fourier-transform pair

$$\frac{y^{N/2-1} e^{-y/\alpha}}{\alpha^{N/2} \Gamma(N/2)} \leftrightarrow (1 - j\alpha v)^{-N/2}$$

- (c) Show that for N large, the χ^2 pdf can be approximated as

$$f_Y(y) = \frac{\exp\left[-\frac{1}{2}\left(\frac{y-N\sigma^2}{\sqrt{4N\sigma^4}}\right)^2\right]}{\sqrt{4N\pi\sigma^4}}, \quad N \gg 1$$

Hint: Use the central-limit theorem. Since the x_i 's are independent,

$$\bar{Y} = \sum_{i=1}^N \overline{X_i^2} = N\sigma^2$$

and

$$\text{var}(Y) = \sum_{i=1}^N \text{var}(X_i^2) = N \text{var}(X_i^2)$$

- (d) Compare the approximation obtained in part (c) with $f_Y(y)$ for $N = 2, 4, 8$.
 (e) Let $R^2 = Y$. Show that the pdf of R for $N = 2$ is Rayleigh.

6.38 Compare the Q -function and the approximation to it for large arguments given by (6.202) by plotting both expressions on a log-log graph. (*Note:* MATLAB is handy for this problem.)

6.39 Determine the cdf for a Gaussian random variable of mean m and variance σ^2 . Express in terms of the Q -function. Plot the resulting cdf for $m = 0$, and $\sigma = 0.5, 1$, and 2 .

6.40 Prove that the Q -function may also be represented as $Q(x) = \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-\frac{x^2}{2\sin^2\phi}\right) d\phi$.

6.41 A random variable X has the probability-density function

$$f_X(x) = \frac{e^{-(x-10)^2/50}}{\sqrt{50\pi}}$$

Express the following probabilities in terms of the Q -function and calculate numerical answers for each:

- (a) $P(|X| \leq 15)$;
 (b) $P(10 < X \leq 20)$;
 (c) $P(5 < X \leq 25)$;
 (d) $P(20 < X \leq 30)$.

6.42

- (a) Prove Chebyshev's inequality. *Hint:* Let $Y = (X - m_x)/\sigma_x$ and find a bound for $P(|Y| < k)$ in terms of k .
 (b) Let X be uniformly distributed over $|x| \leq 1$. Plot $P(|X| \leq k\sigma_x)$ versus k and the corresponding bound given by Chebyshev's inequality.

6.43 If the random variable X is Gaussian, with zero mean and variance σ^2 , obtain numerical values for the following probabilities:

- (a) $P(|X| > \sigma)$;
 (b) $P(|X| > 2\sigma)$;
 (c) $P(|X| > 3\sigma)$.

6.44 Speech is sometimes idealized as having a Laplacian-amplitude pdf. That is, the amplitude is distributed according to

$$f_X(x) = \left(\frac{a}{2}\right) \exp(-a|x|)$$

- (a) Express the variance of X , σ^2 , in terms of a . Show your derivation; don't just simply copy the result given in Table 6.4.
 (b) Compute the following probabilities: $P(|X| > \sigma)$; $P(|X| > 2\sigma)$; $P(|X| > 3\sigma)$.

6.45 Two jointly Gaussian zero-mean random variables, X and Y , have respective variances of 3 and 4 and correlation coefficient $\rho_{XY} = -0.4$. A new random variable is defined as $Z = X + 2Y$. Write down an expression for the pdf of Z .

6.46 Two jointly Gaussian random variables, X and Y , have means of 1 and 2, and variances of 3 and 2, respectively. Their correlation coefficient is $\rho_{XY} = 0.2$. A new random variable is defined as $Z = 3X + Y$. Write down an expression for the pdf of Z .

6.47 Two Gaussian random variables, X and Y , are independent. Their respective means are 5 and 3, and their respective variances are 1 and 2.

- (a) Write down expressions for their marginal pdfs.
 (b) Write down an expression for their joint pdf.
 (c) What is the mean of $Z_1 = X + Y$? $Z_2 = X - Y$?
 (d) What is the variance of $Z_1 = X + Y$? $Z_2 = X - Y$?
 (e) Write down an expression for the pdf of $Z_1 = X + Y$.
 (f) Write down an expression for the pdf of $Z_2 = X - Y$.

6.48 Two Gaussian random variables, X and Y , are independent. Their respective means are 4 and 2, and their respective variances are 3 and 5.

- (a) Write down expressions for their marginal pdfs.
- (b) Write down an expression for their joint pdf.
- (c) What is the mean of $Z_1 = 3X + Y$? $Z_2 = 3X - Y$?
- (d) What is the variance of $Z_1 = 3X + Y$? $Z_2 = 3X - Y$?
- (e) Write down an expression for the pdf of $Z_1 = 3X + Y$.

(f) Write down an expression for the pdf of $Z_2 = 3X - Y$.

6.49 Find the probabilities of the following random variables, with pdfs as given in Table 6.4, exceeding their means. That is, in each case, find the probability that $X \geq m_X$, where X is the respective random variable and m_X is its mean.

- (a) Uniform;
- (b) Rayleigh;
- (b) One-sided exponential.

Computer Exercises

6.1 In this exercise we examine a useful technique for generating a set of samples having a given pdf.

- (a) First, prove the following theorem: If X is a continuous random variable with cdf $F_X(x)$, the random variable

$$Y = F_X(X)$$

is a uniformly distributed random variable in the interval $(0, 1)$.

- (b) Using this theorem, design a random number generator to generate a sequence of exponentially distributed random variables having the pdf

$$f_X(x) = ae^{-ax}u(x)$$

where $u(x)$ is the unit step. Plot histograms of the random numbers generated to check the validity of the random number generator you designed.

6.2 An algorithm for generating a Gaussian random variable from two independent uniform random variables is easily derived.

- (a) Let U and V be two statistically independent random numbers uniformly distributed in $[0, 1]$. Show that the following transformation generates two statistically independent Gaussian random numbers with unit variance and zero mean:

$$\begin{aligned} X &= R \cos(2\pi U) \\ Y &= R \sin(2\pi U) \end{aligned}$$

where

$$R = \sqrt{-2 \ln(V)}$$

Hint: First show that R is Rayleigh.

- (b) Generate 1000 random variable pairs according to the above algorithm. Plot histograms for each

set (i.e., X and Y), and compare with Gaussian pdfs after properly scaling the histograms (i.e., divide each cell by the total number of counts times the cell width so that the histogram approximates a probability-density function). *Hint:* Use the hist function of MATLAB.

6.3 Using the results of Problem 6.26 and the Gaussian random number generator designed in Computer Exercise 6.2, design a Gaussian random number generator that will provide a specified correlation between adjacent samples. Let

$$P(\tau) = e^{-\alpha|\tau|}$$

and plot sequences of Gaussian random numbers for various choices of α . Show how stronger correlation between adjacent samples affects the variation from sample to sample. (*Note:* To get memory over more than adjacent samples, a digital filter should be used with independent Gaussian samples at the input.)

6.4 Check the validity of the central-limit theorem by repeatedly generating n independent uniformly distributed random variables in the interval $(-0.5, 0.5)$, forming the sum given by (6.187), and plotting the histogram. Do this for $N = 5, 10$, and 20 . Can you say anything qualitatively and quantitatively about the approach of the sums to Gaussian random numbers? Repeat for exponentially distributed component random variables (do Computer Exercise 6.1 first). Can you think of a drawback to the approach of summing uniformly distributed random variables to generating Gaussian random variables (*Hint:* Consider the probability of the sum of uniform random variables being greater than $0.5N$ or less than $-0.5N$. What are the same probabilities for a Gaussian random variable?)

CHAPTER 7

RANDOM SIGNALS AND NOISE

The mathematical background reviewed in Chapter 6 on probability theory provides the basis for developing the statistical description of random waveforms. The importance of considering such waveforms, as pointed out in Chapter 1, lies in the fact that noise in communication systems is due to unpredictable phenomena, such as the random motion of charge carriers in conducting materials and other unwanted sources.

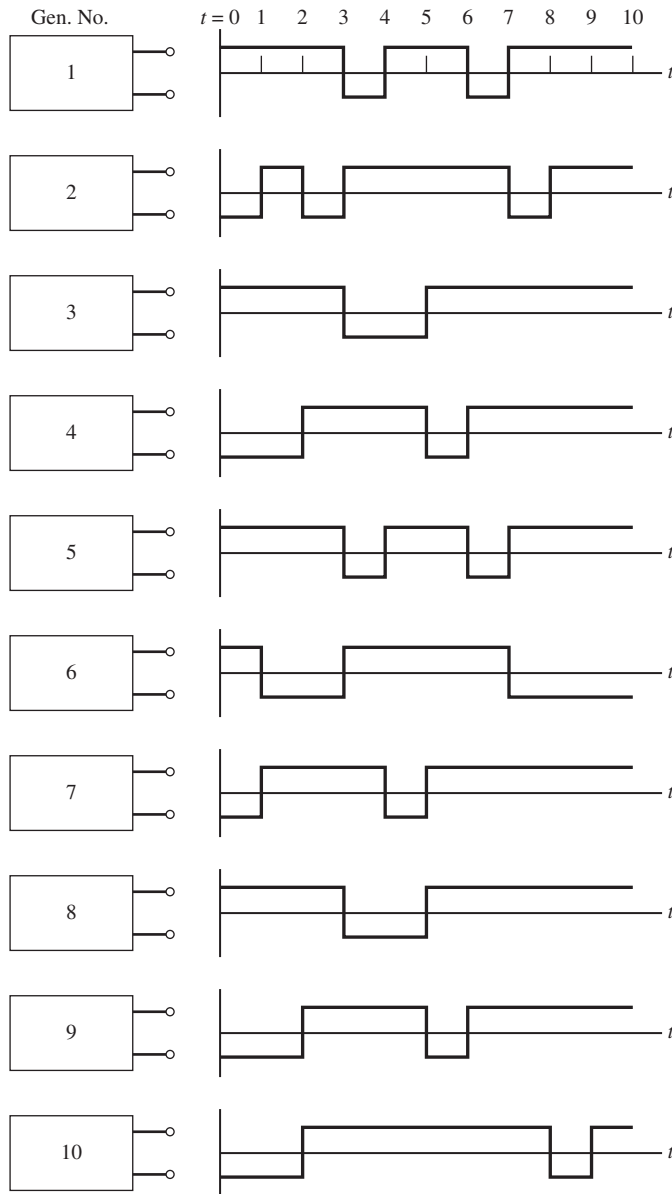
In the relative-frequency approach to probability, we imagined repeating the underlying chance experiment many times, the implication being that the replication process was carried out sequentially in time. In the study of random waveforms, however, the outcomes of the underlying chance experiments are mapped into functions of time, or waveforms, rather than numbers, as in the case of random variables. The particular waveform is not predictable in advance of the experiment, just as the particular value of a random variable is not predictable before the chance experiment is performed. We now address the statistical description of chance experiments that result in waveforms as outputs. To visualize how this may be accomplished, we again think in terms of relative frequency.

7.1 A RELATIVE-FREQUENCY DESCRIPTION OF RANDOM PROCESSES

For simplicity, consider a binary digital waveform generator whose output randomly switches between +1 and -1 in T_0 -second intervals as shown in Figure 7.1. Let $X(t, \zeta_i)$ be the random waveform corresponding to the output of the i th generator. Suppose relative frequency is used to estimate $P(X = +1)$ by examining the outputs of all generators at a particular time. Since the outputs are functions of time, we must specify the time when writing down the relative frequency. The following table may be constructed from an examination of the generator outputs in each time interval shown:

Time Interval:	(0,1)	(1,2)	(2,3)	(3,4)	(4,5)	(5,6)	(6,7)	(7,8)	(8,9)	(9,10)
Relative Frequency:	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{8}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{8}{10}$	$\frac{8}{10}$	$\frac{8}{10}$	$\frac{9}{10}$

From this table it is seen that the relative frequencies change with the time interval. Although this variation in relative frequency could be the result of *statistical irregularity*, we

**Figure 7.1**

A statistically identical set of binary waveform generators with typical outputs.

highly suspect that some phenomenon is making $X = +1$ more probable as time increases. To reduce the possibility that statistical irregularity is the culprit, we might repeat the experiment with 100 generators or 1000 generators. This is obviously a mental experiment in that it would be very difficult to obtain a set of identical generators and prepare them all in identical fashions.

7.2 SOME TERMINOLOGY OF RANDOM PROCESSES

7.2.1 Sample Functions and Ensembles

In the same fashion as is illustrated in Figure 7.1, we could imagine performing any chance experiment many times simultaneously. If, for example, the random quantity of interest is the voltage at the terminals of a noise generator, the random variable X_1 may be assigned to represent the possible values of this voltage at time t_1 and the random variable X_2 the values at time t_2 . As in the case of the digital waveform generator, we can imagine many noise generators all constructed in an identical fashion, insofar as we can make them, and run under identical conditions. Figure 7.2(a) shows typical waveforms generated in such an experiment. Each waveform $X(t, \zeta_i)$ is referred to as a *sample function*, where ζ_i is a member of a sample space \mathcal{S} . The totality of all sample functions is called an *ensemble*. The underlying chance experiment that gives rise to the ensemble of sample functions is called a *random*, or *stochastic, process*. Thus, to every outcome ζ we assign, according to a certain rule, a time function $X(t, \zeta)$. For a specific ζ , say ζ_i , $X(t, \zeta_i)$ signifies a single time function. For a specific time t_j , $X(t_j, \zeta)$ denotes a random variable. For fixed $t = t_j$ and fixed $\zeta = \zeta_i$, $X(t_j, \zeta_i)$ is a *number*. In what follows, we often suppress the ζ .

To summarize, the difference between a random variable and a random process is that for a random variable, an outcome in the sample space is mapped into a number, whereas for a random process it is mapped into a function of time.

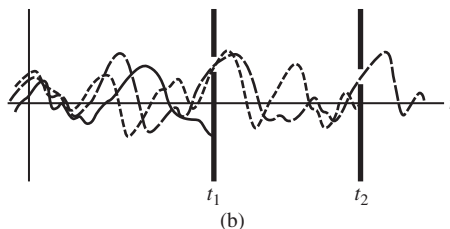
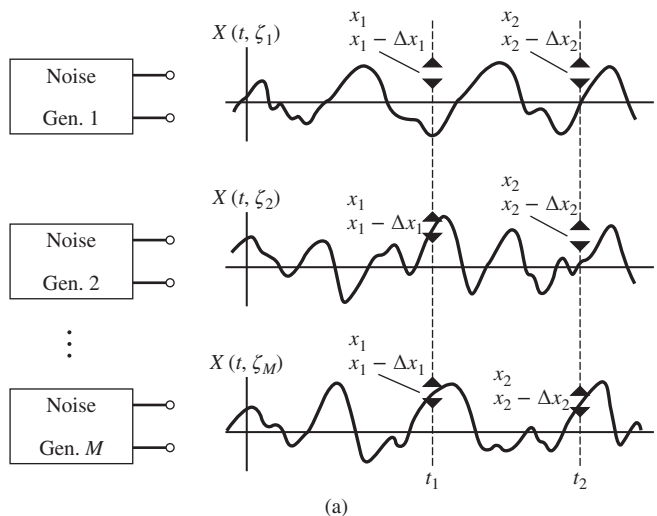


Figure 7.2

Typical sample functions of a random process and illustration of the relative-frequency interpretation of its joint pdf. (a) Ensemble of sample functions. (b) Superposition of the sample functions shown in (a).

7.2.2 Description of Random Processes in Terms of Joint pdfs

A complete description of a random process $\{X(t, \zeta)\}$ is given by the N -fold joint pdf that probabilistically describes the possible values assumed by a typical sample function at times $t_N > t_{N-1} > \dots > t_1$, where N is arbitrary. For $N = 1$, we can interpret this joint pdf $f_{X_1}(x_1, t_1)$ as

$$f_{X_1}(x_1, t_1)dx_1 = P(x_1 - dx_1 < X_1 \leq x_1 \text{ at time } t_1) \quad (7.1)$$

where $X_1 = X(t_1, \zeta)$. Similarly, for $N = 2$, we can interpret the joint pdf $f_{X_1 X_2}(x_1, t_1; x_2, t_2)$ as

$$f_{X_1 X_2}(x_1, t_1; x_2, t_2)dx_1 dx_2 = P(x_1 - dx_1 < X_1 \leq x_1 \text{ at time } t_1, \text{ and } x_2 - dx_2 < X_2 \leq x_2 \text{ at time } t_2) \quad (7.2)$$

where $X_2 = X(t_2, \zeta)$.

To help visualize the interpretation of (7.2), Figure 7.2(b) shows the three sample functions of Figure 7.2(a) superimposed with barriers placed at $t = t_1$ and $t = t_2$. According to the relative-frequency interpretation, the joint probability given by (7.2) is the number of sample functions that pass through the slits in both barriers divided by the total number M of sample functions as M becomes large without bound.

7.2.3 Stationarity

We have indicated the possible dependence of $f_{X_1 X_2}$ on t_1 and t_2 by including them in its argument. If $\{X(t)\}$ were a Gaussian random process, for example, its values at time t_1 and t_2 would be described by (6.187), where $m_X, m_Y, \sigma_X^2, \sigma_Y^2$, and ρ would, in general, depend on t_1 and t_2 .¹ Note that we need a general N -fold pdf to completely describe the random process $\{X(t)\}$. In general, such a pdf depends on N time instants t_1, t_2, \dots, t_N . In some cases, these joint pdfs depend only on the time differences $t_2 - t_1, t_3 - t_1, \dots, t_N - t_1$; that is, the choice of time origin for the random process is immaterial. Such random processes are said to be *statistically stationary in the strict sense*, or simply *stationary*.

For stationary processes, means and variances are independent of time, and the correlation coefficient (or covariance) depends only on the time difference $t_2 - t_1$.² Figure 7.3 contrasts sample functions of stationary and nonstationary processes. It may happen that in some cases the mean and variance of a random process are time-independent and the covariance is a function only of the time difference, but the N -fold joint pdf depends on the time origin. Such random processes are called *wide-sense stationary processes* to distinguish them from strictly stationary processes (that is, processes whose N -fold pdf is independent of time origin). Strict-sense stationarity implies wide-sense stationarity, but the reverse is not necessarily true. An exception occurs for *Gaussian random processes for which wide-sense stationarity does imply strict-sense stationarity*, since the joint Gaussian pdf is completely specified in terms of the means, variances, and covariances of $X(t_1), X(t_2), \dots, X(t_N)$.

¹For a stationary process, all joint moments are independent of time origin. We are interested primarily in the covariance, however.

²At N instants of time, its values would be described by (B.1) of Appendix B.

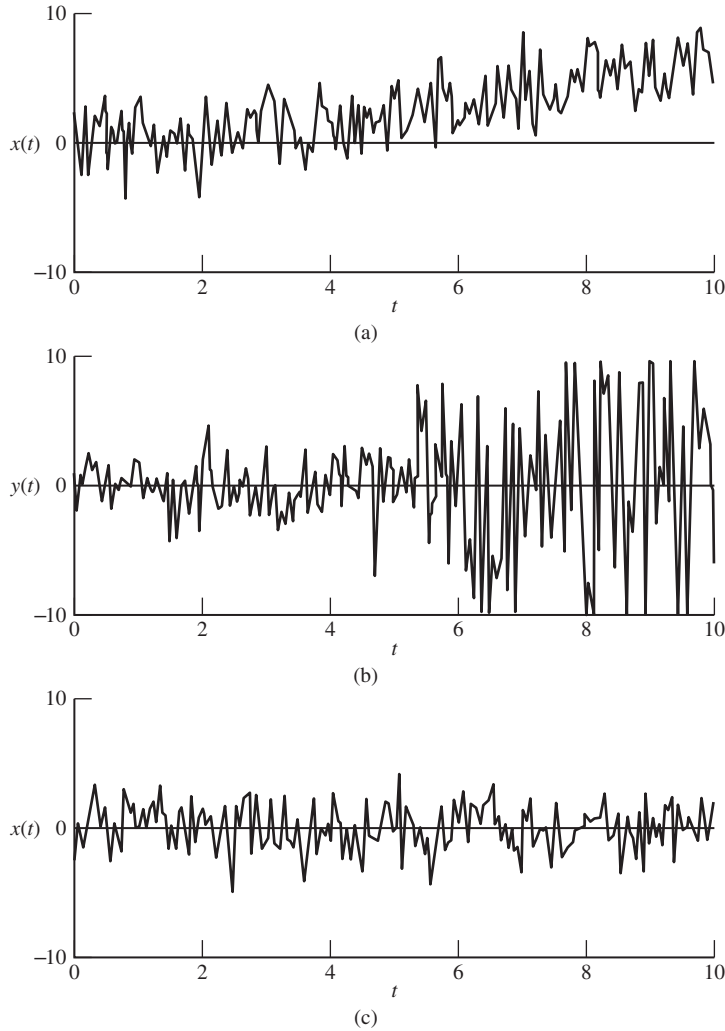


Figure 7.3 Sample functions of nonstationary processes contrasted with a sample function of a stationary process. (a) Time-varying mean. (b) Time-varying variance. (c) Stationary.

7.2.4 Partial Description of Random Processes: Ergodicity

As in the case of random variables, we may not always require a complete statistical description of a random process, or we may not be able to obtain the N -fold joint pdf even if desired. In such cases, we work with various moments, either by choice or by necessity. The most important averages are the mean,

$$m_X(t) = E[X(t)] = \overline{X(t)} \quad (7.3)$$

the variance,

$$\sigma_X^2(t) = E \left\{ [X(t) - \overline{X(t)}]^2 \right\} = \overline{X^2(t)} - \overline{X(t)}^2 \quad (7.4)$$

and the covariance,

$$\begin{aligned}\mu_X(t, t + \tau) &= E \left\{ [X(t) - \overline{X(t)}][X(t + \tau) - \overline{X(t + \tau)}] \right\} \\ &= E[X(t)X(t + \tau)] - \overline{X(t)} \overline{X(t + \tau)}\end{aligned}\quad (7.5)$$

In (7.5), we let $t = t_1$ and $t + \tau = t_2$. The first term on the right-hand side is the *autocorrelation function* computed as a *statistical*, or *ensemble*, *average* (that is, the average is across the sample functions at times t and $t + \tau$). In terms of the joint pdf of the random process, the autocorrelation function is

$$R_X(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1 X_2}(x_1, t_1; x_2, t_2) dx_1 dx_2 \quad (7.6)$$

where $X_1 = X(t_1)$ and $X_2 = X(t_2)$. If the process is wide-sense stationary, $f_{X_1 X_2}$ does not depend on t but rather on the time difference, $\tau = t_2 - t_1$ and as a result, $R_X(t_1, t_2) = R_X(\tau)$ is a function only of τ . A very important question is: ‘‘If the autocorrelation function using the definition of a time average as given in Chapter 2 is used, will the result be the same as the statistical average given by (7.6)?’’ For many processes, referred to as *ergodic*, the answer is affirmative. Ergodic processes are processes for which *time and ensemble averages are interchangeable*. Thus, if $X(t)$ is an ergodic process, all time and the corresponding ensemble averages are interchangeable. In particular,

$$m_X = E[X(t)] = \langle X(t) \rangle \quad (7.7)$$

$$\sigma_X^2 = E \left\{ [X(t) - \overline{X(t)}]^2 \right\} = \left\langle [X(t) - \langle X(t) \rangle]^2 \right\rangle \quad (7.8)$$

and

$$R_X(\tau) = E[X(t)X(t + \tau)] = \langle X(t)X(t + \tau) \rangle \quad (7.9)$$

where

$$\langle v(t) \rangle \triangleq \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T v(t) dt \quad (7.10)$$

as defined in Chapter 2. We emphasize that for ergodic processes *all* time and ensemble averages are interchangeable, not just the mean, variance, and autocorrelation function.

EXAMPLE 7.1

Consider the random process with sample functions³

$$n(t) = A \cos(2\pi f_0 t + \Theta)$$

where f_0 is a constant and Θ is a random variable with the pdf

$$f_{\Theta}(\theta) = \begin{cases} \frac{1}{2\pi}, & |\theta| \leq \pi \\ 0, & \text{otherwise} \end{cases} \quad (7.11)$$

³In this example we violate our earlier established convention that sample functions are denoted by capital letters. This is quite often done if confusion will not result.

Computed as statistical averages, the first and second moments are

$$\begin{aligned}\overline{n(t)} &= \int_{-\infty}^{\infty} A \cos(2\pi f_0 t + \theta) f_{\Theta}(\theta) d\theta \\ &= \int_{-\pi}^{\pi} A \cos(2\pi f_0 t + \theta) \frac{d\theta}{2\pi} = 0\end{aligned}\quad (7.12)$$

and

$$\overline{n^2(t)} = \int_{-\pi}^{\pi} A^2 \cos^2(2\pi f_0 t + \theta) \frac{d\theta}{2\pi} = \frac{A^2}{4\pi} \int_{-\pi}^{\pi} [1 + \cos(4\pi f_0 t + 2\theta)] d\theta = \frac{A^2}{2}\quad (7.13)$$

respectively. The variance is equal to the second moment, since the mean is zero.

Computed as time averages, the first and second moments are

$$\langle n(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T A \cos(2\pi f_0 t + \Theta) dt = 0\quad (7.14)$$

and

$$\langle n^2(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T A^2 \cos^2(2\pi f_0 t + \Theta) dt = \frac{A^2}{2}\quad (7.15)$$

respectively. In general, the time average of some function of an ensemble member of a random process is a random variable. In this example, $\langle n(t) \rangle$ and $\langle n^2(t) \rangle$ are constants! We suspect that this random process is stationary and ergodic, even though the preceding results do not *prove* this. It turns out that this is indeed true.

To continue the example, consider the pdf

$$f_{\Theta}(\theta) = \begin{cases} \frac{2}{\pi}, & |\theta| \leq \frac{1}{4}\pi \\ 0, & \text{otherwise} \end{cases}\quad (7.16)$$

For this case, the expected value, or mean, of the random process computed at an arbitrary time t is

$$\begin{aligned}\overline{n(t)} &= \int_{-\pi/4}^{\pi/4} A \cos(2\pi f_0 t + \theta) \frac{2}{\pi} d\theta \\ &= \frac{2}{\pi} A \sin(2\pi f_0 t + \theta) \Big|_{-\pi/4}^{\pi/4} = \frac{2\sqrt{2}A}{\pi} \cos \omega_0 t\end{aligned}\quad (7.17)$$

The second moment, computed as a statistical average, is

$$\begin{aligned}\overline{n^2(t)} &= \int_{-\pi/4}^{\pi/4} A^2 \cos^2(2\pi f_0 t + \theta) \frac{2}{\pi} d\theta \\ &= \int_{-\pi/4}^{\pi/4} \frac{A^2}{\pi} [1 + \cos(4\pi f_0 t + 2\theta)] d\theta \\ &= \frac{A^2}{2} + \frac{A^2}{\pi} \cos 4\pi f_0 t\end{aligned}\quad (7.18)$$

Since stationarity of a random process implies that all moments are independent of time origin, these results show that this process is not stationary. In order to comprehend the physical reason for this, you should sketch some typical sample functions. In addition, this process cannot be ergodic, since ergodicity requires stationarity. Indeed, the time average first and second moments are still $\langle n(t) \rangle = 0$ and $\langle n^2(t) \rangle = \frac{1}{2}A^2$, respectively. Thus, we have exhibited two time averages that are not equal to the corresponding statistical averages. ■

7.2.5 Meanings of Various Averages for Ergodic Processes

It is useful to pause at this point and summarize the meanings of various averages for an ergodic process:

1. The mean $\overline{X(t)} = \langle X(t) \rangle$ is the dc component.
2. $\overline{X(t)^2} = \langle X(t)^2 \rangle$ is the dc power.
3. $\overline{X^2(t)} = \langle X^2(t) \rangle$ is the total power.
4. $\sigma_X^2 = \overline{X^2(t)} - \overline{X(t)}^2 = \langle X^2(t) \rangle - \langle X(t) \rangle^2$ is the power in the ac (time-varying) component.
5. The total power $\overline{X^2(t)} = \sigma_X^2 + \overline{X(t)}^2$ is the ac power plus the dc power.

Thus, in the case of ergodic processes, we see that these moments are measurable quantities in the sense that they can be replaced by the corresponding time averages and a finite-time approximation to these time averages can be measured in the laboratory.

EXAMPLE 7.2

To illustrate some of the definitions given above with regard to correlation functions, let us consider a random telegraph waveform $X(t)$, as illustrated in Figure 7.4. The sample functions of this random process have the following properties:

1. The values taken on at any time instant t_0 are either $X(t_0) = A$ or $X(t_0) = -A$ with equal probability.
2. The number k of switching instants in any time interval T obeys a Poisson distribution, as defined by (6.182), with the attendant assumptions leading to this distribution. (That is, the probability of more than one switching instant occurring in an infinitesimal time interval dt is zero, with the probability of exactly one switching instant occurring in dt being αdt , where α is a constant. Furthermore, successive switching occurrences are independent.)

If τ is any positive time increment, the autocorrelation function of the random process defined by the preceding properties can be calculated as

$$\begin{aligned} R_X(\tau) &= E[X(t) X(t + \tau)] \\ &= A^2 P[X(t) \text{ and } X(t + \tau) \text{ have the same sign}] \\ &\quad + (-A^2) P[X(t) \text{ and } X(t + \tau) \text{ have different signs}] \\ &= A^2 P[\text{even number of switching instants in } (t, t + \tau)] \\ &\quad - A^2 P[\text{odd number of switching instants in } (t, t + \tau)] \end{aligned}$$

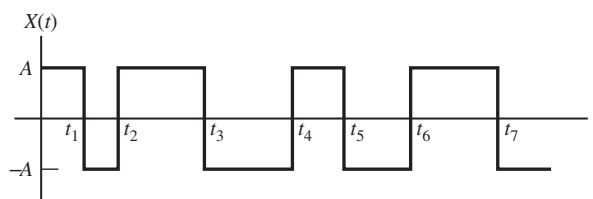


Figure 7.4
Sample function of a random telegraph waveform.

$$\begin{aligned}
&= A^2 \sum_{\substack{k=0 \\ k \text{ even}}}^{\infty} \frac{(\alpha\tau)^k}{k!} \exp(-\alpha\tau) - A^2 \sum_{\substack{k=0 \\ k \text{ odd}}}^{\infty} \frac{(\alpha\tau)^k}{k!} \exp(-\alpha\tau) \\
&= A^2 \exp(-\alpha\tau) \sum_{k=0}^{\infty} \frac{(-\alpha\tau)^k}{k!} \\
&= A^2 \exp(-\alpha\tau) \exp(-\alpha\tau) = A^2 \exp(-2\alpha\tau) \tag{7.19}
\end{aligned}$$

The preceding development was carried out under the assumption that τ was positive. It could have been similarly carried out with τ negative, such that

$$R_X(\tau) = E[X(t)X(t - |\tau|)] = E[X(t - |\tau|)X(t)] = A^2 \exp(-2\alpha|\tau|) \tag{7.20}$$

This is a result that holds for all τ . That is, $R_X(\tau)$ is an even function of τ , which we will show in general shortly. ■

7.3 CORRELATION AND POWER SPECTRAL DENSITY

The autocorrelation function, computed as a statistical average, has been defined by (7.6). If a process is ergodic, the autocorrelation function computed as a time average, as first defined in Chapter 2, is equal to the statistical average of (7.6). In Chapter 2, we defined the power spectral density $S(f)$ as the Fourier transform of the autocorrelation function $R(\tau)$. The *Wiener-Khinchine* theorem is a formal statement of this result for stationary random processes, for which $R(t_1, t_2) = R(t_2 - t_1) = R(\tau)$. For such processes, previously defined as wide-sense stationary, the power spectral density and autocorrelation function are Fourier-transform pairs. That is,

$$S(f) \xleftrightarrow{\mathfrak{F}} R(\tau) \tag{7.21}$$

If the process is ergodic, $R(\tau)$ can be calculated as either a time or an ensemble average.

Since $R_X(0) = \overline{X^2(t)}$ is the average power contained in the process, we have from the inverse Fourier transform of $S_X(f)$ that

$$\text{Average power} = R_X(0) = \int_{-\infty}^{\infty} S_X(f) df \tag{7.22}$$

which is reasonable, since the definition of $S_X(f)$ is that it is *power density* with respect to frequency.

7.3.1 Power Spectral Density

An intuitively satisfying, and in some cases computationally useful, expression for the power spectral density of a stationary random process can be obtained by the following approach. Consider a particular sample function, $n(t, \zeta_i)$, of a stationary random process. To obtain a function giving power density versus frequency using the Fourier transform, we consider a

truncated version, $n_T(t, \zeta_i)$, defined as⁴

$$n_T(t, \zeta_i) = \begin{cases} n(t, \zeta_i) & |t| < \frac{1}{2}T \\ 0, & \text{otherwise} \end{cases} \quad (7.23)$$

Since sample functions of stationary random processes are power signals, the Fourier transform of $n(t, \zeta_i)$ does not exist, which necessitates defining $n_T(t, \zeta_i)$. The Fourier transform of a truncated sample function is

$$N_T(f, \zeta_i) = \int_{-T/2}^{T/2} n(t, \zeta_i) e^{-j2\pi f t} dt \quad (7.24)$$

and its energy spectral density, according to Equation (2.90), is $|N_T(f, \zeta_i)|^2$. The time-average power density over the interval $[-\frac{1}{2}T, \frac{1}{2}T]$ for this sample function is $|N_T(f, \zeta_i)|^2 / T$. Since this time-average power density depends on the particular sample function chosen, we perform an ensemble average and take the limit as $T \rightarrow \infty$ to obtain the distribution of power density with frequency. This is defined as the power spectral density $S_n(f)$, which can be expressed as

$$S_n(f) = \lim_{T \rightarrow \infty} \frac{\overline{|N_T(f, \zeta_i)|^2}}{T} \quad (7.25)$$

The operations of taking the limit and taking the ensemble average in (7.25) cannot be interchanged.

EXAMPLE 7.3

Let us find the power spectral density of the random process considered in Example 7.1 using (7.25). In this case,

$$n_T(t, \Theta) = A\Pi\left(\frac{t}{T}\right) \cos\left[2\pi f_0\left(t + \frac{\Theta}{2\pi f_0}\right)\right] \quad (7.26)$$

By the time-delay theorem of Fourier transforms and using the transform pair

$$\cos 2\pi f_0 t \leftrightarrow \frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0) \quad (7.27)$$

we obtain

$$\mathfrak{F}[\cos(2\pi f_0 t + \Theta)] = \frac{1}{2}\delta(f - f_0) e^{j\Theta} + \frac{1}{2}\delta(f + f_0) e^{-j\Theta} \quad (7.28)$$

We also recall from Chapter 2 (Example 2.8) that $\Pi(t/T) \leftrightarrow T \text{sinc} T f$, so, by the multiplication theorem of Fourier transforms,

$$\begin{aligned} N_T(f, \Theta) &= (AT \text{sinc} T f) * \left[\frac{1}{2}\delta(f - f_0) e^{j\Theta} + \frac{1}{2}\delta(f + f_0) e^{-j\Theta} \right] \\ &= \frac{1}{2} AT \left[e^{j\Theta} \text{sinc}(f - f_0) T + e^{-j\Theta} \text{sinc}(f + f_0) T \right] \end{aligned} \quad (7.29)$$

⁴Again, we use a lowercase letter to denote a random process for the simple reason that we need to denote the Fourier transform of $n(t)$ by an uppercase letter.

Therefore, the energy spectral density of the truncated sample function is

$$\begin{aligned} |N_T(f, \Theta)|^2 = & \left(\frac{1}{2}AT\right)^2 \{ \text{sinc}^2 T(f - f_0) + e^{2j\Theta} \text{sinc} T(f - f_0) \text{sinc} T(f + f_0) \\ & + e^{-2j\Theta} \text{sinc} T(f - f_0) \text{sinc} T(f + f_0) + \text{sinc}^2 T(f + f_0) \} \end{aligned} \quad (7.30)$$

In obtaining $\overline{|N_T(f, \Theta)|^2}$, we note that

$$\overline{\exp(\pm j2\Theta)} = \int_{-\pi}^{\pi} e^{\pm j2\Theta} \frac{d\theta}{2\pi} = \int_{-\pi}^{\pi} (\cos 2\theta \pm j \sin 2\theta) \frac{d\theta}{2\pi} = 0 \quad (7.31)$$

Thus, we obtain

$$\overline{|N_T(f, \Theta)|^2} = \left(\frac{1}{2}AT\right)^2 [\text{sinc}^2 T(f - f_0) + \text{sinc}^2 T(f + f_0)] \quad (7.32)$$

and the power spectral density is

$$S_n(f) = \lim_{T \rightarrow \infty} \frac{1}{4} A^2 [T \text{sinc}^2 T(f - f_0) + T \text{sinc}^2 T(f + f_0)] \quad (7.33)$$

However, a representation of the delta function is $\lim_{T \rightarrow \infty} T \text{sinc}^2 Tu = \delta(u)$. [See Figure 2.4(b).] Thus,

$$S_n(f) = \frac{1}{4} A^2 \delta(f - f_0) + \frac{1}{4} A^2 \delta(f + f_0) \quad (7.34)$$

The average power is $\int_{-\infty}^{\infty} S_n(f) df = \frac{1}{2} A^2$, the same as obtained in Example 7.1. ■

7.3.2 The Wiener–Khinchine Theorem

The Wiener–Khinchine theorem states that the autocorrelation function and power spectral density of a stationary random process are Fourier-transform pairs. It is the purpose of this subsection to provide a formal proof of this statement.

To simplify the notation in the proof of the Wiener–Khinchine theorem, we rewrite (7.25) as

$$S_n(f) = \lim_{T \rightarrow \infty} \frac{E \left\{ \left| \mathfrak{F} [n_{2T}(t)] \right|^2 \right\}}{2T} \quad (7.35)$$

where, for convenience, we have truncated over a $2T$ -second interval and dropped ζ in the argument of $n_{2T}(t)$. Note that

$$\begin{aligned} \left| \mathfrak{F} [n_{2T}(t)] \right|^2 &= \left| \int_{-T}^T n(t) e^{-j\omega t} dt \right|^2, \quad \omega = 2\pi f \\ &= \int_{-T}^T \int_{-T}^T n(t) n(\sigma) e^{-j\omega(t-\sigma)} dt d\sigma \end{aligned} \quad (7.36)$$

where the product of two integrals has been written as an iterated integral. Taking the ensemble average and interchanging the orders of averaging and integration, we obtain

$$\begin{aligned} E \left\{ \left| \mathfrak{F} [n_{2T}(t)] \right|^2 \right\} &= \int_{-T}^T \int_{-T}^T E \{ n(t) n(\sigma) \} e^{-j\omega(t-\sigma)} dt d\sigma \\ &= \int_{-T}^T \int_{-T}^T R_n(t - \sigma) e^{-j\omega(t-\sigma)} dt d\sigma \end{aligned} \quad (7.37)$$

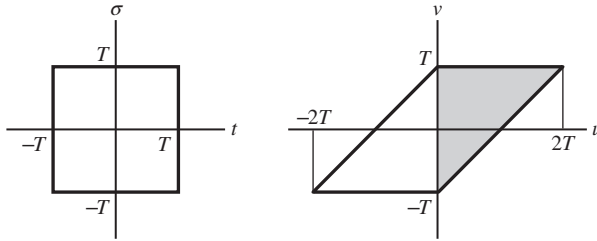


Figure 7.5
Regions of integration for
Equation (7.37).

by the definition of the autocorrelation function. The change of variables $u = t - \sigma$ and $v = t$ is now made with the aid of Figure 7.5. In the uv plane we integrate over v first and then over u by breaking the integration over u up into two integrals, one for u negative and one for u positive. Thus,

$$\begin{aligned}
 & E \left\{ \left| \mathfrak{F} [n_{2T}(t)] \right|^2 \right\} \\
 &= \int_{u=-2T}^0 R_n(u) e^{-j\omega u} \left(\int_{-T}^{u+T} dv \right) du + \int_{u=0}^{2T} R_n(u) e^{-j\omega u} \left(\int_{u-T}^T dv \right) du \\
 &= \int_{-2T}^0 (2T+u) R_n(u) e^{-j\omega u} + \int_0^{2T} (2T-u) R_n(u) e^{-j\omega u} du \\
 &= 2T \int_{-2T}^{2T} \left(1 - \frac{|u|}{2T} \right) R_n(u) e^{-j\omega u} du \tag{7.38}
 \end{aligned}$$

The power spectral density is, by (7.35),

$$S_n(f) = \lim_{T \rightarrow \infty} \int_{-2T}^{2T} \left(1 - \frac{|u|}{2T} \right) R_n(u) e^{-j\omega u} du \tag{7.39}$$

which is the limit as $T \rightarrow \infty$ results in (7.21).

EXAMPLE 7.4

Since the power spectral density and the autocorrelation function are Fourier-transform pairs, the autocorrelation function of the random process defined in Example 7.1 is, from the result of Example 7.3, given by

$$\begin{aligned}
 R_n(\tau) &= \mathfrak{F}^{-1} \left[\frac{1}{4} A^2 \delta(f - f_0) + \frac{1}{4} A^2 \delta(f + f_0) \right] \\
 &= \frac{1}{2} A^2 \cos(2\pi f_0 \tau) \tag{7.40}
 \end{aligned}$$

Computing $R_n(\tau)$ as an ensemble average, we obtain

$$\begin{aligned}
 R_n(\tau) &= E \{ n(t)n(t+\tau) \} \\
 &= \int_{-\pi}^{\pi} A^2 \cos(2\pi f_0 t + \theta) \cos[2\pi f_0(t+\tau) + \theta] \frac{d\theta}{2\pi}
 \end{aligned}$$

$$\begin{aligned}
&= \frac{A^2}{4\pi} \int_{-\pi}^{\pi} \{\cos 2\pi f_0 \tau + \cos[2\pi f_0(2t + \tau) + 2\theta]\} d\theta \\
&= \frac{1}{2} A^2 \cos(2\pi f_0 \tau)
\end{aligned} \tag{7.41}$$

which is the same result as that obtained using the Wiener–Khinchine theorem. ■

7.3.3 Properties of the Autocorrelation Function

The properties of the autocorrelation function for a stationary random process $X(t)$ were stated in Chapter 2, at the end of Section 2.6, and all time averages may now be replaced by statistical averages. These properties are now easily proved.

Property 1 states that $|R(\tau)| \leq R(0)$ for all τ . To show this, consider the nonnegative quantity

$$[X(t) \pm X(t + \tau)]^2 \geq 0 \tag{7.42}$$

where $\{X(t)\}$ is a stationary random process. Squaring and averaging term by term, we obtain

$$\overline{X^2(t) \pm 2X(t)X(t + \tau) + X^2(t + \tau)} \geq 0 \tag{7.43}$$

which reduces to

$$2R(0) \pm 2R(\tau) \geq 0 \text{ or } -R(0) \leq R(\tau) \leq R(0) \tag{7.44}$$

because $\overline{X^2(t)} = \overline{X^2(t + \tau)} = R(0)$ by the stationarity of $\{X(t)\}$.

Property 2 states that $R(-\tau) = R(\tau)$. This is easily proved by noting that

$$R(\tau) \triangleq \overline{X(t)X(t + \tau)} = \overline{X(t' - \tau)X(t')} = \overline{X(t')X(t' - \tau)} \triangleq R(-\tau) \tag{7.45}$$

where the change of variables $t' = t + \tau$ has been made.

Property 3 states that $\lim_{|\tau| \rightarrow \infty} R(\tau) = \overline{X(t)}^2$ if $\{X(t)\}$ does not contain a periodic component. To show this, we note that

$$\begin{aligned}
\lim_{|\tau| \rightarrow \infty} R(\tau) &\triangleq \lim_{|\tau| \rightarrow \infty} \overline{X(t)X(t + \tau)} \\
&\cong \overline{X(t)} \overline{X(t + \tau)}, \text{ where } |\tau| \text{ is large} \\
&= \overline{X(t)}^2
\end{aligned} \tag{7.46}$$

where the second step follows intuitively because the interdependence between $X(t)$ and $X(t + \tau)$ becomes smaller as $|\tau| \rightarrow \infty$ (if no periodic components are present), and the last step results from the stationarity of $\{X(t)\}$.

Property 4, which states that $R(\tau)$ is periodic if $\{X(t)\}$ is periodic, follows by noting from the time-average definition of the autocorrelation function given by Equation (2.161) that periodicity of the integrand implies periodicity of $R(\tau)$.

Finally, Property 5, which says that $\Im[R(\tau)]$ is nonnegative, is a direct consequence of the Wiener–Khinchine theorem (7.21) and (7.25) from which it is seen that the power spectral density is nonnegative.

EXAMPLE 7.5

Processes for which

$$S(f) = \begin{cases} \frac{1}{2}N_0, & |f| \leq B \\ 0, & \text{otherwise} \end{cases} \quad (7.47)$$

where N_0 is constant, are commonly referred to as *bandlimited white noise*, since, as $B \rightarrow \infty$, all frequencies are present, in which case the process is simply called *white*. N_0 is the single-sided power spectral density of the nonbandlimited process. For a bandlimited white-noise process,

$$\begin{aligned} R(\tau) &= \int_{-B}^B \frac{1}{2}N_0 \exp(j2\pi f\tau) df \\ &= \frac{N_0}{2} \frac{\exp(j2\pi f\tau)}{j2\pi\tau} \Big|_{-B}^B = BN_0 \frac{\sin(2\pi B\tau)}{2\pi B\tau} \\ &= BN_0 \text{sinc}2B\tau \end{aligned} \quad (7.48)$$

As $B \rightarrow \infty$, $R(\tau) \rightarrow \frac{1}{2}N_0\delta(\tau)$. That is, no matter how close together we sample a white-noise process, the samples have zero correlation. If, in addition, the process is Gaussian, the samples are independent. A white-noise process has infinite power and is therefore a mathematical idealization, but it is nevertheless useful in systems analysis. ■

7.3.4 Autocorrelation Functions for Random Pulse Trains

As another example of calculating autocorrelation functions, consider a random process with sample functions that can be expressed as

$$X(t) = \sum_{k=-\infty}^{\infty} a_k p(t - kT - \Delta) \quad (7.49)$$

where $\dots a_{-1}, a_0, a_1, \dots, a_k \dots$ is a doubly-infinite sequence of random variables with

$$E[a_k a_{k+m}] = R_m \quad (7.50)$$

The function $p(t)$ is a deterministic pulse-type waveform where T is the separation between pulses; Δ is a random variable that is independent of the value of a_k and uniformly distributed in the interval $(-T/2, T/2)$.⁵ The autocorrelation function of this waveform is

$$\begin{aligned} R_X(\tau) &= E[X(t)X(t+\tau)] \\ &= E \left\{ \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} a_k a_{k+m} p(t - kT - \Delta) p[t + \tau - (k+m)T - \Delta] \right\} \end{aligned} \quad (7.51)$$

⁵Including the random variable Δ in the definition of the sample functions for the process guarantees wide-sense stationarity. If it weren't included, $X(t)$ would be what is referred to as a cyclostationary random process.

Taking the expectation inside the double sum and making use of the independence of the sequence $\{a_k a_{k+m}\}$ and the delay variable Δ , we obtain

$$\begin{aligned} R_X(\tau) &= \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} E[a_k a_{k+m}] E\{p(t - kT - \Delta) p[t + \tau - (k + m)T - \Delta]\} \\ &= \sum_{m=-\infty}^{\infty} R_m \sum_{k=-\infty}^{\infty} \int_{-T/2}^{T/2} p(t - kT - \Delta) p[t + \tau - (k + m)T - \Delta] \frac{d\Delta}{T} \end{aligned} \quad (7.52)$$

The change of variables $u = t - kT - \Delta$ inside the integral results in

$$\begin{aligned} R_X(\tau) &= \sum_{m=-\infty}^{\infty} R_m \sum_{k=-\infty}^{\infty} \int_{t-(k+1/2)T}^{t-(k-1/2)T} p(u) p(u + \tau - mT) \frac{du}{T} \\ &= \sum_{m=-\infty}^{\infty} R_m \left[\frac{1}{T} \int_{-\infty}^{\infty} p(u + \tau - mT) p(u) du \right] \end{aligned} \quad (7.53)$$

Finally we have

$$R_X(\tau) = \sum_{m=-\infty}^{\infty} R_m r(\tau - mT) \quad (7.54)$$

where

$$r(\tau) \triangleq \frac{1}{T} \int_{-\infty}^{\infty} p(t + \tau) p(t) dt \quad (7.55)$$

is the pulse-correlation function. We consider the following example as an illustration.

EXAMPLE 7.6

In this example we consider a situation where the sequence $\{a_k\}$ has memory built into it by the relationship

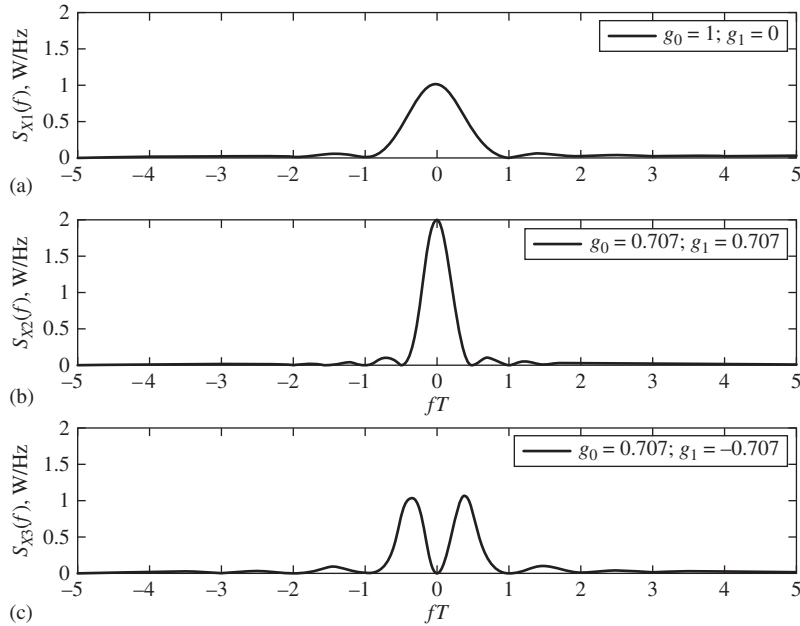
$$a_k = g_0 A_k + g_1 A_{k-1} \quad (7.56)$$

where g_0 and g_1 are constants and the A_k 's are random variables such that $A_k = \pm A$ where the sign is determined by a random coin toss independently from pulse to pulse for all k (note that if $g_1 = 0$, there is no memory). It can be shown that

$$E[a_k a_{k+m}] = \begin{cases} (g_0^2 + g_1^2) A^2, & m = 0 \\ g_0 g_1 A^2, & m = \pm 1 \\ 0, & \text{otherwise} \end{cases} \quad (7.57)$$

The assumed pulse shape is $p(t) = \Pi\left(\frac{t}{T}\right)$ so that the pulse-correlation function is

$$\begin{aligned} r(\tau) &= \frac{1}{T} \int_{-\infty}^{\infty} \Pi\left(\frac{t+\tau}{T}\right) \Pi\left(\frac{t}{T}\right) dt \\ &= \frac{1}{T} \int_{-T/2}^{T/2} \Pi\left(\frac{t+\tau}{T}\right) dt = \Lambda\left(\frac{\tau}{T}\right) \end{aligned} \quad (7.58)$$

**Figure 7.6**

Power spectra of binary-valued waveforms. (a) Case in which there is no memory. (b) Case in which there is reinforcing memory between adjacent pulses. (c) Case where the memory between adjacent pulses is antipodal.

where, from Chapter 2, $\Lambda\left(\frac{\tau}{T}\right)$ is a unit-height triangular pulse symmetrical about $t = 0$ of width $2T$. Thus, the autocorrelation function (7.58) becomes

$$R_X(\tau) = A^2 \left\{ [g_0^2 + g_1^2] \Lambda\left(\frac{\tau}{T}\right) + g_0 g_1 \left[\Lambda\left(\frac{\tau+T}{T}\right) + \Lambda\left(\frac{\tau-T}{T}\right) \right] \right\} \quad (7.59)$$

Applying the Wiener–Khinchine theorem, the power spectral density of $X(t)$ is found to be

$$S_X(f) = \mathfrak{F}[R_X(\tau)] = A^2 T \text{sinc}^2(fT) [g_0^2 + g_1^2 + 2g_0 g_1 \cos(2\pi fT)] \quad (7.60)$$

Figure 7.6 compares the power spectra for the two cases: (1) $g_0 = 1$ and $g_1 = 0$ (i.e., no memory); (2) $g_0 = g_1 = 1/\sqrt{2}$ (reinforcing memory between adjacent pulses). For case 1, the resulting power spectral density is

$$S_X(f) = A^2 T \text{sinc}^2(fT) \quad (7.61)$$

while for case (2) it is

$$S_X(f) = 2A^2 T \text{sinc}^2(fT) \cos^2(\pi fT) \quad (7.62)$$

In both cases, g_0 and g_1 have been chosen to give a total power of 1 W, which is verified from the plots by numerical integration. Note that in case 2 memory has confined the power spectrum more than without it. Yet a third case is shown in the bottom plot for which (3) $g_0 = -g_1 = 1/\sqrt{2}$. Now the spectral width is doubled over case 2, but a spectral null appears at $f = 0$.

Other values for g_0 and g_1 can be assumed, and memory between more than just adjacent pulses also can be assumed. ■

7.3.5 Cross-Correlation Function and Cross-Power Spectral Density

Suppose we wish to find the power in the sum of two noise voltages $X(t)$ and $Y(t)$. We might ask if we can simply add their separate powers. The answer is, in general, no. To see why, consider

$$n(t) = X(t) + Y(t) \quad (7.63)$$

where $X(t)$ and $Y(t)$ are two stationary random voltages that may be related (that is, that are not necessarily statistically independent). The power in the sum is

$$\begin{aligned} E[n^2(t)] &= E\{[X(t) + Y(t)]^2\} \\ &= E[X^2(t)] + 2E[X(t)Y(t)] + E[Y^2(t)] \\ &= P_X + 2P_{XY} + P_Y \end{aligned} \quad (7.64)$$

where P_X and P_Y are the powers of $X(t)$ and $Y(t)$, respectively, and P_{XY} is the cross power. More generally, we define the *cross-correlation function* as

$$R_{XY}(\tau) = E\{X(t)Y(t + \tau)\} \quad (7.65)$$

In terms of the cross-correlation function, $P_{XY} = R_{XY}(0)$. A *sufficient* condition for P_{XY} to be zero, so that we may simply add powers to obtain total power, is that

$$R_{XY}(0) = 0, \text{ for all } \tau \quad (7.66)$$

Such processes are said to be *orthogonal*. If two processes are statistically independent and at least one of them has zero mean, they are orthogonal. However, orthogonal processes are not necessarily statistically independent.

Cross-correlation functions can be defined for nonstationary processes also, in which case we have a function of two independent variables. We will not need to be this general in our considerations.

A useful symmetry property of the cross-correlation function for jointly stationary processes is

$$R_{XY}(\tau) = R_{YX}(-\tau) \quad (7.67)$$

which can be shown as follows. By definition,

$$R_{XY}(\tau) = E[X(t)Y(t + \tau)] \quad (7.68)$$

Defining $t' = t + \tau$, we obtain

$$R_{XY}(\tau) = E[Y(t')X(t' - \tau)] \triangleq R_{YX}(-\tau) \quad (7.69)$$

since the choice of time origin is immaterial for stationary processes.

The cross-power spectral density of two stationary random processes is defined as the Fourier transform of their cross-correlation function:

$$S_{XY}(f) = \mathfrak{F}[R_{XY}(\tau)] \quad (7.70)$$

It provides, in the frequency domain, the same information about the random processes as does the cross-correlation function.

7.4 LINEAR SYSTEMS AND RANDOM PROCESSES

7.4.1 Input-Output Relationships

In the consideration of the transmission of stationary random waveforms through fixed linear systems, a basic tool is the relationship of the output power spectral density to the input power spectral density, given as

$$S_y(f) = |H(f)|^2 S_x(f) \quad (7.71)$$

The autocorrelation function of the output is the inverse Fourier transform of $S_y(f)$:⁶

$$R_y(\tau) = \mathfrak{F}^{-1}[S_y(f)] = \int_{-\infty}^{\infty} |H(f)|^2 S_x(f) e^{j2\pi f\tau} df \quad (7.72)$$

$H(f)$ is the system's frequency response function; $S_x(f)$ is the power spectral density of the input $x(t)$; $S_y(f)$ is the power spectral density of the output $y(t)$; and $R_y(\tau)$ is the autocorrelation function of the output. The analogous result for energy signals was proved in Chapter 2 [Equation (2.190)], and the result for power signals was simply stated.

A proof of (7.71) could be carried out by employing (7.25). We will take a somewhat longer route, however, and obtain several useful intermediate results. In addition, the proof provides practice in manipulating convolutions and expectations.

We begin by obtaining the cross-correlation function between input and output, $R_{xy}(\tau)$, defined as

$$R_{xy}(\tau) = E[x(t)y(t + \tau)] \quad (7.73)$$

Using the superposition integral, we have

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t - u) du \quad (7.74)$$

where $h(t)$ is the system's impulse response. Equation (7.74) relates each sample function of the input and output processes, so we can write (7.73) as

$$R_{xy}(\tau) = E \left\{ x(t) \int_{-\infty}^{\infty} h(u)x(t + \tau - u) du \right\} \quad (7.75)$$

Since the integral does not depend on t , we can take $x(t)$ inside and interchange the operations of expectation and convolution. (Both are simply integrals over different variables.) Since $h(u)$ is not random, (7.75) becomes

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} h(u) E \{ x(t)x(t + \tau - u) \} du \quad (7.76)$$

By definition of the autocorrelation function of $x(t)$,

$$E[x(t)x(t + \tau - u)] = R_x(\tau - u) \quad (7.77)$$

⁶For the remainder of this chapter we use lowercase x and y to denote input and output random-process signals in keeping with Chapter 2 notation.

Thus, (7.76) can be written as

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} h(u)R_x(\tau - u) du \triangleq h(\tau) * R_x(\tau) \quad (7.78)$$

That is, the cross-correlation function of input with output is *the autocorrelation function of the input convolved with the system's impulse response*, an easily remembered result. Since (7.78) is a convolution, the Fourier transform of $R_{xy}(\tau)$, the cross-power spectral density of $x(t)$ with $y(t)$ is

$$S_{xy}(f) = H(f)S_x(f) \quad (7.79)$$

From the time-reversal theorem of Table F.6, the cross-power spectral density $S_{yx}(f)$ is

$$S_{yx}(f) = \mathfrak{F}[R_{yx}(\tau)] = \mathfrak{F}[R_{xy}(-\tau)] = S_{xy}^*(f) \quad (7.80)$$

Employing (7.79) and using the relationships $H^*(f) = H(-f)$ and $S_x^*(f) = S_x(f)$ [where $S_x(f)$ is real], we obtain

$$S_{yx}(f) = H(-f)S_x(f) = H^*(f)S_x(f) \quad (7.81)$$

where the order of the subscripts is important. Taking the inverse Fourier transform of (7.81) with the aid of the convolution theorem of Fourier transforms in Table F.6, and again using the time-reversal theorem, we obtain

$$R_{yx}(\tau) = h(-\tau) * R_x(\tau) \quad (7.82)$$

Let us pause to emphasize what we have obtained. By definition, $R_{xy}(\tau)$ can be written as

$$R_{xy}(\tau) \triangleq E\{x(t) \underbrace{[h(t) * x(t + \tau)]}_{y(t + \tau)}}\} \quad (7.83)$$

Combining this with (7.78), we have

$$E\{x(t)[h(t) * x(t + \tau)]\} = h(\tau) * R_x(\tau) \triangleq h(\tau) * E\{x(t)x(t + \tau)\} \quad (7.84)$$

Similarly, (7.82) becomes

$$\begin{aligned} R_{yx}(\tau) &\triangleq E\{\underbrace{[h(t) * x(t)]}_{y(t)}x(t + \tau)\} = h(-\tau) * R_x(\tau) \\ &\triangleq h(-\tau) * E\{x(t)x(t + \tau)\} \end{aligned} \quad (7.85)$$

Thus, bringing the convolution operation outside the expectation gives a convolution of $h(\tau)$ with the autocorrelation function if $h(t) * x(t + \tau)$ is inside the expectation, or a convolution of $h(-\tau)$ with the autocorrelation function if $h(t) * x(t)$ is inside the expectation.

These results are combined to obtain the autocorrelation function of the output of a linear system in terms of the input autocorrelation function as follows:

$$R_y(\tau) \triangleq E\{y(t)y(t + \tau)\} = E\{y(t)[h(t) * x(t + \tau)]\} \quad (7.86)$$

which follows because $y(t + \tau) = h(t) * x(t + \tau)$. Using (7.84) with $x(t)$ replaced by $y(t)$, we obtain

$$\begin{aligned} R_y(\tau) &= h(\tau) * E\{y(t)x(t + \tau)\} \\ &= h(\tau) * R_{yx}(\tau) \\ &= h(\tau) * \{h(-\tau) * R_x(\tau)\} \end{aligned} \quad (7.87)$$

where the last line follows by substituting from (7.82). Written in terms of integrals, (7.87) is

$$R_y(\tau) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(u) h(v) R_x(\tau + v - u) dv du \quad (7.88)$$

The Fourier transform of (7.87) is the output power spectral density and is easily obtained as follows:

$$\begin{aligned} S_y(f) &\triangleq \mathfrak{F}[R_y(\tau)] = \mathfrak{F}[h(\tau) * R_{yx}(\tau)] \\ &= H(f)S_{yx}(f) \\ &= |H(f)|^2 S_x(f) \end{aligned} \quad (7.89)$$

where (7.81) has been substituted to obtain the last line.

EXAMPLE 7.7

The input to a filter with impulse response $h(t)$ and frequency response function $H(f)$ is a white-noise process with power spectral density,

$$S_x(f) = \frac{1}{2} N_0, \quad -\infty < f < \infty \quad (7.90)$$

The cross-power spectral density between input and output is

$$S_{xy}(f) = \frac{1}{2} N_0 H(f) \quad (7.91)$$

and the cross-correlation function is

$$R_{xy}(\tau) = \frac{1}{2} N_0 h(\tau) \quad (7.92)$$

Hence, we could measure the impulse response of a filter by driving it with white noise and determining the cross-correlation function of input with output. Applications include system identification and channel measurement. ■

7.4.2 Filtered Gaussian Processes

Suppose the input to a linear system is a stationary random process. What can we say about the output statistics? For general inputs and systems, this is usually a difficult question to answer. However, *if the input to a linear system is Gaussian, the output is also Gaussian.*

A nonrigorous demonstration of this is carried out as follows. The sum of two independent Gaussian random variables has already been shown to be Gaussian. By repeated application of this result, we can find that the sum of any number of independent Gaussian random variables is Gaussian.⁷ For a fixed linear system, the output $y(t)$ in terms of the input $x(t)$ is given by

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} x(\tau) h(t - \tau) d\tau \\ &= \lim_{\Delta\tau \rightarrow 0} \sum_{k=-\infty}^{\infty} x(k \Delta\tau) h(t - k \Delta\tau) \Delta\tau \end{aligned} \quad (7.93)$$

⁷This also follows from Appendix B, (B.13).

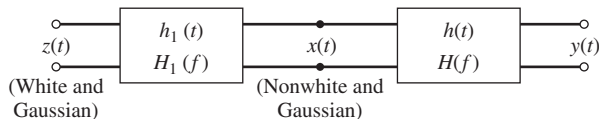


Figure 7.7
Cascade of two linear systems with Gaussian input.

where $h(t)$ is the impulse response. By writing the integral as a sum, we have demonstrated that if $x(t)$ is a white Gaussian process, the output is also Gaussian (but not white) because, at any time t , the right-hand side of (7.93) is simply a linear combination of independent Gaussian random variables. (Recall Example 7.5, where the autocorrelation function of white noise was shown to be an impulse. Also recall that uncorrelated Gaussian random variables are independent.)

If the input is not white, we can still show that the output is Gaussian by considering the cascade of two linear systems, as shown in Figure 7.7. The system in question is the one with the impulse response $h(t)$. To show that its output is Gaussian, we note that the cascade of $h_1(t)$ with $h(t)$ is a linear system with the impulse response

$$h_2(t) = h_1(t) * h(t) \tag{7.94}$$

This system's input, $z(t)$, is Gaussian and white. Therefore, its output, $y(t)$, is also Gaussian by application of the theorem just proved. However, the output of the system with impulse response $h_1(t)$ is Gaussian by application of the same theorem, but not white. Hence, the output of a linear system with nonwhite Gaussian input is Gaussian.

EXAMPLE 7.8

The input to the lowpass RC filter shown in Figure 7.8 is white Gaussian noise with the power spectral density $S_{n_i}(f) = \frac{1}{2}N_0, -\infty < f < \infty$. The power spectral density of the output is

$$S_{n_o}(f) = S_{n_i}(f) |H(f)|^2 = \frac{\frac{1}{2}N_0}{1 + (f/f_3)^2} \tag{7.95}$$

where $f_3 = (2\pi RC)^{-1}$ is the filter's 3-dB cutoff frequency. Inverse Fourier-transforming $S_{n_o}(f)$, we obtain $R_{n_o}(\tau)$, the output autocorrelation function, which is

$$R_{n_o}(\tau) = \frac{\pi f_3 N_0}{2} e^{-2\pi f_3 |\tau|} = \frac{N_0}{4RC} e^{-|\tau|/RC}, \quad \frac{1}{RC} = 2\pi f_3 \tag{7.96}$$

The square of the mean of $n_o(t)$ is

$$\overline{n_o(t)}^2 = \lim_{|\tau| \rightarrow \infty} R_{n_o}(\tau) = 0 \tag{7.97}$$

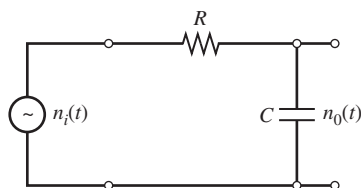


Figure 7.8
A lowpass RC filter with a white-noise input.

and the mean-squared value, which is also equal to the variance since the mean is zero, is

$$\overline{n_0^2(t)} = \sigma_{n_0}^2 = R_{n_0}(0) = \frac{N_0}{4RC} \quad (7.98)$$

Alternatively, we can find the average power at the filter output by integrating the power spectral density of $n_0(t)$. The same result is obtained as above:

$$\overline{n_0^2(t)} = \int_{-\infty}^{\infty} \frac{\frac{1}{2}N_0}{1 + (f/f_3)^2} df = \frac{N_0}{2\pi RC} \int_0^{\infty} \frac{dx}{1 + x^2} = \frac{N_0}{4RC} \quad (7.99)$$

Since the input is Gaussian, the output is Gaussian as well. The first-order pdf is

$$f_{n_0}(y, t) = f_{n_0}(y) = \frac{e^{-2RCy^2/N_0}}{\sqrt{\pi N_0/2RC}} \quad (7.100)$$

by employing Equation (6.194). The second-order pdf at time t and $t + \tau$ is found by substitution into Equation (6.189). Letting X be a random variable that refers to the values the output takes on at time t and Y be a random variable that refers to the values the output takes on at time $t + \tau$, we have, from the preceding results,

$$m_x = m_y = 0 \quad (7.101)$$

$$\sigma_x^2 = \sigma_y^2 = \frac{N_0}{4RC} \quad (7.102)$$

and the correlation coefficient is

$$\rho(\tau) = \frac{R_{n_0}(\tau)}{R_{n_0}(0)} = e^{-|\tau|/RC} \quad (7.103)$$

Referring to Example 7.2, one can see that the random telegraph waveform has the same autocorrelation function as that of the output of the lowpass RC filter of Example 7.8 (with constants appropriately chosen). This demonstrates that processes with drastically different sample functions can have the same second-order averages. ■

7.4.3 Noise-Equivalent Bandwidth

If we pass white noise through a filter that has the frequency response function $H(f)$, the average power at the output, by (7.72), is

$$P_{n_0} = \int_{-\infty}^{\infty} \frac{1}{2}N_0 |H(f)|^2 df = N_0 \int_0^{\infty} |H(f)|^2 df \quad (7.104)$$

where $\frac{1}{2}N_0$ is the two-sided power spectral density of the input. If the filter were ideal with bandwidth B_N and midband (maximum) gain⁸ H_0 , as shown in Figure 7.9, the noise power at the output would be

$$P_{n_0} = H_0^2 \left(\frac{1}{2}N_0 \right) (2B_N) = N_0 B_N H_0^2 \quad (7.105)$$

The question we now ask is the following: What is the bandwidth of an ideal, fictitious filter that has the same midband gain as $H(f)$ and that passes the same noise power? If the midband

⁸ Assumed to be finite.

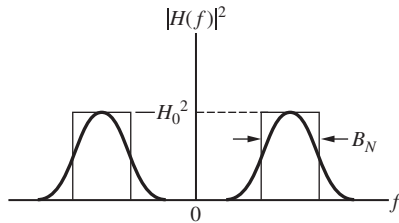


Figure 7.9
Comparison between $|H(f)|^2$ and an idealized approximation.

gain of $H(f)$ is H_0 , the answer is obtained by equating the preceding two results. Thus,

$$B_N = \frac{1}{H_0^2} \int_0^{\infty} |H(f)|^2 df \quad (7.106)$$

is the single-sided bandwidth of the fictitious filter. B_N is called the *noise-equivalent bandwidth* of $H(f)$.

It is sometimes useful to determine the noise-equivalent bandwidth of a system using time-domain integration. Assume a lowpass system with maximum gain at $f = 0$ for simplicity. By Rayleigh's energy theorem [see (2.88)], we have

$$\int_{-\infty}^{\infty} |H(f)|^2 df = \int_{-\infty}^{\infty} |h(t)|^2 dt \quad (7.107)$$

Thus, (7.106) can be written as

$$B_N = \frac{1}{2H_0^2} \int_{-\infty}^{\infty} |h(t)|^2 dt = \frac{\int_{-\infty}^{\infty} |h(t)|^2 dt}{2 \left[\int_{-\infty}^{\infty} h(t) dt \right]^2} \quad (7.108)$$

where it is noted that

$$H_0 = H(f)|_{f=0} = \int_{-\infty}^{\infty} h(t) e^{-j2\pi ft} \Big|_{f=0} = \int_{-\infty}^{\infty} h(t) dt \quad (7.109)$$

For some systems, (7.108) is easier to evaluate than (7.106).

EXAMPLE 7.9

Assume that a filter has the amplitude response function illustrated in Figure 7.10(a). Note that assumed filter is noncausal. The purpose of this problem is to provide an illustration of the computation of B_N for a simple filter. The first step is to square $|H(f)|$ to give $|H(f)|^2$ as shown in Figure 7.10(b). By simple geometry, the area under $|H(f)|^2$ for nonnegative frequencies is

$$A = \int_0^{\infty} |H(f)|^2 df = 50 \quad (7.110)$$

Note also that the maximum gain of the actual filter is $H_0 = 2$. For the ideal filter with amplitude response denoted by $H_e(f)$, which is ideal bandpass centered at 15 Hz of single-sided bandwidth B_N and passband gain H_0 , we want

$$\int_0^{\infty} |H(f)|^2 df = H_0^2 B_N \quad (7.111)$$

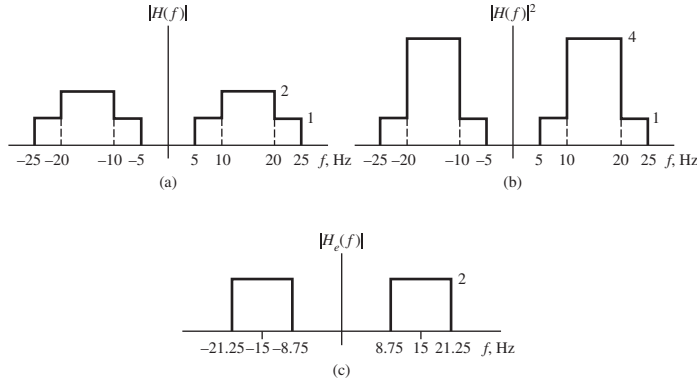


Figure 7.10
Illustrations for Example 7.9.

or

$$50 = 2^2 B_N \tag{7.112}$$

from which

$$B_N = 12.5 \text{ Hz} \tag{7.113}$$



EXAMPLE 7.10

The noise-equivalent bandwidth of an n th-order Butterworth filter for which

$$|H_n(f)|^2 = \frac{1}{1 + (f/f_3)^{2n}} \tag{7.114}$$

is

$$\begin{aligned} B_N(n) &= \int_0^\infty \frac{1}{1 + (f/f_3)^{2n}} df = f_3 \int_0^\infty \frac{1}{1 + x^{2n}} dx \\ &= \frac{\pi f_3 / 2n}{\sin(\pi/2n)}, \quad n = 1, 2, \dots \end{aligned} \tag{7.115}$$

where f_3 is the 3-dB frequency of the filter. For $n = 1$, (7.115) gives the result for a lowpass RC filter, namely $B_N(1) = \frac{\pi}{2} f_3$. As n approaches infinity, $H_n(f)$ approaches the frequency response function of an ideal lowpass filter of single-sided bandwidth f_3 . The noise-equivalent bandwidth is

$$\lim_{n \rightarrow \infty} B_N(n) = f_3 \tag{7.116}$$

as it should be by its definition. As the cutoff of a filter becomes sharper, its noise-equivalent bandwidth approaches its 3-dB bandwidth.



EXAMPLE 7.11

To illustrate the application of (7.108), consider the computation of the noise-equivalent bandwidth of a first-order Butterworth filter in the time domain. Its impulse response is

$$h(t) = \mathfrak{F}^{-1} \left[\frac{1}{1 + jf/f_3} \right] = 2\pi f_3 e^{-2\pi f_3 t} u(t) \quad (7.117)$$

According to (7.108), the noise-equivalent bandwidth of this filter is

$$B_N = \frac{\int_0^\infty (2\pi f_3)^2 e^{-4\pi f_3 t} dt}{2 \left[\int_0^\infty 2\pi f_3 e^{-2\pi f_3 t} dt \right]^2} = \frac{2\pi f_3}{2} \frac{\int_0^\infty e^{-v} dv}{2 \left(\int_0^\infty e^{-u} du \right)^2} = \frac{\pi f_3}{2} \quad (7.118)$$

which checks with (7.115) if $n = 1$ is substituted. ■

COMPUTER EXAMPLE 7.1

Equation (7.106) gives a fixed number for the noise-equivalent bandwidth. However, if the filter transfer function is unknown or cannot be easily integrated, it follows that the noise-equivalent bandwidth can be estimated by placing a finite-length segment of white noise on the input of the filter and measuring the input and output variances. The estimate of the noise-equivalent bandwidth is then the ratio of the output variance to the input variance. The following MATLAB program simulates the process. Note that unlike (7.106), the noise-equivalent bandwidth is now a random variable. The variance of the estimate can be reduced by increasing the length of the noise segment.

```
% File: c7ce1.m
clear all
npts = 50000;           % number of points generated
fs = 2000;             % sampling frequency
f3 = 20;               % 3-dB break frequency
N = 4;                % filter order
Wn = f3/(fs/2);       % scaled 3-dB frequency
in = randn(1,npts);   % vector of noise samples
[B,A] = butter(N,Wn); % filter parameters
out=filter(B,A,in);  % filtered noise samples
vin=var(in);          % variance of input noise samples
vout=var(out);        % input noise samples
Bnexp=(vout/vin)*(fs/2); % estimated noise-equivalent bandwidth
Bntheor=(pi*f3/2/N)/sin(pi/2/N); % true noise-equivalent bandwidth
a = ['The experimental estimate of Bn is ',num2str(Bnexp),' Hz.'];
b = ['The theoretical value of Bn is ',num2str(Bntheor),' Hz.'];
disp(a)
disp(b)
% End of script file.
```

Executing the program gives

```
>> c6cel
The experimental estimate of Bn is 20.5449 Hz.
The theoretical value of Bn is 20.5234 Hz.
```

■

7.5 NARROWBAND NOISE

7.5.1 Quadrature-Component and Envelope-Phase Representation

In most communication systems operating at a carrier frequency f_0 , the bandwidth of the channel, B , is small compared with f_0 . In such situations, it is convenient to represent the noise in terms of quadrature components as

$$n(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta) \quad (7.119)$$

where $\omega_0 = 2\pi f_0$ and θ is an arbitrary phase angle. In terms of envelope and phase components, $n(t)$ can be written as

$$n(t) = R(t) \cos[2\pi f_0 t + \phi(t) + \theta] \quad (7.120)$$

where

$$R(t) = \sqrt{n_c^2 + n_s^2} \quad (7.121)$$

and

$$\phi(t) = \tan^{-1} \left[\frac{n_s(t)}{n_c(t)} \right] \quad (7.122)$$

Actually, any random process can be represented in either of these forms, but if a process is narrowband, $R(t)$ and $\phi(t)$ can be interpreted as the slowly varying envelope and phase, respectively, as sketched in Figure 7.11.

Figure 7.12 shows the block diagram of a system for producing $n_c(t)$ and $n_s(t)$ where θ is, as yet, an arbitrary phase angle. Note that the composite operations used in producing $n_c(t)$ and $n_s(t)$ constitute linear systems (superposition holds from input to output). Thus, if $n(t)$ is a Gaussian process, so are $n_c(t)$ and $n_s(t)$. (The system of Figure 7.12 is to be interpreted as relating input and output processes sample function by sample function.)

We will prove several properties of $n_c(t)$ and $n_s(t)$. Most important, of course, is whether equality really holds in (7.119) and in what sense. It is shown in Appendix C that

$$E \left\{ [n(t) - [n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)]]^2 \right\} = 0 \quad (7.123)$$

That is, the mean-squared error between a sample function of the actual noise process and the right-hand side of (7.119) is zero (averaged over the ensemble of sample functions).

More useful when using the representation in (7.119), however, are the following properties:

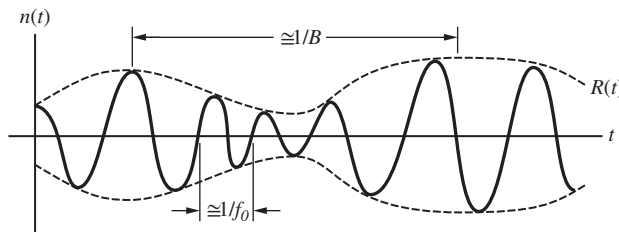


Figure 7.11
A typical narrowband noise waveform.

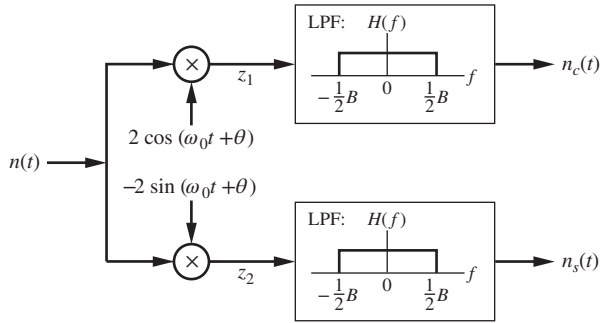


Figure 7.12
The operations involved in producing $n_c(t)$ and $n_s(t)$.

MEANS

$$\overline{n(t)} = \overline{n_c(t)} = \overline{n_s(t)} = 0 \tag{7.124}$$

VARIANCES

$$\overline{n^2(t)} = \overline{n_c^2(t)} = \overline{n_s^2(t)} \triangleq N \tag{7.125}$$

POWER SPECTRAL DENSITIES

$$S_{n_c}(f) = S_{n_s}(f) = \text{Lp} [S_n(f - f_0) + S_n(f + f_0)] \tag{7.126}$$

CROSS-POWER SPECTRAL DENSITY

$$S_{n_c n_s}(f) = j \text{Lp} [S_n(f - f_0) - S_n(f + f_0)] \tag{7.127}$$

where $\text{Lp} [\]$ denotes the lowpass part of the quantity in brackets; $S_n(f)$, $S_{n_c}(f)$, and $S_{n_s}(f)$ are the power spectral densities of $n(t)$, $n_c(t)$, and $n_s(t)$, respectively; $S_{n_c n_s}(f)$ is the cross-power spectral density of $n_c(t)$ and $n_s(t)$. From (7.127), we see that

$$R_{n_c n_s}(\tau) \equiv 0 \quad \text{for all } \tau, \text{ if } \text{Lp}[S_n(f - f_0) - S_n(f + f_0)] = 0 \tag{7.128}$$

This is an especially useful property in that it tells us that $n_c(t)$ and $n_s(t)$ are uncorrelated if the power spectral density of $n(t)$ is symmetrical about $f = f_0$ where $f > 0$. If, in addition, $n(t)$ is Gaussian, $n_c(t)$ and $n_s(t)$ will be independent Gaussian processes because they are uncorrelated, and the joint pdf of $n_c(t)$ and $n_s(t + \tau)$ for any delay τ , will simply be of the form

$$f(n_c, t; n_s, t + \tau) = \frac{1}{2\pi N} e^{-(n_c^2 + n_s^2)/2N} \tag{7.129}$$

If $S_n(f)$ is not symmetrical about $f = f_0$, where $f > 0$, then (7.129) holds only for $\tau = 0$ or those values of τ for which $R_{n_c n_s}(\tau) = 0$.

Using the results of Example 6.15, the envelope and phase functions of (7.120) have the joint pdf

$$f(r, \phi) = \frac{r}{2\pi N} e^{-r^2/2N}, \text{ for } r > 0 \text{ and } |\phi| \leq \pi \tag{7.130}$$

which holds for the same conditions as for (7.129).

7.5.2 The Power Spectral Density Function of $n_c(t)$ and $n_s(t)$

To prove (7.126), we first find the power spectral density of $z_1(t)$, as defined in Figure 7.12, by computing its autocorrelation function and Fourier-transforming the result. To simplify the derivation, it is assumed that θ is a uniformly distributed random variable in $[0, 2\pi)$ and is statistically independent of $n(t)$.⁹

The autocorrelation function of $z_1(t) = 2n(t) \cos(\omega_0 t + \theta)$ is

$$\begin{aligned} R_{z_1}(\tau) &= E\{4n(t)n(t+\tau) \cos(2\pi f_0 t + \theta) \cos[2\pi f_0(t+\tau) + \theta]\} \\ &= 2E[n(t)n(t+\tau)] \cos 2\pi f_0 \tau \\ &\quad + 2E[n(t)n(t+\tau) \cos(4\pi f_0 t + 2\pi f_0 \tau + 2\theta)] \\ &= 2R_n(\tau) \cos 2\pi f_0 \tau \end{aligned} \quad (7.131)$$

where $R_n(\tau)$ is the autocorrelation function of $n(t)$ and $\omega_0 = 2\pi f_0$ in Figure 6.12. In obtaining (7.131), we used appropriate trigonometric identities in addition to the independence of $n(t)$ and θ . Thus, by the multiplication theorem of Fourier transforms, the power spectral density of $z_1(t)$ is

$$\begin{aligned} S_{z_1}(f) &= S_n(f) * [\delta(f - f_0) + \delta(f + f_0)] \\ &= S_n(f - f_0) + S_n(f + f_0) \end{aligned} \quad (7.132)$$

of which only the lowpass part is passed by $H(f)$. Thus, the result for $S_{n_c}(f)$ expressed by (7.126) follows. A similar proof can be carried out for $S_{n_s}(f)$. Equation (7.125) follows by integrating (7.126) over all f .

Next, let us consider (7.127). To prove it, we need an expression for $R_{z_1 z_2}(\tau)$, the cross-correlation function of $z_1(t)$ and $z_2(t)$. (See Figure 7.12.) By definition, and from Figure 7.12,

$$\begin{aligned} R_{z_1 z_2}(\tau) &= E\{z_1(t)z_2(t+\tau)\} \\ &= E\{4n(t)n(t+\tau) \cos(2\pi f_0 t + \theta) \sin[2\pi f_0(t+\tau) + \theta]\} \\ &= 2R_n(\tau) \sin 2\pi f_0 \tau \end{aligned} \quad (7.133)$$

where we again used appropriate trigonometric identities and the independence of $n(t)$ and θ . Letting $h(t)$ be the impulse response of the lowpass filters in Figure 7.12 and employing (7.84) and (7.85), the cross-correlation function of $n_c(t)$ and $n_s(t)$ can be written as

$$\begin{aligned} R_{n_c n_s}(\tau) &\triangleq E[n_c(t)n_s(t+\tau)] = E\{[h(t) * z_1(t)]n_s(t+\tau)\} \\ &= h(-\tau) * E\{z_1(t)n_s(t+\tau)\} \\ &= h(-\tau) * E\{z_1(t)[h(t) * z_2(t+\tau)]\} \\ &= h(-\tau) * h(\tau) * E\{z_1(t)z_2(t+\tau)\} \\ &= h(-\tau) * [h(\tau) * R_{z_1 z_2}(\tau)] \end{aligned} \quad (7.134)$$

⁹This might be satisfactory for modeling noise where the phase can be viewed as completely random. In other situations, where knowledge of the phase makes this an inappropriate assumption, a cyclostationary model may be more appropriate.

The Fourier transform of $R_{n_c n_s}(\tau)$ is the cross-power spectral density, $S_{n_c n_s}(f)$, which, from the convolution theorem, is given by

$$\begin{aligned} S_{n_c n_s}(f) &= H(f) \mathfrak{F}[h(-\tau) * R_{z_1 z_2}(\tau)] \\ &= H(f) H^*(f) S_{z_1 z_2}(f) \\ &= |H(f)|^2 S_{z_1 z_2}(f) \end{aligned} \quad (7.135)$$

From (7.133) and the frequency translation theorem, it follows that

$$\begin{aligned} S_{z_1 z_2}(f) &= \mathfrak{F}[jR_n(\tau) (e^{j2\pi f_0 \tau} - e^{-j2\pi f_0 \tau})] \\ &= j [S_n(f - f_0) - S_n(f + f_0)] \end{aligned} \quad (7.136)$$

Thus, from (7.135),

$$\begin{aligned} S_{n_c n_s}(f) &= j |H(f)|^2 [S_n(f - f_0) - S_n(f + f_0)] \\ &= j \text{Lp} [S_n(f - f_0) - S_n(f + f_0)] \end{aligned} \quad (7.137)$$

which proves (7.127). Note that since the cross-power spectral density $S_{n_c n_s}(f)$ is imaginary, the cross-correlation function $R_{n_c n_s}(\tau)$ is odd. Thus, $R_{n_c n_s}(0)$ is zero if the cross-correlation function is continuous at $\tau = 0$, which is the case for bandlimited signals.

EXAMPLE 7.12

Let us consider a bandpass random process with the power spectral density shown in Figure 7.13(a). Choosing the center frequency of $f_0 = 7$ Hz results in $n_c(t)$ and $n_s(t)$ being uncorrelated. Figure 7.13(b) shows $S_{z_1}(f)$ [or $S_{z_2}(f)$] for $f_0 = 7$ Hz with $S_{n_c}(f)$ [or $S_{n_s}(f)$], that is, the lowpass part of $S_{z_1}(f)$, shaded. The integral of $S_n(f)$ is $2(6)(2) = 24$ W, which is the same result obtained from integrating the shaded portion of Figure 7.13(b).

Now suppose f_0 is chosen as 5 Hz. Then $S_{z_1}(f)$ and $S_{z_2}(f)$ are as shown in Figure 7.12(c), with $S_{n_c}(f)$ shown shaded. From Equation (7.127), it follows that $-jS_{n_c n_s}(f)$ is the shaded portion of Figure 7.12(d). Because of the asymmetry that results from the choice of f_0 , $n_c(t)$ and $n_s(t)$ are not uncorrelated. As a matter of interest, we can calculate $R_{n_c n_s}(\tau)$ easily by using the transform pair

$$2AW \text{sinc} 2W\tau \leftrightarrow A\Pi\left(\frac{f}{2W}\right) \quad (7.138)$$

and the frequency-translation theorem. From Figure 7.12(d), it follows that

$$S_{n_c n_s}(f) = 2j \left\{ -\Pi\left[\frac{1}{4}(f - 3)\right] + \Pi\left[\frac{1}{4}(f + 3)\right] \right\} \quad (7.139)$$

which results in the cross-correlation function

$$\begin{aligned} R_{n_c n_s}(\tau) &= 2j (-4\text{sinc} 4\tau e^{j6\pi\tau} + 4\text{sinc} 4\tau e^{-j6\pi\tau}) \\ &= 16 \text{sinc}(4\tau) \sin(6\pi\tau) \end{aligned} \quad (7.140)$$

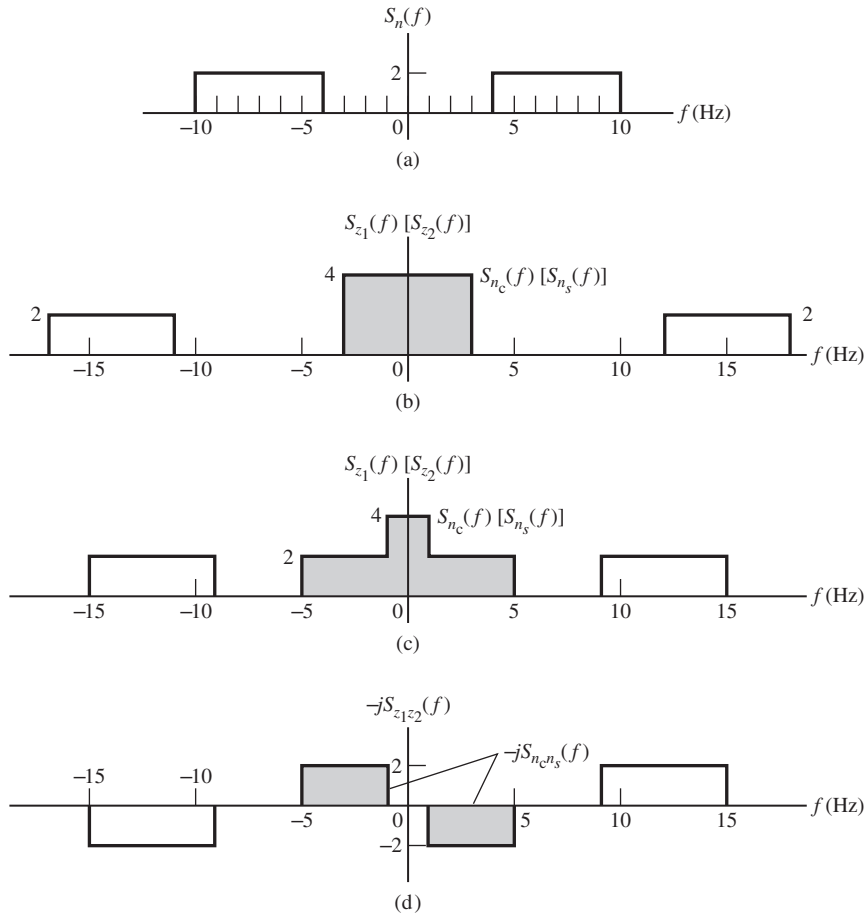


Figure 7.13
 Spectra for Example 7.11. (a) Bandpass spectrum. (b) Lowpass spectra for $f_0 = 7$ Hz. (c) Lowpass spectra for $f_0 = 5$ Hz. (d) Cross spectra for $f_0 = 5$ Hz.

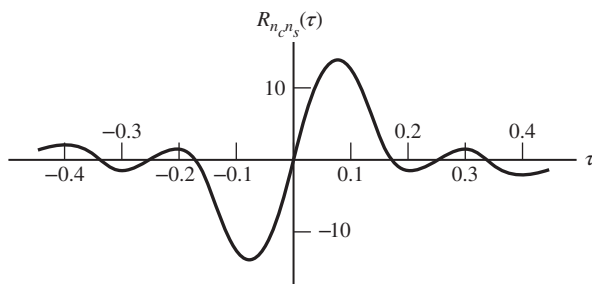


Figure 7.14
 Cross-correlation function of $n_c(t)$ and $n_s(t)$ for Example 7.11.

This cross-correlation function is shown in Figure 7.14. Although $n_c(t)$ and $n_s(t)$ are not uncorrelated, we see that τ may be chosen such that $R_{n_c n_s}(\tau) = 0$ for particular values of τ ($\tau = 0, \pm 1/6, \pm 1/3, \dots$). ■

7.5.3 Ricean Probability Density Function

A useful random-process model for many applications, for example, signal fading, is the sum of a random phased sinusoid and bandlimited Gaussian random noise. Thus, consider a sample function of this process expressed as

$$z(t) = A \cos(\omega_0 t + \theta) + n_c(t) \cos(\omega_0 t) - n_s(t) \sin(\omega_0 t) \quad (7.141)$$

where $n_c(t)$ and $n_s(t)$ are Gaussian quadrature components of the bandlimited, stationary, Gaussian random process $n_c(t) \cos(\omega_0 t) - n_s(t) \sin(\omega_0 t)$, A is a constant amplitude, and θ is a random variable uniformly distributed in $[0, 2\pi)$. The pdf of the envelope of this stationary random process at any time t is said to be Ricean after its originator, S. O. Rice. The first term is often referred to as the specular component and the latter two terms make up the diffuse component. This is in keeping with the idea that (7.141) results from transmitting an unmodulated sinusoidal signal through a dispersive channel, with the specular component being a direct-ray reception of that signal while the diffuse component is the resultant of multiple independent reflections of the transmitted signal (the central-limit theorem of probability can be invoked to justify that the quadrature components of this diffuse part are Gaussian random processes). Note that if $A = 0$, the pdf of the envelope of (7.141) is Rayleigh.

The derivation of the Ricean pdf proceeds by expanding the first term of (7.141) using the trigonometric identity for the cosine of the sum of two angles to rewrite it as

$$\begin{aligned} z(t) &= A \cos \theta \cos(2\pi f_0 t) - A \sin \theta \sin(2\pi f_0 t) + n_c(t) \cos(2\pi f_0 t) - n_s(t) \sin(2\pi f_0 t) \\ &= [A \cos \theta + n_c(t)] \cos(2\pi f_0 t) - [A \sin \theta + n_s(t)] \sin(2\pi f_0 t) \\ &= X(t) \cos(2\pi f_0 t) - Y(t) \sin(2\pi f_0 t) \end{aligned} \quad (7.142)$$

where

$$X(t) = A \cos \theta + n_c(t) \quad \text{and} \quad Y(t) = A \sin \theta + n_s(t) \quad (7.143)$$

These random processes, given θ , are independent Gaussian random processes with variance σ^2 . Their means are $E[X(t)] = A \cos \theta$ and $E[Y(t)] = A \sin \theta$, respectively. The goal is to find the pdf of

$$R(t) = \sqrt{X^2(t) + Y^2(t)} \quad (7.144)$$

Given θ , the joint pdf of $X(t)$ and $Y(t)$ is the product of their respective marginal pdfs since they are independent. Using the means and variance given above, this becomes

$$\begin{aligned} f_{XY}(x, y) &= \frac{\exp[-(x - A \cos \theta)^2 / 2\sigma^2]}{\sqrt{2\pi\sigma^2}} \frac{\exp[-(y - A \sin \theta)^2 / 2\sigma^2]}{\sqrt{2\pi\sigma^2}} \\ &= \frac{\exp\{-[x^2 + y^2 - 2A(\cos \theta + \sin \theta)x + A^2] / 2\sigma^2\}}{2\pi\sigma^2} \end{aligned} \quad (7.145)$$

Now make the change of variables

$$\left. \begin{aligned} x &= r \cos \phi \\ y &= r \sin \phi \end{aligned} \right\}, \quad r \geq 0 \quad \text{and} \quad 0 \leq \phi < 2\pi \quad (7.146)$$

Recall that transformation of a joint pdf requires multiplication by the Jacobian of the transformation, which in this case is just r . Thus, the joint pdf of the random variables R and Φ is

$$\begin{aligned} f_{R\Phi}(r, \phi) &= \frac{r \exp \left\{ - \left[r^2 + A^2 - 2rA (\cos \theta \cos \phi + \sin \theta \sin \phi) \right] / 2\sigma^2 \right\}}{2\pi\sigma^2} \\ &= \frac{r}{2\pi\sigma^2} \exp \left\{ - \left[r^2 + A^2 - 2rA \cos(\theta - \phi) \right] / 2\sigma^2 \right\} \end{aligned} \quad (7.147)$$

The pdf over R alone may be obtained by integrating over ϕ with the aid of the definition

$$I_0(u) = \frac{1}{2\pi} \int_0^{2\pi} \exp(u \cos \alpha) d\alpha \quad (7.148)$$

where $I_0(u)$ is referred to as the modified Bessel function of order zero. Since the integrand of (7.148) is periodic with period 2π , the integral can be over any 2π range. The result of the integration of (7.147) over ϕ produces

$$f_R(r) = \frac{r}{\sigma^2} \exp \left\{ - \left[r^2 + A^2 \right] / 2\sigma^2 \right\} I_0 \left(\frac{Ar}{\sigma^2} \right), \quad r \geq 0 \quad (7.149)$$

Since the result is independent of θ , this is the marginal pdf of R alone. From (7.148), it follows that $I_0(0) = 1$ so that with $A = 0$ (7.149) reduces to the Rayleigh pdf, as it should.

Often, (7.149) is expressed in terms of the parameter $K = \frac{A^2}{2\sigma^2}$, which is the ratio of the powers in the steady component [first term of (7.141)] to the random Gaussian component [second and third terms of (7.141)]. When this is done, (7.149) becomes

$$f_R(r) = \frac{r}{\sigma^2} \exp \left\{ - \left[\frac{r^2}{2\sigma^2} + K \right] \right\} I_0 \left(\sqrt{2K} \frac{r}{\sigma} \right), \quad r \geq 0 \quad (7.150)$$

As K becomes large, (7.150) approaches a Gaussian pdf. The parameter K is often referred to as the Ricean K -factor.

From (7.144) it follows that

$$\begin{aligned} E[R^2] &= E[X^2] + E[Y^2] \\ &= E \left\{ \left[A \cos \theta + n_c(t) \right]^2 + \left[A \sin \theta + n_s(t) \right]^2 \right\} \\ &= E \left[A^2 \cos^2 \theta + A^2 \sin^2 \theta \right] + 2AE \left[n_c(t) \cos \theta + n_s(t) \sin \theta \right] + E \left[n_c^2(t) \right] \\ &\quad + E \left[n_s^2(t) \right] \\ &= A^2 + 2\sigma^2 \\ &= 2\sigma^2 (1 + K) \end{aligned} \quad (7.151)$$

Other moments for a Ricean random variable must be expressed in terms of confluent hypergeometric functions.¹⁰

¹⁰See, for example, J. Proakis, *Digital Communications*, 4th ed., New York: McGraw Hill, 2001.

Further Reading

Papoulis (1991) is a recommended book for random processes. The references given in Chapter 6 also provide further reading on the subject matter of this chapter.

Summary

1. A random process is completely described by the N -fold joint pdf of its amplitudes at the arbitrary times t_1, t_2, \dots, t_N . If this pdf is invariant under a shift of the time origin, the process is said to be *statistically stationary in the strict sense*.

2. The autocorrelation function of a random process, computed as a statistical average, is defined as

$$R(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 x_2 f_{X_1 X_2}(x_1, t_1; x_2, t_2) dx_1 dx_2$$

where $f_{X_1 X_2}(x_1, t_1; x_2, t_2)$ is the joint amplitude pdf of the process at times t_1 and t_2 . If the process is stationary,

$$R(t_1, t_2) = R(t_2 - t_1) = R(\tau)$$

where $\tau \triangleq t_2 - t_1$.

3. A process whose statistical average mean and variance are time-independent and whose autocorrelation function is a function only of $t_2 - t_1 = \tau$ is termed *wide-sense stationary*. *Strict-sense stationary processes are also wide-sense stationary. The converse is true only for special cases; for example, wide-sense stationarity for a Gaussian process guarantees strict-sense stationarity.*

4. A process for which statistical averages and time averages are equal is called *ergodic*. *Ergodicity implies stationarity, but the reverse is not necessarily true.*

5. The Wiener–Khinchine theorem states that the autocorrelation function and the power spectral density of a stationary random process are a Fourier-transform pair. An expression for the power spectral density of a random process that is often useful is

$$S_n(f) = \lim_{T \rightarrow \infty} \frac{1}{T} E \left\{ \left| \mathfrak{F} [n_T(t)] \right|^2 \right\}$$

where $n_T(t)$ is a sample function truncated to T seconds, centered about $t = 0$.

6. The autocorrelation function of a random process is a real, even function of the delay variable τ with an absolute maximum at $\tau = 0$. It is periodic for periodic random processes, and its Fourier transform is nonnegative for all frequencies. As $\tau \rightarrow \pm\infty$, the autocorrelation function

approaches the square of the mean of the random process unless the random process is periodic. $R(0)$ gives the total average power in a process.

7. White noise has a constant power spectral density $\frac{1}{2}N_0$ for all f . Its autocorrelation function is $\frac{1}{2}N_0\delta(\tau)$. For this reason, it is sometimes called delta-correlated noise. It has infinite power and is therefore a mathematical idealization. However, it is, nevertheless, a useful approximation in many cases.

8. The cross-correlation function of two stationary random processes $X(t)$ and $Y(t)$ is defined as

$$R_{XY}(\tau) = E[X(t)Y(t + \tau)]$$

Their cross-power spectral density is

$$S_{XY}(f) = \mathfrak{F}[R_{XY}(\tau)]$$

They are said to be *orthogonal* if $R_{XY}(\tau) = 0$ for all τ .

9. Consider a linear system with the impulse response $h(t)$ and the frequency response function $H(f)$ with random input $x(t)$ and output $y(t)$. Then

$$S_Y(f) = |H(f)|^2 S_X(f)$$

$$R_Y(\tau) = \mathfrak{F}^{-1} [S_Y(f)] = \int_{-\infty}^{\infty} |H(f)|^2 S_X(f) e^{j2\pi f\tau} df$$

$$R_{XY}(\tau) = h(\tau) * R_X(\tau)$$

$$S_{XY}(f) = H(f)S_X(f)$$

$$R_{YX}(\tau) = h(-\tau) * R_X(\tau)$$

$$S_{YX}(f) = H^*(f)S_X(f)$$

where $S(f)$ denotes the spectral density, $R(\tau)$ denotes the autocorrelation function, and the asterisk denotes convolution.

10. The output of a linear system with Gaussian input is Gaussian.

11. The noise-equivalent bandwidth of a linear system with a frequency response function $H(f)$ is defined as

$$B_N = \frac{1}{H_0^2} \int_0^{\infty} |H(f)|^2 df$$

where H_0 represents the maximum value of $|H(f)|$. If the input is white noise with the single-sided power spectral density N_0 , the output power is

$$P_0 = H_0^2 N_0 B_N$$

An equivalent expression for the noise-equivalent bandwidth written in terms of the impulse response of the filter is

$$B_N = \frac{\int_{-\infty}^{\infty} |h(t)|^2 dt}{2 \left[\int_{-\infty}^{\infty} h(t) dt \right]^2}$$

12. The quadrature-component representation of a bandlimited random process $n(t)$ is

$$n(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)$$

where θ is an arbitrary phase angle. The envelope-phase representation is

$$n(t) = R(t) \cos(2\pi f_0 t + \phi(t) + \theta)$$

where $R^2(t) = n_c^2(t) + n_s^2(t)$ and $\tan[\phi(t)] = n_s(t)/n_c(t)$. If the process is narrowband, n_c , n_s , R , and ϕ vary slowly with respect to $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$. If the power spectral density of $n(t)$ is $S_n(f)$, the power spectral densities

of $n_c(t)$ and $n_s(t)$ are

$$S_{n_c}(f) = S_{n_s}(f) = \text{Lp}[S_n(f - f_0) + S_n(f + f_0)]$$

where $\text{Lp}[]$ denotes the low-frequency part of the quantity in the brackets. If $\text{Lp}[S_n(f + f_0) - S_n(f - f_0)] = 0$, then $n_c(t)$ and $n_s(t)$ are orthogonal. The average powers of $n_c(t)$, $n_s(t)$, and $n(t)$ are equal. The processes $n_c(t)$ and $n_s(t)$ are given by

$$n_c(t) = \text{Lp}[2n(t) \cos(2\pi f_0 t + \theta)]$$

and

$$n_s(t) = -\text{Lp}[2n(t) \sin(2\pi f_0 t + \theta)]$$

Since these operations are linear, $n_c(t)$ and $n_s(t)$ will be Gaussian if $n(t)$ is Gaussian. Thus, $n_c(t)$ and $n_s(t)$ are independent if $n(t)$ is zero-mean Gaussian with a power spectral density that is symmetrical about $f = f_0$ for $f > 0$.

13. The Ricean pdf gives the distribution of envelope values assumed by the sum of a sinusoid with phase uniformly distributed in $[0, 2\pi)$ plus bandlimited Gaussian noise. It is convenient in various applications including modeling of fading channels.

Drill Problems

7.1 A random process is defined by the sample functions $X_i(t) = A_i t + B_i$, where t is time in seconds, the A_i s are independent random variables for each i , which are Gaussian with 0 means and unit variances, and the B_i s are independent random variables for each i uniformly distributed in $[-0.5, 0.5]$.

- Sketch several typical sample functions.
- Is the random process stationary?
- Is the random process ergodic?
- Write down an expression for its mean at an arbitrary time t .
- Write down an expression for its mean-squared value at an arbitrary time t .
- Write down an expression for its variance at an arbitrary time t .

7.2 White Gaussian noise of double-sided power spectral density 1 W/Hz is passed through a filter with frequency response function $H(f) = (1 + j2\pi f)^{-1}$.

- What is the power spectral density, $S_Y(f)$, of the output process?;

- What is the autocorrelation function, $R_Y(\tau)$, of the output process?
- What is the mean of the output process?
- What is the variance of the output process?
- Is the output process stationary?
- What is the first-order pdf of the output process?
- Comment on the similarities and dissimilarities of the output process and the random process considered in Example 7.2.

7.3 For each case given below, tell whether the given function can be a satisfactory autocorrelation function. If it is not satisfactory, give the reason(s).

- $R_a(\tau) = \Pi(\tau/\tau_0)$ where τ_0 is a constant;
- $R_b(\tau) = \Lambda(\tau/\tau_0)$ where τ_0 is a constant;
- $R_c(\tau) = A \cos(2\pi f_0 \tau)$ where A and f_0 are constants;
- $R_d(\tau) = A + B \cos(2\pi f_0 \tau)$ where A , B , and f_0 are constants;
- $R_e(\tau) = A \sin(2\pi f_0 \tau)$ where A and f_0 are constants;

342 Chapter 7 • Random Signals and Noise

(f) $R_f(\tau) = A \sin^2(2\pi f_0 \tau)$ where A and f_0 are constants.

7.4 A filter with frequency response function $H(f) = (1 + j2\pi f)^{-1}$ is driven by a white-noise process with double-sided power spectral density of 1 W/Hz.

- (a) What is the cross-power spectral density of input with output?
- (b) What is the cross-correlation function of input with output?
- (c) What is the power spectral density of the output?
- (d) What is the autocorrelation function of the output?

7.5 A bandpass random process has power spectral density $S(f) = \Pi\left(\frac{f-10}{4}\right) + \Pi\left(\frac{f+10}{4}\right)$.

- (a) Find its autocorrelation function.
- (b) It is to be represented in inphase-quadrature form; that is, $x(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)$. If f_0 is chosen as 10 Hz, what is the cross-spectral density, $S_{n_c n_s}(f)$?

(c) If f_0 is chosen as 8 Hz, what is the cross-spectral density, $S_{n_c n_s}(f)$?

(d) If f_0 is chosen as 12 Hz, what is the cross-spectral density, $S_{n_c n_s}(f)$?

(e) What is the cross-correlation function corresponding to part (c)?

(f) What is the cross-correlation function corresponding to part (d)?

7.6 A filter has frequency response function $H(f) = \Lambda(f/2)$. What is its noise-equivalent bandwidth?

7.7 A bandlimited signal consists of a steady sinusoidal component of power 10 W and a narrowband Gaussian component centered on the steady component of power 5 W. Find the following:

- (a) The steady to random power ratio, K .
- (b) The total received power.
- (c) The pdf of the envelope process.
- (d) The probability that the envelope will exceed 10 V (requires numerical integration).

Problems

Section 7.1

7.1 A fair die is thrown. Depending on the number of spots on the up face, the following random processes are generated. Sketch several examples of sample functions for each case.

$$(a) X(t, \zeta) = \begin{cases} 2A, & 1 \text{ or } 2 \text{ spots up} \\ 0, & 3 \text{ or } 4 \text{ spots up} \\ -2A, & 5 \text{ or } 6 \text{ spots up} \end{cases}$$

$$(b) X(t, \zeta) = \begin{cases} 3A, & 1 \text{ spot up} \\ 2A, & 2 \text{ spots up} \\ A, & 3 \text{ spots up} \\ -A, & 4 \text{ spots up} \\ -2A, & 5 \text{ spots up} \\ -3A, & 6 \text{ spots up} \end{cases}$$

$$(c) X(t, \zeta) = \begin{cases} 4A, & 1 \text{ spot up} \\ 2A, & 2 \text{ spots up} \\ At, & 3 \text{ spots up} \\ -At, & 4 \text{ spots up} \\ -2A, & 5 \text{ spots up} \\ -4A, & 6 \text{ spots up} \end{cases}$$

Section 7.2

7.2 Referring to Problem 7.1, what are the following probabilities for each case?

- (a) $F_X(X \leq 2A, t = 4)$
- (b) $F_X(X \leq 0, t = 4)$
- (c) $F_X(X \leq 2A, t = 2)$

7.3 A random process is composed of sample functions that are square waves, each with constant amplitude A , period T_0 , and random delay τ as sketched in Figure 7.15. The pdf of τ is

$$f(\tau) = \begin{cases} 1/T_0, & |\tau| \leq T_0/2 \\ 0, & \text{otherwise} \end{cases}$$

- (a) Sketch several typical sample functions.
- (b) Write the first-order pdf for this random process at some arbitrary time t_0 . (*Hint*: Because of the random delay τ , the pdf is independent of t_0 . Also, it might be easier to deduce the cdf and differentiate it to get the pdf.)

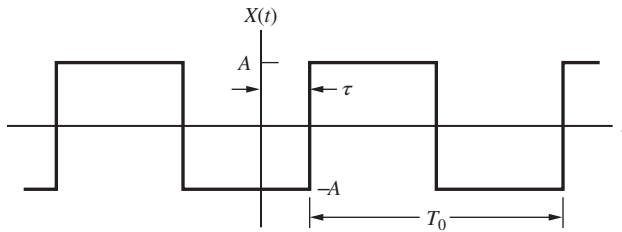


Figure 7.15

7.4 Let the sample functions of a random process be given by

$$X(t) = A \cos 2\pi f_0 t$$

where ω_0 is fixed and A has the pdf

$$f_A(a) = \frac{e^{-a^2/2\sigma_a^2}}{\sqrt{2\pi}\sigma_a}$$

This random process is passed through an ideal integrator to give a random process $Y(t)$.

- Find an expression for the sample functions of the output process $Y(t)$.
- Write down an expression for the pdf of $Y(t)$ at time t_0 . *Hint:* Note that $\sin 2\pi f_0 t_0$ is just a constant.
- Is $Y(t)$ stationary? Is it ergodic?

7.5 Consider the random process of Problem 7.3.

- Find the time-average mean and the autocorrelation function.
- Find the ensemble-average mean and the autocorrelation function.
- Is this process wide-sense stationary? Why or why not?

7.6 Consider the random process of Example 7.1 with the pdf of θ given by

$$p(\theta) = \begin{cases} 2/\pi, & \pi/2 \leq \theta \leq \pi \\ 0, & \text{otherwise} \end{cases}$$

- Find the statistical-average and time-average mean and variance.
- Find the statistical-average and time-average autocorrelation functions.
- Is this process ergodic?

7.7 Consider the random process of Problem 7.4.

- Find the time-average mean and the autocorrelation function.
- Find the ensemble-average mean and the autocorrelation function.
- Is this process wide-sense stationary? Why or why not?

7.8 The voltage of the output of a noise generator whose statistics are known to be closely Gaussian and stationary is measured with a dc voltmeter and a true root-mean-square (rms) voltmeter that is ac coupled. The dc meter reads 6 V, and the true rms meter reads 7 V. Write down an expression for the first-order pdf of the voltage at any time $t = t_0$. Sketch and dimension the pdf.

Section 7.3

7.9 Which of the following functions are suitable autocorrelation functions? Tell why or why not. (ω_0 , τ_0 , τ_1 , A , B , C , and f_0 are positive constants.)

- $A \cos \omega_0 \tau$
- $A \Lambda(\tau/\tau_0)$, where $\Lambda(x)$ is the unit-area triangular function defined in Chapter 2
- $A \Pi(\tau/\tau_0)$, where $\Pi(x)$ is the unit-area pulse function defined in Chapter 2
- $A \exp(-\tau/\tau_0) u(\tau)$ where $u(x)$ is the unit-step function
- $A \exp(-|\tau|/\tau_0)$
- $A \text{sinc}(f_0 \tau) = \frac{\sin(\pi f_0 \tau)}{\pi f_0 \tau}$

7.10 A bandlimited white-noise process has a double-sided power spectral density of 2×10^{-5} W/Hz in the frequency range $|f| \leq 1$ kHz. Find the autocorrelation function of the noise process. Sketch and fully dimension the resulting autocorrelation function.

344 Chapter 7 • Random Signals and Noise

7.11 Consider a random binary pulse waveform as analyzed in Example 7.6, but with half-cosine pulses given by $p(t) = \cos(2\pi t/2T)\Pi(t/T)$. Obtain and sketch the autocorrelation function for the two cases considered in Example 7.6, namely,

- (a) $a_k = \pm A$ for all k , where A is a constant, with $R_m = A^2$, $m = 0$, and $R_m = 0$ otherwise.
- (b) $a_k = A_k + A_{k-1}$ with $A_k = \pm A$ and $E[A_k A_{k+m}] = A^2$, $m = 0$, and zero otherwise.
- (c) Find and sketch the power spectral density for each preceding case.

7.12 Two random processes are given by

$$X(t) = n(t) + A \cos(2\pi f_0 t + \theta)$$

and

$$Y(t) = n(t) + A \sin(2\pi f_0 t + \theta)$$

where A and f_0 are constants and θ is a random variable uniformly distributed in the interval $[-\pi, \pi]$. The first term, $n(t)$, represents a stationary random noise process with autocorrelation function $R_n(\tau) = B\Lambda(\tau/\tau_0)$, where B and τ_0 are nonnegative constants.

- (a) Find and sketch their autocorrelation functions. Assume values for the various constants involved.
- (b) Find and sketch the cross-correlation function of these two random processes.

7.13 Given two independent, wide-sense stationary random processes $X(t)$ and $Y(t)$ with autocorrelation functions $R_X(\tau)$ and $R_Y(\tau)$, respectively.

- (a) Show that the autocorrelation function $R_Z(\tau)$ of their product $Z(t) = X(t)Y(t)$ is given by

$$R_Z(\tau) = R_X(\tau)R_Y(\tau)$$

- (b) Express the power spectral density of $Z(t)$ in terms of the power spectral densities of $X(t)$ and $Y(t)$, denoted as $S_X(f)$ and $S_Y(f)$, respectively.
- (c) Let $X(t)$ be a bandlimited stationary noise process with power spectral density $S_X(f) = 10\Pi(f/200)$, and let $Y(t)$ be the process defined by sample functions of the form

$$Y(t) = 5 \cos(50\pi t + \theta)$$

where θ is a uniformly distributed random variable in the interval $(0, 2\pi)$. Using the results derived in parts (a) and (b), obtain the autocorrelation function and power spectral density of $Z(t) = X(t)Y(t)$.

7.14 A random signal has the autocorrelation function

$$R(\tau) = 9 + 3\Lambda(\tau/5)$$

where $\Lambda(x)$ is the unit-area triangular function defined in Chapter 2. Determine the following:

- (a) The ac power.
- (b) The dc power.
- (c) The total power.
- (d) The power spectral density. Sketch it and label carefully.

7.15 A random process is defined as $Y(t) = X(t) + X(t - T)$, where $X(t)$ is a wide-sense stationary random process with autocorrelation function $R_X(T)$ and power spectral density $S_X(f)$.

- (a) Show that $R_Y(\tau) = 2R_X(\tau) + R_X(\tau + T) + R_X(\tau - T)$.
- (b) Show that $S_Y(f) = 4S_X(f) \cos^2(\pi f T)$.
- (c) If $X(t)$ has autocorrelation function $R_X(\tau) = 5\Lambda(\tau)$, where $\Lambda(\tau)$ is the unit-area triangular function, and $T = 0.5$, find and sketch the power spectral density of $Y(t)$ as defined in the problem statement.

7.16 The power spectral density of a wide-sense stationary random process is given by

$$S_X(f) = 10\delta(f) + 25\text{sinc}^2(5f) + 5\delta(f - 10) + 5\delta(f + 10)$$

- (a) Sketch and fully dimension this power spectral density function.
- (b) Find the power in the dc component of the random process.
- (c) Find the total power.
- (d) Given that the area under the main lobe of the sinc-squared function is approximately 0.9 of the total area, which is unity if it has unity amplitude, find the fraction of the total power contained in this process for frequencies between 0 and 0.2 Hz.

7.17 Given the following functions of τ :

$$R_{X_1}(\tau) = 4 \exp(-\alpha|\tau|) \cos 2\pi\tau$$

$$R_{X_2}(\tau) = 2 \exp(-\alpha|\tau|) + 4 \cos 2\pi b\tau$$

$$R_{X_3}(f) = 5 \exp(-4\tau^2)$$

- (a) Sketch each function and fully dimension.
- (b) Find the Fourier transforms of each and sketch. With the information of part (a) and the Fourier transforms justify that each is suitable for an autocorrelation function.
- (c) Determine the value of the dc power, if any, for each one.
- (d) Determine the total power for each.
- (e) Determine the frequency of the periodic component, if any, for each.

Section 7.4

7.18 A stationary random process $n(t)$ has a power spectral density of 10^{-6} W/Hz, $-\infty < f < \infty$. It is passed through an ideal lowpass filter with frequency response function $H(f) = \Pi(f/500 \text{ kHz})$, where $\Pi(x)$ is the unit-area pulse function defined in Chapter 2.

- (a) Find and sketch the power spectral density of the output?
- (b) Obtain and sketch the autocorrelation function of the output.
- (c) What is the power of the output process? Find it two different ways.

7.19 An ideal finite-time integrator is characterized by the input-output relationship

$$Y(t) = \frac{1}{T} \int_{t-T}^t X(\alpha) d\alpha$$

- (a) Justify that its impulse response is $h(t) = \frac{1}{T} [u(t) - u(t-T)]$.
- (b) Obtain its frequency response function. Sketch it.
- (c) The input is white noise with two-sided power spectral density $N_0/2$. Find the power spectral density of the output of the filter.
- (d) Show that the autocorrelation function of the output is

$$R_0(\tau) = \frac{N_0}{2T} \Lambda(\tau/T)$$

where $\Lambda(x)$ is the unit-area triangular function defined in Chapter 2.

- (e) What is the equivalent noise bandwidth of the integrator?
- (f) Show that the result for the output noise power obtained using the equivalent noise bandwidth found in part (e) coincides with the result found

from the autocorrelation function of the output found in part (d).

7.20 White noise with two-sided power spectral density $N_0/2$ drives a second-order Butterworth filter with frequency response function magnitude

$$|H_{2\text{bu}}(f)| = \frac{1}{\sqrt{1 + (f/f_3)^4}}$$

where f_3 is its 3-dB cutoff frequency.

- (a) What is the power spectral density of the filter's output?
- (b) Show that the autocorrelation function of the output is

$$R_0(r) = \frac{\pi f_3 N_0}{2} \exp\left(-\sqrt{2}\pi f_3 |\tau|\right) \cos\left[\sqrt{2}\pi f_3 |\tau| - \pi/4\right]$$

Plot as a function of $f_3 \tau$. *Hint:* Use the integral given below:

$$\int_0^\infty \frac{\cos(ax)}{b^4 + x^4} dx = \frac{\sqrt{2}\pi}{4b^3} \exp\left(-ab/\sqrt{2}\right) \left[\cos\left(ab/\sqrt{2}\right) + \sin\left(ab/\sqrt{2}\right)\right], \quad a, b > 0$$

- (c) Does the output power obtained by taking $\lim_{\tau \rightarrow 0} R_0(\tau)$ check with that calculated using the equivalent noise bandwidth for a Butterworth filter as given by (7.115)?

7.21 A power spectral density given by

$$S_Y(f) = \frac{f^2}{f^4 + 100}$$

is desired. A white-noise source of two-sided power spectral density 1 W/Hz is available. What is the frequency response function of the filter to be placed at the noise-source output to produce the desired power spectral density?

7.22 Obtain the autocorrelation functions and power spectral densities of the outputs of the following systems with the input autocorrelation functions or power spectral densities given.

- (a) Transfer function:

$$H(f) = \Pi(f/2B)$$

Autocorrelation function of input:

$$R_X(\tau) = \frac{N_0}{2} \delta(\tau)$$

N_0 and B are positive constants.

(b) Impulse response:

$$h(t) = A \exp(-\alpha t) u(t)$$

Power spectral density of input:

$$S_X(f) = \frac{B}{1 + (2\pi\beta f)^2}$$

$A, \alpha, B,$ and β are positive constants.

7.23 The input to a lowpass filter with impulse response

$$h(t) = \exp(-10t) u(t)$$

is white, Gaussian noise with single-sided power spectral density of 2 W/Hz. Obtain the following:

- (a) The mean of the output.
- (b) The power spectral density of the output.
- (c) The autocorrelation function of the output.
- (d) The probability density function of the output at an arbitrary time t_1 .
- (e) The joint probability density function of the output at times t_1 and $t_1 + 0.03$ s.

7.24 Find the noise-equivalent bandwidths for the following first- and second-order lowpass filters in terms of their 3-dB bandwidths. Refer to Chapter 2 to determine the magnitudes of their transfer functions.

- (a) Chebyshev
- (b) Bessel

7.25 A second-order Butterworth filter has 3-dB bandwidth of 500 Hz. Determine the unit impulse response of the filter and use it to compute the noise-equivalent bandwidth of the filter. Check your result against the appropriate special case of Example 7.9.

7.26 Determine the noise-equivalent bandwidths for the four filters having transfer functions given below:

- (a) $H_a(f) = \Pi(f/4) + \Pi(f/2)$
- (b) $H_b(f) = 2\Lambda(f/50)$
- (c) $H_c(f) = \frac{10}{10 + j2\pi f}$
- (d) $H_d(f) = \Pi(f/10) + \Lambda(f/5)$

7.27 A filter has frequency response function

$$H(f) = H_0(f - 500) + H_0(f + 500)$$

where

$$H_0(f) = 2\Lambda(f/100)$$

Find the noise-equivalent bandwidth of the filter.

7.28 Determine the noise-equivalent bandwidths of the systems having the following transfer functions. *Hint:* Use the time-domain approach.

(a) $H_a(f) = \frac{10}{(j2\pi f + 2)(j2\pi f + 25)}$

(b) $H_b(f) = \frac{100}{(j2\pi f + 10)^2}$

Section 7.5

7.29 Noise $n(t)$ has the power spectral density shown in Figure 7.16. We write

$$n(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)$$

Make plots of the power spectral densities of $n_c(t)$ and $n_s(t)$ for the following cases:

- (a) $f_0 = f_1$
- (b) $f_0 = f_2$
- (c) $f_0 = \frac{1}{2}(f_2 + f_1)$
- (d) For which of these cases are $n_c(t)$ and $n_s(t)$ uncorrelated?

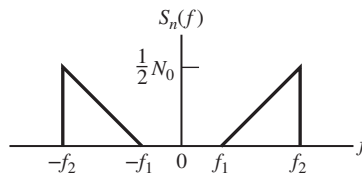


Figure 7.16

7.30

- (a) If $S_n(f) = \alpha^2 / (\alpha^2 + 4\pi^2 f^2)$ show that $R_n(\tau) = K e^{-\alpha|\tau|}$. Find K .
- (b) Find $R_n(\tau)$ if

$$S_n(f) = \frac{\frac{1}{2}\alpha^2}{\alpha^2 + 4\pi^2 (f - f_0)^2} + \frac{\frac{1}{2}\alpha^2}{\alpha^2 + 4\pi^2 (f + f_0)^2}$$

- (c) If $n(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)$, find $S_{n_c}(f)$, and $S_{n_c n_s}(f)$, where $S_n(f)$ is as given in part (b). Sketch each spectral density.

7.31 The double-sided power spectral density of noise $n(t)$ is shown in Figure 7.17. If $n(t) = n_c(t) \cos(2\pi f_0 t + \theta) - n_s(t) \sin(2\pi f_0 t + \theta)$, find and plot $S_{n_c}(f)$, $S_{n_s}(f)$, and $S_{n_c n_s}(f)$ for the following cases:

- (a) $f_0 = \frac{1}{2}(f_1 + f_2)$
- (b) $f_0 = f_1$
- (c) $f_0 = f_2$

- (d) Find $R_{n_c n_s}(\tau)$ for each case for where $S_{n_c n_s}(f)$ is not zero. Plot.

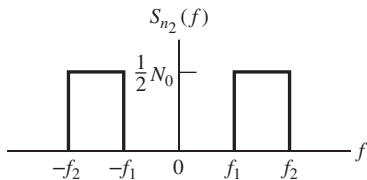


Figure 7.17

7.32 A noise waveform $n_1(t)$ has the bandlimited power spectral density shown in Figure 7.18. Find and plot the power spectral density of $n_2(t) = n_1(t) \cos(\omega_0 t + \theta) - n_1(t) \sin(\omega_0 t + \theta)$, where θ is a uniformly distributed random variable in $(0, 2\pi)$.

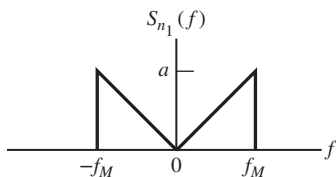


Figure 7.18

Section 7.6

7.33 Consider a signal-plus-noise process of the form

$$z(t) = A \cos 2\pi(f_0 + f_d)t + n(t)$$

where $\omega_0 = 2\pi f_0$, with

$$n(t) = n_c(t) \cos \omega_0 t - n_s(t) \sin \omega_0 t$$

an ideal bandlimited white-noise process with double-sided power spectral density equal to $\frac{1}{2}N_0$, for $f_0 - \frac{B}{2} \leq |f| \leq f_0 + \frac{B}{2}$, and zero otherwise. Write $z(t)$ as $z(t) = A \cos[2\pi(f_0 + f_d)t] + n'_c(t) \cos[2\pi(f_0 + f_d)t] - n'_s(t) \sin[2\pi(f_0 + f_d)t]$

- (a) Express $n'_c(t)$ and $n'_s(t)$ in terms of $n_c(t)$ and $n_s(t)$. Using the techniques developed in Section 7.5, find the power spectral densities of $n'_c(t)$ and $n'_s(t)$, $S_{n'_c}(f)$ and $S_{n'_s}(f)$.
- (b) Find the cross-spectral density of $n'_c(t)$ and $n'_s(t)$, $S_{n'_c n'_s}(f)$, and the cross-correlation function, $R_{n'_c n'_s}(\tau)$. Are $n'_c(t)$ and $n'_s(t)$ correlated? Are $n'_c(t)$ and $n'_s(t)$, sampled at the same instant independent?

Problems Extending Text Material

7.34 A random process is composed of sample functions of the form

$$x(t) = n(t) \sum_{k=-\infty}^{\infty} \delta(t - kT_s) = \sum_{k=-\infty}^{\infty} n_k \delta(t - kT_s)$$

where $n(t)$ is a wide-sense stationary random process with the auto correlation function $R_n(\tau)$, and $n_k = n(kT_s)$.

- (a) If T_s is chosen to satisfy

$$R_n(kT_s) = 0, \quad k = 1, 2, \dots$$

so that the samples $n_k = n(kT_s)$ are orthogonal, use Equation (7.35) to show that the power spectral density of $x(t)$ is

$$S_x(f) = \frac{R_n(0)}{T_s} = f_s R_n(0) = f_s \overline{n^2(t)}, \quad -\infty < f < \infty$$

- (b) If $x(t)$ is passed through a filter with impulse response $h(t)$ and frequency response function $H(f)$, show that the power spectral density of the output random process, $y(t)$, is

$$S_y(f) = f_s \overline{n^2(t)} |H(f)|^2, \quad -\infty < f < \infty \quad (7.152)$$

7.35 Consider the system shown in Figure 7.19 as a means of approximately measuring $R_x(\tau)$ where $x(t)$ is stationary.

- (a) Show that $E[y] = R_x(\tau)$.
- (b) Find an expression for σ_y^2 if $x(t)$ is Gaussian and has zero mean. Hint: If x_1, x_2, x_3 , and x_4 are Gaussian with zero mean, it can be shown that $E[x_1 x_2 x_3 x_4] = E[x_1 x_2]E[x_3 x_4] + E[x_1 x_3]E[x_2 x_4] + E[x_1 x_4]E[x_2 x_3]$

7.36 A useful average in the consideration of noise in FM demodulation is the cross-correlation

$$R_{yy}(\tau) \triangleq E \left\{ y(t) \frac{dy(t + \tau)}{dt} \right\}$$

where $y(t)$ is assumed stationary.

- (a) Show that

$$R_{yy}(\tau) = \frac{dR_y(\tau)}{d\tau}$$

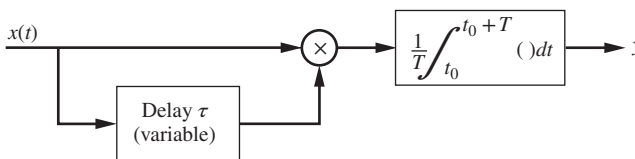


Figure 7.19

where $R_y(\tau)$ is the autocorrelation function of $y(t)$. (Hint: The frequency response function of a differentiator is $H(f) = j2\pi f$.)

- (b) If $y(t)$ is Gaussian, write down the joint pdf of

$$Y \triangleq y(t) \text{ and } Z \triangleq \frac{dy(t)}{dt}$$

at any time t , assuming the ideal lowpass power spectral density

$$S_y(f) = \frac{1}{2} N_0 \Pi\left(\frac{f}{2B}\right)$$

Express your answer in terms of N_0 and B .

- (c) Can one obtain a result for the joint pdf of y and $\frac{dy(t)}{dt}$ if $y(t)$ is obtained by passing white noise through a lowpass RC filter? Why or why not?

Computer Exercises

7.1 In this computer exercise we reexamine Example 7.1. A random process is defined by

$$X(t) = A \cos(2\pi f_0 t + \theta)$$

Using a random number generator program generate 20 values of θ uniformly distributed in the range $0 \leq \theta < 2\pi$. Using these 20 values of θ generate 20 sample functions of the process $X(t)$. Using these 20 sample functions do the following:

- Plot the sample functions on a single set of axes.
- Determine $E\{X(t)\}$ and $E\{X^2(t)\}$ as time averages.
- Determine $E\{X(t)\}$ and $E\{X^2(t)\}$ as ensemble averages.
- Compare the results with those obtained in Example 7.1.

7.2 Repeat the previous computer exercise with 20 values of θ uniformly distributed in the range $-\frac{\pi}{4} \leq \theta < \frac{\pi}{4}$.

7.3 Check the correlation between the random variables X and Y generated by the random number generator of Computer Exercise 6.2 by computing the sample

correlation coefficient of 1000 pairs according to the definition

$$\rho(X, Y) = \frac{1}{(N-1)\hat{\sigma}_1\hat{\sigma}_2} \sum_{n=1}^N (X_n - \hat{\mu}_X)(Y_n - \hat{\mu}_Y)$$

where

$$\hat{\mu}_X = \frac{1}{N} \sum_{n=1}^N X_n$$

$$\hat{\mu}_Y = \frac{1}{N} \sum_{n=1}^N Y_n$$

$$\hat{\sigma}_X^2 = \frac{1}{N-1} \sum_{n=1}^N (X_n - \hat{\mu}_X)^2$$

and

$$\hat{\sigma}_Y^2 = \frac{1}{N-1} \sum_{n=1}^N (Y_n - \hat{\mu}_Y)^2$$

7.4 Write a MATLAB program to plot the Ricean pdf. Use the form (7.150) and plot for $K = 1, 10, 100$ on the same axes. Use r/σ as the independent variable and plot $\sigma^2 f(r)$.